EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

DOANet: a deep dilated convolutional neural network approach for search and rescue with drone-embedded sound source localization

Alif Bin Abdul Qayyum, K. M. Naimul Hassan, Adrita Anika, Md. Farhan Shadiq, Md Mushfiqur Rahman, Md. Tariqul Islam, Sheikh Asif Imran, Shahruk Hossain and Mohammad Ariful Haque*

Abstract

Drone-embedded sound source localization (SSL) has interesting application perspective in challenging search and rescue scenarios due to bad lighting conditions or occlusions. However, the problem gets complicated by severe drone ego-noise that may result in negative signal-to-noise ratios in the recorded microphone signals. In this paper, we present our work on drone-embedded SSL using recordings from an 8-channel cube-shaped microphone array embedded in an unmanned aerial vehicle (UAV). We use angular spectrum-based TDOA (time difference of arrival) estimation methods such as generalized cross-correlation phase-transform (GCC-PHAT), minimum-variance-distortion-less-response (MVDR) as baseline, which are state-of-the-art techniques for SSL. Though we improve the baseline method by reducing ego-noise using speed correlated harmonics cancellation (SCHC) technique, our main focus is to utilize deep learning techniques to solve this challenging problem. Here, we propose an end-to-end deep learning model, called DOANet, for SSL. DOANet is based on a one-dimensional dilated convolutional neural network that computes the azimuth and elevation angles of the target sound source from the raw audio signal. The advantage of using DOANet is that it does not require any hand-crafted audio features or ego-noise reduction for DOA estimation. We then evaluate the SSL performance using the proposed and baseline methods and find that the DOANet shows promising results compared to both the angular spectrum methods with and without SCHC. To evaluate the different methods, we also introduce a well-known parameter—area under the curve (AUC) of cumulative histogram plots of angular deviations—as a performance indicator which, to our knowledge, has not been used as a performance indicator for this sort of problem before.

Keywords: DOA estimation, DNN, Sound source localization, UAV, DREGON, Dilated CNN

1 Introduction

Unmanned aerial vehicles (UAVs), ubiquitously known as drones, have found great use in a wide range of applications—from casual use in photography to search and rescue operations where human lives are at stake. Reports by the United Nations and other humanitarian organizations document the successful deployment of UAVs in relief efforts after natural disasters such as the major earthquakes in Haiti and Nepal in 2010 and 2015, respectively [1, 2]. UAVs have been effective because of their ability to reach areas not easily accessible by humans. They can also cover a larger area than a group of human rescuers could on foot. In search and rescue scenarios, UAVs have typically been equipped with cameras that help locate areas with rubble and debris where people might be trapped. More recently, there has been research on

*Correspondence: arifulhoque@eee.buet.ac.bd
Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, BUET Central Road, Dhaka, 1000, Bangladesh

using embedded microphone arrays in the UAVs to triangulate the sound coming from emergency whistles or humans trapped beneath debris [3–7]. It is evident that a sound source localization (SSL)-based detection system can compliment the visual detection in scenarios where the field of view may be occluded due to obstacles or bad lighting or even operations carried out at night. However, SSL is made difficult by the presence of high ego-noise generated by the rotors and propellers of the UAV. In this article, we report on our efforts to improve upon existing techniques employed in SSL systems for UAVs.

SSL algorithms generally utilize the time difference of arrival (TDOA) feature from multiple microphone pairs [7]. The TDOA can be estimated using various algorithms such as multiple signal classification (MUSIC) and generalized cross-correlation (GCC). For noise-robust SSL, a generalized eigenvalue decomposition-based multiple signal classification (GEVD-MUSIC) algorithm combined with an adaptive estimation method of the noise correlation matrix was proposed by [8]. In the context of UAVs, the drone contains multiple sensors that can provide additional real-time data about the UAV itself such as its rotor speeds and trajectory. It is natural to conclude that incorporating the additional data about the UAV dynamics can benefit SSL. As such, a method for combining information from the GCC between multiple microphone inputs, the dynamics of the UAV, and the Doppler shift in sound frequency due to motion was proposed by [3]. Since the UAV is a remote platform with limited computational capability, SSL algorithms must be computationally efficient so that sound sources can be triangulated in real time. Such an algorithm was proposed by [4] which involved a modified version of the MUSIC algorithm based on incremental generalized singular value decomposition (iGSVD-MUSIC). Furthermore, in order to locate and track a moving sound source, an approach involving time-frequency spatial filtering combined with a particle filter was described to perform well under noisy conditions by [6].

One of the primary challenges involved with SSL using UAVs is the low signal-to-noise ratio (SNR) due to the presence of several noise sources including high "ego-noise" which is the noise emanating from all the moving parts of the UAV such as the rotors and propellers. For accurate SSL, the ego-noise must be compensated for somehow, perhaps via signal enhancement or noise reduction. Recent studies have approached this problem in different ways. A method of noise estimation using learned dictionaries of ego-noise was proposed by [9]. Another study reported on using time-frequency spatial filtering combined with beamforming and blind source separation techniques [10]. Other works have utilized order analysis-based denoising algorithms [11], adaptive signal processing, and pitch shifting [12] methods. These proposed techniques all involve some form of handcrafted modeling and fine-tuning, which makes the task of ensuring robustness under different practical noise conditions difficult and laborious. There is also the possibility that the noise spectrum might overlap with the target sound source spectrum; attempts to filter the noise might inadvertently distort the target source and hence hamper SSL. More recently, there has been promising work in data-driven approaches using deep neural networks for ego-noise reduction which provides a way to bypass these problems [13, 14].

SSL using neural networks (NNs) directly is still a nascent research area, especially in the context we are considering. Generic localization methods using different neural network architectures such as convolutional neural networks (CNNs) [15] and residual neural networks (ResNets) [16] have been proposed. In other domains, such as image classification and segmentation, it is reported that CNNs with dilated kernels [17] perform better than "vanilla" CNNs [18, 19]. To the best of our knowledge, dilated CNN-based SSL has not yet been proposed.

In this article, we present our method for SSL, which was developed for the IEEE Signal Processing Cup (SP Cup) 2019 titled "Search and Rescue with drone-embedded SSL" [20]. Our proposed system called DOANet (Direction of Arrival Network) uses a one-dimensional dilated CNN fed on raw audio signals from a microphone array, to estimate the elevation and azimuth angles of a sound source while the UAV is both static and moving. We compare our system against the baseline system provided by the SP Cup organizers. The baseline method is described in greater detail in Section 3.

2 Problem setup

The problem scenario we considered for our work involved locating the direction of a speech sound source from a UAV which was either hovering (static condition) or flying (in-flight condition). The data we used for our work was shared with us by the SP Cup organizers, a novel dataset called DREGON (DRone EGonoise and localizatiON) containing recordings of a sound source made from a quadcopter UAV in static and in-flight conditions in a low-reverberant large room [7]. That is, all recordings were made with the UAV flying in an indoor environment; as such, the scope of our experiments described in this article was limited to indoor environments. The recordings were made using a cube-shaped 8-microphone array mounted below the UAV as illustrated in Fig. 1. The constellation of 8 microphones formed two parallel horizontal squares, and each of them was twisted in opposite directions in the azimuth plane, as shown in Fig. 2. The DREGON dataset is discussed in more detail in Section 5.1.

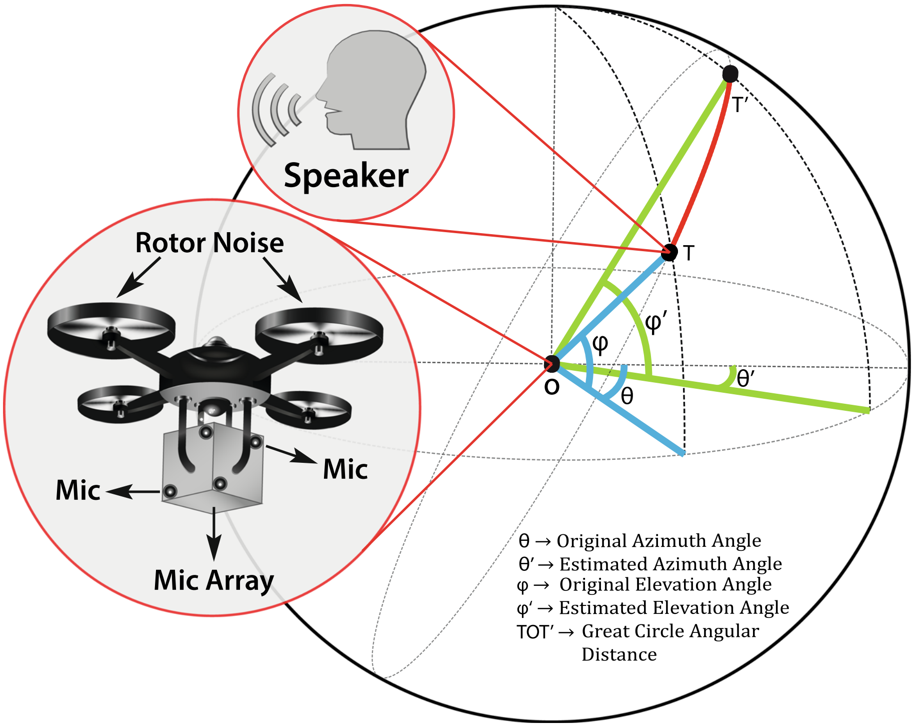
Speaker
Rotor Noise
Mic
Mic
Mic Array
T′
T
φ′
φ
θ′
θ
O
θ → Original Azimuth Angle
θ′ → Estimated Azimuth Angle
φ → Original Elevation Angle
φ′ → Estimated Elevation Angle
TOT′ → Great Circle Angular Distance

Fig. 1 Graphical representation of drone-embedded SSL. The UAV in this figure was taken from Freepik.com with permission

For 3D DOA estimation, we need to predict the azimuth and elevation angle. A naive way to evaluate the predicted DOA is to calculate the deviation of the estimated angles from the true values. A better evaluation metric is obtained by calculating the great-circle angular distance between predicted and true direction. It is a measure of angular deviation between two points in a spherical coordinate system which considers both the azimuth and elevation angles of the predicted and true direction. A visual representation of the azimuth angle, the elevation angle, and the great-circle angular distance is illustrated in Fig. 1.

As mentioned previously, the issue that makes SSL most daunting is the presence of ego-noise originating from

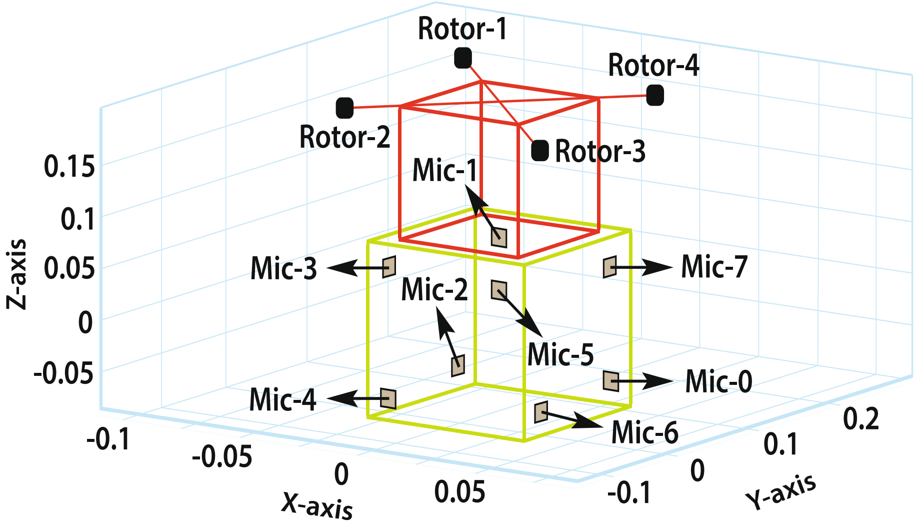

Fig. 2 Positional plot of all microphones and rotors (in meters)

the rotors and propellers of the UAV while flying or hovering. These noise sources are usually very close to the microphones resulting in negative SNR which pose quite a challenge when trying to discern the target sound source. The noisy signal received by ith microphone, $y_i(t)$, can be modeled as:

$$y_i(t) = s_i(t) + \sum_{j}^{N} n_{ij}(t) \tag{1}$$

where $s_i(t)$ is the received signal originated from the target sound source, $n_{ij}(t)$ is the received signal originated from jth noise source, $i = 1, 2, ..., 8$, and $j = 1, ..., N$. The most significant sources of ego-noise for a quadcopter UAV are its 4 rotors. So for simplicity, we can assume that N is equal to 4. In this work, our objective is to estimate the direction of sound source in terms of azimuth (θ) and elevation (φ) angles using the noisy audio signals, $y_i(t)$, where $y_i(t)$ is recorded in either in-flight or static UAV conditions.

For in-flight condition, the DREGON dataset contained recordings of two kinds of sound sources—white noise and human speech. We focused on the speech sound source in our work since SSL is more challenging for speech compared to white noise owing to the dynamic frequency content in the former. Along with actual in-flight UAV recordings, the DREGON dataset also contained recordings where the UAV was stationary and individual rotors were turned on one at a time and set to different speeds. There was no target sound source when these recordings were made. These recordings thus served as direct recordings of the rotor noise at different speeds and were utilized to analyze the characteristics of rotor noise as well as generate synthetic noisy recordings for training. For each recording in the DREGON dataset, we were also given metadata which included the actual DOA label and UAV rotor speeds at different timestamps.

3 Baseline

We compared our proposed system, DOANet, against the baseline system provided by the organizers of the SP Cup 2019. This baseline system utilized angular spectrum techniques which are described in detail in the following subsection. In our initial efforts, we found that we were able to improve the baseline system by first applying an algorithm utilizing the UAV rotor speeds to dynamically denoise the recordings. This is discussed in Section 3.2. We compared DOANet against this modified baseline system as well.

3.1 Baseline: angular spectrum method

The most common method of SSL using multiple microphones is to use time difference of arrival (TDOA) calculated between microphone pairs [7]. Assuming the sound source is far away, a one-to-one relation exists between direction of arrival (DOA) and TDOA for each microphone pair. Thus, the problem of SSL using multiple microphones is essentially a problem of TDOA estimation from microphone pairs. Generally, TDOA is addressed using the short-time Fourier transform (STFT) of the two signals. Compared to deterministic TDOA estimation, probabilistic approaches called angular spectrum-based methods perform better where a function of TDOA is generated and calculated for every possible TDOA [21].

Let us consider a microphone pair (i,j) from M microphones. Let $Y_i(t,f)$ and $Y_j(t,f)$ represent the STFT of noisy microphone signals $y_i(t)$ and $y_j(t)$, respectively, as denoted in Eq. 1. For the microphone pair, a set of TDOA values can be linked with all possible DOA (θ, φ), where θ and φ represent the azimuth and elevation angles. To do so, a set of points $S(x, y, z)$ is taken on the 3D plane covering a uniform grid of (θ, φ):

$$S(x, y, z) = S\left(\cos(\varphi)\cos(\theta), \cos(\varphi)\sin(\theta), \sin(\varphi)\right)$$

Denoting the displacement vector from jth to ith microphone by $\boldsymbol{d_{ij}}$ and wave propagation speed by c, the TDOA between the two microphone for each possible DOA, $\tau_{ij}(\theta, \varphi)$, can be computed as follows:

$$\tau_{ij}\left(\theta, \varphi\right) = \frac{\boldsymbol{d_{ij}} \cdot S\left(x, y, z\right)}{c} \tag{2}$$

The next step is to construct a function of $\tau_{ij}(\theta, \varphi)$ utilizing $Y_i(t,f)$ and $Y_j(t,f)$ which will peak for true τ_{ij}. This function is called local angular spectrum function and is denoted by $\phi_{ij}(t, f, \tau)$. One way to do this is a technique called generalized cross-correlation phase-transform (GCC-PHAT) [21] which produces the following function:

$$\phi_{ij}^{\text{GCC-PHAT}}\left(t, f, \tau\right) = \Re\left(\frac{Y_i(t,f)\overline{Y_j(t,f)}}{\mid Y_i(t,f)\overline{Y_j(t,f)}\mid} e^{-2j\pi f \tau_{ij}}\right) \tag{3}$$

For robust DOA estimation, $\phi_{ij}(t, f, \tau)$ is summed over all frequencies, microphone pairs, and time frames. In cases where the sound source may not be active throughout all time frames, taking the maximum is preferred to summing over time the total time span [21]. Thus, we obtain a global angular spectrum $\phi(\theta, \varphi)$ for each possible direction:

$$\phi(\theta, \varphi) = \sum_{t} or \max \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \sum_{f} \phi_{ij}\left(t, f, \tau\right) \tag{4}$$

Finally, DOA is estimated by the local peak finding method from $\phi(\theta, \varphi)$.

There are several techniques for building $\phi_{ij}(t,f,\tau)$ other than GCC-PHAT. The GCC-PHAT method is however the most popular choice [21]. The baseline system provided by the SP Cup organizers also considered six other techniques for building the local angular spectrum function, $\phi_{ij}(t,f,\tau)$. Generalized cross-correlation with a non-linear function (GCC-NONLIN) is a slightly modified version of GCC-PHAT where a non-linear function is applied on GCC-PHAT to emphasize large values. The other five methods are SNR based and have been proposed in [21]. The general scheme involves calculating the directional SNR by extracting target signal and noise power for every possible direction and using the assumption that SNR is likely to peak for the true direction. Such methods have the advantage of ignoring erroneous contribution from other directions. Among the five SNR-based methods, two of them use beamformer-based methods to separate the target signal and noise, one is a statistical method, and the rest are hybrids of the beamformer and statistical methods.

The two beamformer methods are the minimum-variance-distortion-less-response (MVDR) and delay-and-sum (DS) methods which work based on Capon (or MVDR) and classical (or Bartlett) beamformers, respectively [22, 23]. MVDR beamformer generally performs better than classical beamformer as all degrees of freedom are used to maximize energy on the specific direction [24]. However, these beamformer-based methods tend to overestimate the SNR at low frequencies. This problem is addressed by the diffuse noise model (DNM) method where SNR is estimated a priori using a statistical mixture model of one predominant source and noise. Weighted minimum-variance-distortion-less-response (MVDRW) and weighted delay-and-sum (DSW) methods are formulated by combining DNM with MVDR and DS, respectively.

3.2 Modified baseline: speed correlated harmonics cancellation with angular spectrum

Acoustic noise in recorded audio during UAV flights consists of three major components [7]. These components are ego-noise, air flow noise from the propellers, and wind noise; ego-noise is the most significant in terms of noise power spectrum and is principally generated by the rotors of the UAV. The DREGON dataset contained recordings where the UAV was kept stationary, with individual rotors turned on one at a time and ramped up to various speeds. These recordings served as noise samples for each rotor. The paper detailing the DREGON dataset [7] showed that the peaks of power spectral density for these individual rotor recordings varied proportionally with the rotor speed.

In our literature review, we came across works that also noted this type of relationship and utilized it for UAV noise harmonics cancellation [11, 12, 14]. The noise power spectrum of one of the rotors at different speeds is shown in Fig. 3. We analyzed all the available recordings of rotor noise in the DREGON dataset and used simple linear regression between the first harmonic of the rotor noise and rotor speed to obtain the following relationship:

$$f_0^{\text{ego}}(r_s) = \alpha \cdot r_s \tag{5}$$

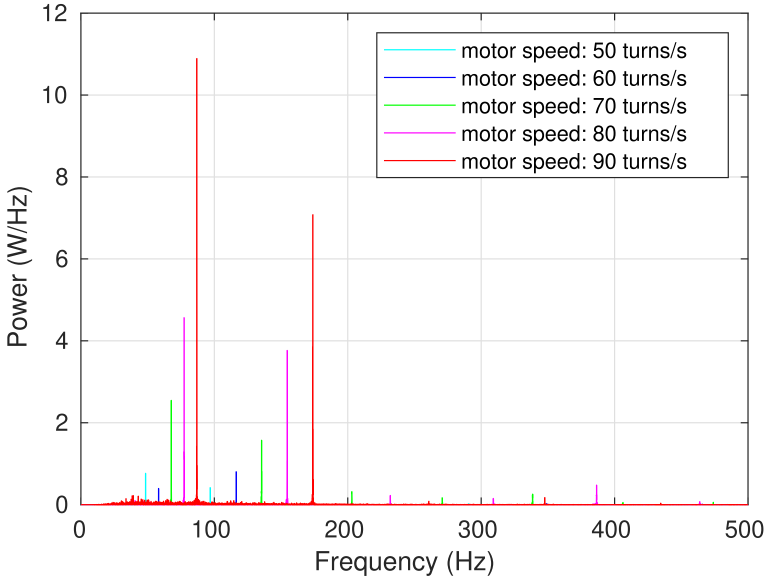

Fig. 3 Frequency bins of ego-noise for different rotor speeds

where $f_0^{\text{ego}}(r_s)$ is the first harmonic of ego-noise as a function of the rotor speed, and r_s and α are the proportionality constant, computed from the gradient of the plot of $f_0^{\text{ego}}(r_s)$ vs. r_s. The value of α thus obtained was 0.98.

In the modified baseline method, we designed band-pass filters and applied them on the target sound source recordings to extract the harmonics given by Eq. 5 for different rotor speeds the UAV happened to be flying or hovering at during the recording. The signals obtained from the filters were subtracted from the original audio. The resulting signal was likely to have a better SNR. This denoised signal was then fed through the original baseline system described in Section 3.1. This process is illustrated in Fig. 4. Since this modification to the baseline system involves suppressing the noise that is correlated with rotor speed, we refer to this method as speed correlated harmonics cancellation (SCHC).

4 Proposed system

We propose an end-to-end one-dimensional dilated convolutional neural network, called DOANet. Our network accepts multi-channel raw audio signals from the microphone array and estimates the DOA of the sound source by predicting the azimuth and elevation angles. The SSL system using DOANet is illustrated in Fig. 5. Over the course of our work, we found having two separate models for predicting azimuth and elevation angle separately worked better than trying to do so using a single model. So DOANet is composed of two networks which are almost identical (discussed further in Section 4.2), each taking on the task of predicting the azimuth and elevation angles independently.

The raw 8 channel audio signals are first passed through a channel selection block which can be configured to select the appropriate channels. The selected channels are then windowed and propagated through the DOANet model.

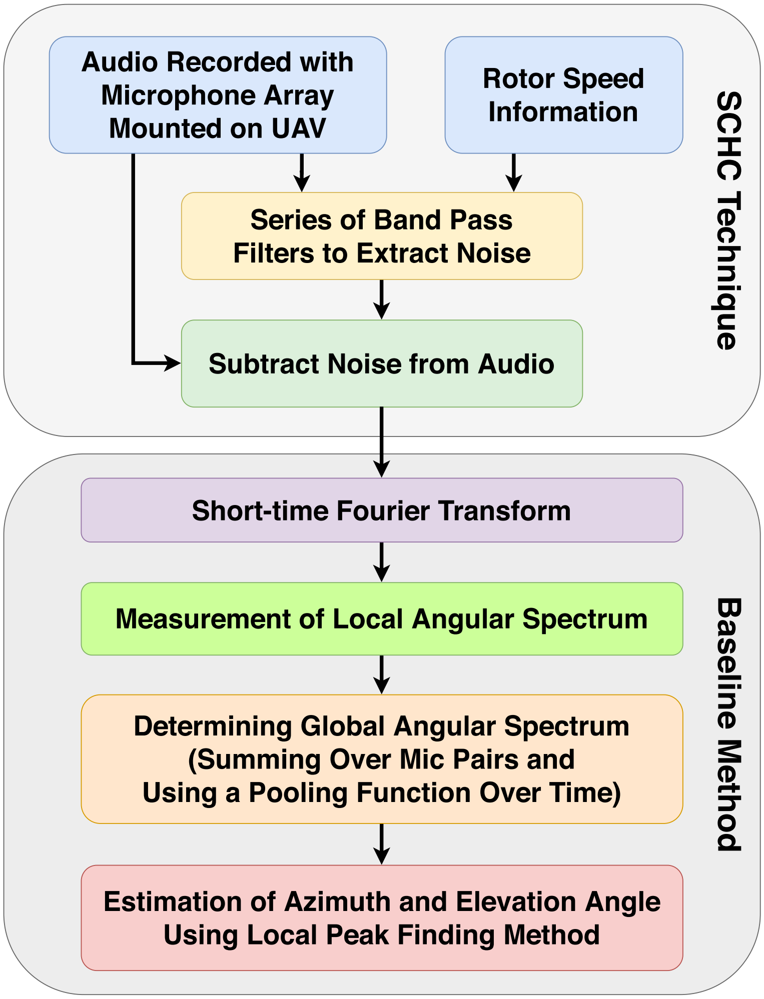
Audio Recorded with Microphone Array Mounted on UAV
Rotor Speed Information
Series of Band Pass Filters to Extract Noise
Subtract Noise from Audio
SCHC Technique
Short-time Fourier Transform
Measurement of Local Angular Spectrum
Determining Global Angular Spectrum (Summing Over Mic Pairs and Using a Pooling Function Over Time)
Estimation of Azimuth and Elevation Angle Using Local Peak Finding Method
Baseline Method

Fig. 4 Block diagram of the modified baseline method

4.1 Channel selection

We have two configurable modes for the channel selection block: CS (channel separation) and ACU (all channel utilization). In the ACU mode, DOANet uses all 8 audio channels. In the CS configuration, we create two different sets of audio signals—the first set consists of microphones 0, 1, 4, and 5 and is referred to as CS0145 in the rest of this article; the second set consists of the remaining microphones 2, 3, 6, and 7 and is referred to as CS2367. The spatial location and orientation of the microphones are illustrated in Fig. 2. These two sets were chosen to ensure maximum spatial diversity of selected microphones. We trained separate networks for each of these sets.

4.2 Network architecture for DOANet

The networks within DOANet follow the typical architecture used in CNN-based state-of-the-art systems. However, instead of applying the usual convolution operation, we use *dilated* convolutions. The convolutional kernel or filter is expanded over different sample ranges using different dilation rates. As the dilation rate is increased, the gap between original convolution filter elements gets wider. This allows a kernel of the same size to incorporate information from a larger context [18, 19]. We were motivated to use dilated convolution for audio signals as it had been successfully applied in speech and music synthesis [25], and speech recognition [26] from raw audio signals.

The detailed network architecture of DOANet is illustrated in Fig. 6. Overall, there are 9 convolutional layers, each one followed by a ReLU activation function and a batch normalization layer [27, 28]. Each layer has a higher dilation rate than the previous. Compared to the general scheme of using powers of 2, dilation rates following a Fibonacci sequence were shown to perform slightly better [18]. We thus used the following sequence of dilation rates: 1, 2, 3, 5, 8, 13, 21, 34, and 55. After two consecutive convolutions, there is a max pooling layer with a filter size of 2, except the first max pooling layer whose filter size is 3. After the final convolution, we have a global max pooling layer. The pooled output is passed through a couple of layers with fully connected neurons (dense layers) and *tanh* activation function at the last layer which generate the network's output.

As mentioned earlier, we use two different networks trained independently for predicting the azimuth and elevation angles in the DOANet. These networks primarily differ in their hyper-parameters which are shown in Fig. 6. Another difference is that the network for estimating azimuth angle has two output nodes which map to the x-axis and y-axis projections of the azimuth angle. The reasoning for this is discussed in Section 4.3. The total

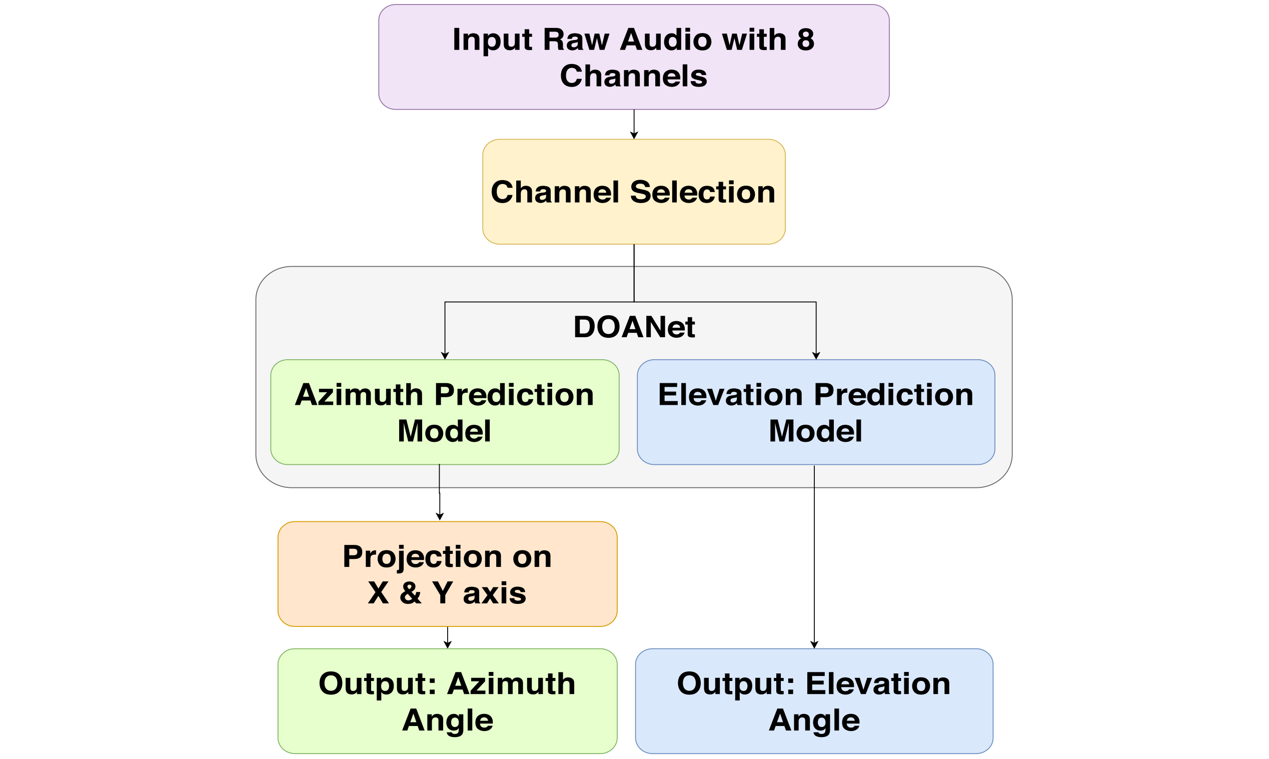

Fig. 5 DOANet-based sound source localization

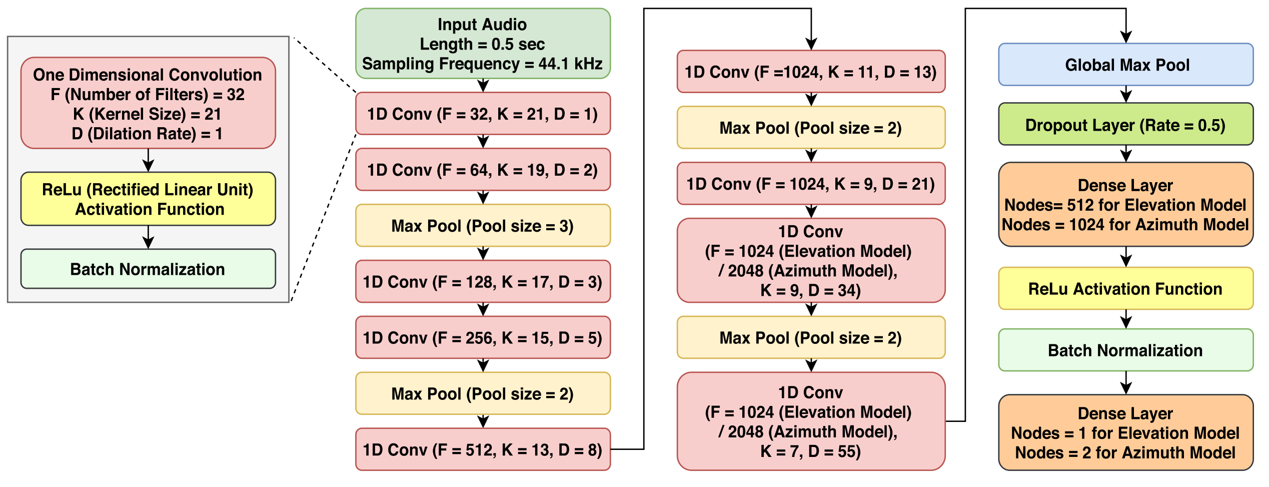
One Dimensional Convolution
F (Number of Filters) = 32
K (Kernel Size) = 21
D (Dilation Rate) = 1
ReLu (Rectified Linear Unit) Activation Function
Batch Normalization
Input Audio
Length = 0.5 sec
Sampling Frequency = 44.1 kHz
1D Conv (F = 32, K = 21, D = 1)
1D Conv (F = 64, K = 19, D = 2)
Max Pool (Pool size = 3)
1D Conv (F = 128, K = 17, D = 3)
1D Conv (F = 256, K = 15, D = 5)
Max Pool (Pool size = 2)
1D Conv (F = 512, K = 13, D = 8)
1D Conv (F =1024, K = 11, D = 13)
Max Pool (Pool size = 2)
1D Conv (F = 1024, K = 9, D = 21)
1D Conv
(F = 1024 (Elevation Model)
/ 2048 (Azimuth Model),
K = 9, D = 34)
Max Pool (Pool size = 2)
1D Conv
(F = 1024 (Elevation Model)
/ 2048 (Azimuth Model),
K = 7, D = 55)
Global Max Pool
Dropout Layer (Rate = 0.5)
Dense Layer
Nodes= 512 for Elevation Model
Nodes = 1024 for Azimuth Model
ReLu Activation Function
Batch Normalization
Dense Layer
Nodes = 1 for Elevation Model
Nodes = 2 for Azimuth Model

Fig. 6 Network architecture of DOANet

number of parameters for each model is summarized in Table 1.

4.3 DOA estimation from DOANet

The elevation angle prediction network of DOANet outputs a number between -1 and 1 which correspond to the scaled elevation angle (actual elevation angle between -90 and $+90°$ divided by 90). However, the output of the azimuth angle network is not the scaled azimuth angle; instead, it is the x-axis and y-axis projection of a unit length two-dimensional vector. For an azimuth angle, θ projections on the x- and y-axes are $x = \cos\theta$ and $y = \sin\theta$. We observed that having the network predict the projections worked better than making it predict the angle. We hypothesize that this may offer the network more flexibility in learning the DOA on the xy plane, since the projections on the two axes are independent. We use the trigonometric relation $\theta = \tan^{-1}(y/x)$ to calculate the azimuth angle from the predicted projection values. Thus, the predicted elevation angle and azimuth angle together provide DOANet's estimate of the DOA of the sound source.

5 Experiments

In this section, we describe the dataset, experimental setup, and evaluation metric used in our study.

Table 1 Total number of parameters for DOANet models

Model	Trainable	Non-trainable	Total
Azimuth	67,938,850	16,320	67,955,170
Elevation	34,902,177	11,200	34,913,377

5.1 SP Cup 2019 data

For training and evaluating our system, we used a subset of the DREGON dataset [7] compiled by the organizers of the IEEE SP Cup 2019 [29]. The dataset contained multi-channel audio files recorded in a large low-reverberant room, using the microphone array embedded on a quadcopter UAV. A speaker was placed at the center of the room which played different audio clips taken from the TIMIT dataset [30] containing human speech. The dataset also contained recordings where the speaker played white noise instead of human speech, but we did not include them in this study since SSL is more challenging for speech compared to white noise owing to the dynamic frequency content in speech. The recordings were grouped into two categories: static task and in-flight task. Files in the static task category were recorded with the UAV hovering in a fixed position. Similarly, the in-flight task category contained files recorded when the UAV was flying around the room. The dataset also contained metadata for each recording related to the position of the UAV in the room tracked with 3D Motion Capture Hardware and UAV rotor speeds at different timestamps within the recordings. The dataset was shared with us by the SP Cup organizers in two phases: primary round data and final round data. The final round data was only used for evaluation, while the primary round data was used for training and validation. The summary of the audio data split into train, validation, and test sets is shown in Table 2, while the following sections detail how the data was prepared.

5.1.1 Primary round data

The primary round data contained 300 static audio files around 2 to 3 s long and 16 in-flight audio files which were

Table 2 Audio data points created from SP Cup 2019 data

Task	Train	Validation	Test
Static	1126	569	120
In-flight	180	60	80

4 s in duration. The static files were randomly divided into training and validation sets with 200 files for training and 100 files for validation. The train and validation data for in-flight files were divided in a 3:1 ratio. For training DOANet, we segmented all the static audio files into 0.5-s clips. The in-flight files were also segmented in the same way with metadata (DOA labels, rotor speeds) at 15 timestamps as follows: 0.25 s, 0.5 s, 0.75 s, 1 s, 1.25 s, 1.5 s, 1.75 s, 2 s, 2.25 s, 2.5 s, 2.75 s, 3 s, 3.25 s, 3.5 s, and 3.75 s. As a result, we obtained 1126 and 569 static train and validation data points, respectively. For in-flight data, we had 180 and 60 train and validation data points, respectively.

5.1.2 Final round data

The final round data added a further 20 static audio files with duration ranging from 2 to 4 s and 1 in-flight audio file with a duration of 20 s. The static audio files were split in the same way as the primary round data, resulting in a total of 120 data points. The in-flight speech audio file had a total of 80 timestamps for which metadata was provided. The timestamps were at intervals of 0.25 s, each covering 0.5 s of the recording. The entirety of the final round data was used only for evaluating the trained DOANet.

5.2 Synthetic data

The amount of audio data provided in the SP Cup 2019 was not sufficient for properly training a deep neural network such as DOANet. Therefore, we created a synthetic static audio dataset using the open-source *pyroomacoustics* package [31]. This package allowed us to simulate indoor environments where we could place a sound source, noise sources, and microphones at different positions in the virtual space.

We created a virtual 10 m × 10 m × 5 m room which was comparable to the environment where the DREGON recordings were made. We also constructed a virtual UAV to mimic the one used in the DREGON dataset, with an 8-microphone array and 4 noise sources located at the four rotor positions in the relative positions as shown in Fig. 2 and described in [7]. We wanted our synthetic data to match the DREGON dataset as much as possible. To that end, we extracted the rotor ego-noise from static audio files in the primary round data (Section 5.1.1) using a generalized sidelobe canceller (GSC) beamformer [32]. The noise sources placed at the rotor locations were made to emit these extracted noise signals.

We then added a sound source located at the floor of the room and made it emit random clips of human speech from the TIMIT dataset, to act as our target sound source. The virtual UAV was then placed at random positions in the virtual room, and the simulated recordings of the microphone array were generated. The positions of the virtual sound source and UAV were used to calculate the DOA labels for each recording. Using this simulation technique, we were able to generate a large 8-channel synthetic audio dataset containing 2980 recordings to train DOANet.

5.3 DOANet model training

The DOANet model was built using the *Keras* [33] and *Tensorflow* [34] frameworks and trained on Kaggle Notebooks' GPU instances. The model was trained in three stages. In the first stage, DOANet was trained from scratch on the synthetic data described in Section 5.2 until the model converged, i.e., until validation objective function plateaued. In the subsequent two stages, the model was fine-tuned using the training and validation partitions of the DREGON dataset described in Section 5.1.1; first, the static data points were used and then the in-flight data points. This training scheme was inspired by the "curriculum learning" approach proposed in [35], and we found that it helped the final model converge faster and more accurately than when training with all the data mixed together in a single stage.

The objective function for the training algorithm was the mean squared error (MSE) between the predictions and ground truth labels (i.e., scaled elevation angles, and x-y projections of the azimuth angles). We used the popular "Adam" [36] optimizer algorithm with an initial learning rate of 0.01 when training from scratch. During the fine-tuning stages, the initial learning rate was set at 0.001. We decreased the learning rate by a factor of 10 every time the objective function on the validation data stalled or started getting worse for consecutive training iterations. We did this up to three times before stopping the training run. On average, the first stage of training lasted for 50 epochs on synthetic data and fine-tuning stages for 35 epochs on the real data. The total training time was about 6 h for each azimuth model and about 4 h for each elevation model.

5.4 Performance evaluation

The proposed system, along with the baseline and modified baseline systems described in Section 3.1 and 3.2, respectively, was evaluated on the final round data of the DREGON dataset as described in Section 5.1.2. For the modified baseline using SCHC, we computed the proportionality constant in Eq. 5 from the training data and obtained a value of $\alpha = 0.98$. We limited the number of bandpass filters used to extract the ego-noise harmonics to 10 after determining that no gain in accuracy was obtained beyond this number.

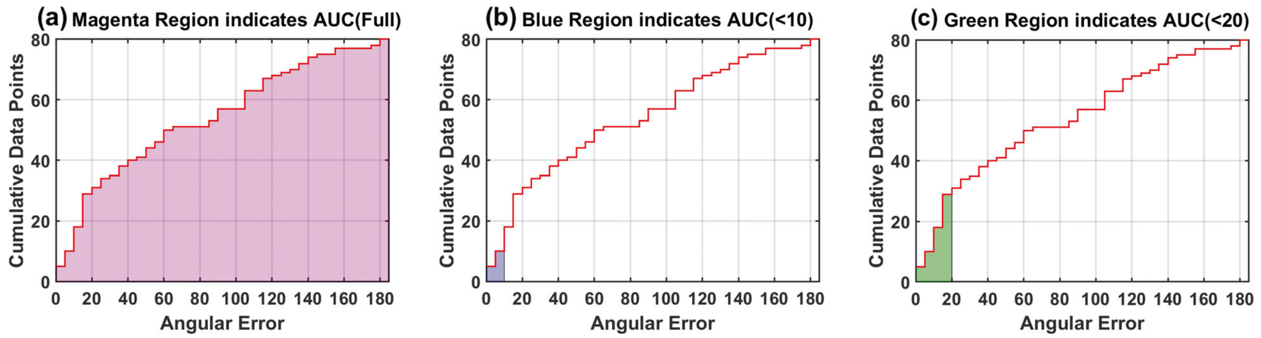

Fig. 7 Typical cumulative histograms and the shaded area in each subplot is the AUC. The normalized AUC is obtained by dividing the shaded area by the total area of the rectangular box. In these plots, the *x*-axis indicates the angular value and *y*-axis indicates the cumulative number of data points. Fractional AUCs are calculated and referred to as AUC($< \delta$), indicating the area under the curve from angular deviation = 0 to angular deviation = δ. For example, the AUC, AUC($<$ 10), and AUC($<$ 20) are shown in **a**, **b**, and **c**

For each system, we calculated the azimuth and elevation angle deviation and great-circle angular distance as described in Section 2. The possible range of values for these metrics (180°) was divided into 36 equal bins of 5°. The values obtained were plotted in cumulative histograms using these 36 bins. Finally, we calculated the normalized area under the curve (AUC) for all three systems and compared them.

Figure 7 shows typical cumulative histograms, and the shaded area in each subplot is the AUC. The normalized AUC is obtained by dividing the shaded area by the total area of the rectangular box. In these plots, the *x*-axis indicates the angular value and *y*-axis indicates the cumulative number of data points. Using the normalized AUC values, we sorted out the best technique or scheme for the baseline systems and our proposed system.

We chose AUC as our key performance indicator over conventional accuracy (number of correct predictions divided by total predictions) because of its inherent quality of measuring the system's *consistency* in predicting the

Table 3 Nomenclature used in presenting results

Technique	Nomenclature
DNM + SCHC	DNM with SCHC
DS + SCHC	DS with SCHC
DSW + SCHC	DSW with SCHC
GCC-NONLIN + SCHC	GCC-NONLIN with SCHC
GCC-PHAT + SCHC	GCC-PHAT with SCHC
MVDR + SCHC	MVDR with SCHC
MVDRW + SCHC	MVDRW with SCHC
DOANet + CS0145	DOANet with channel separation (channels = 0, 1, 4, 5)
DOANet + CS2367	DOANet with channel separation (channels = 2, 3, 6, 7)
DOANet + ACU	DOANet with all channel utilization
DOANet + CS0145(A) + CS0145(E)	DOANet with channel separation (channels for azimuth = 0, 1, 4, 5 and for elevation = 0, 1, 4, 5)
DOANet + CS0145(A) + CS2367(E)	DOANet with channel separation (channels for azimuth = 0, 1, 4, 5 and for elevation = 2, 3, 6, 7)
DOANet + CS0145(A) + ACU(E)	DOANet with channel separation for azimuth (channels = 0, 1, 4, 5) and all channel utilization for elevation
DOANet + CS2367(A) + CS0145(E)	DOANet with channel separation (channels for azimuth = 2, 3, 6, 7 and for elevation = 0, 1, 4, 5)
DOANet + CS2367(A) + CS2367(E)	DOANet with channel separation (channels for azimuth = 2, 3, 6, 7 and for elevation = 2, 3, 6, 7)
DOANet + CS2367(A) + ACU(E)	DOANet with channel separation for azimuth (channels = 0, 1, 4, 5) and all channel utilization for elevation
DOANet + ACU(A) + CS0145(E)	DOANet with all channel utilization for azimuth and channel separation for elevation (channels = 0, 1, 4, 5)
DOANet + ACU(A) + CS2367(E)	DOANet with all channel utilization for azimuth and channel separation for elevation (channels = 2, 3, 6, 7)
DOANet + ACU(A) + ACU(E)	DOANet with all channel utilization for both azimuth and elevation

accurate DOA. Generally, we consider a prediction accurate if the angular deviation of the prediction is within a predefined margin of error. But the problem with this approach is that "slightly wrong" and "grossly wrong" are treated the same. Likewise, the granularity, in how correct a prediction is, is not preserved either. To avoid this, we opted to use AUC for comparing different systems and analyzing the consistency in a system's ability to correctly estimate the DOA.

We also calculate fractional AUCs referred to as AUC($<$ δ), indicating the area under the curve from angular deviation = 0 to angular deviation = δ. For example, the AUC, AUC($<$ 10), and AUC($<$ 20) are shown in Fig. 7a–c. From these figures, we can infer that *higher* AUC value results in *lower* standard deviation for angular error of azimuth and elevation and great-circle angular distance.

6 Results

This section presents and compares the results for the baseline, modified baseline, and proposed systems configured in different schemes. The nomenclature used for specifying different configurations and techniques for which results are presented is shown in Table 3. The best performing techniques for the baseline system (angular spectrum methods), modified baseline system (angular spectrum methods with SCHC), and DOANet are shown with blue, green, and red colors, respectively, in all the tables in the subsequent sections.

Table 4 AUC of static azimuth angle deviation

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.6917	0.0183	0.0417
DS	**0.7706**	0.0213	0.0509
DSW	0.6977	0.0208	0.0461
GCC-NONLIN	0.7465	**0.0238**	**0.0521**
GCC-PHAT	0.7569	0.0231	0.0516
MVDR	0.7500	0.0218	0.0498
MVDRW	0.6931	0.0194	0.0428
DNM + SCHC	0.6734	0.0183	0.0417
DS + SCHC	**0.7789**	0.0213	0.0509
DSW + SCHC	0.7000	0.0208	0.0461
GCC-NONLIN + SCHC	0.7498	**0.0248**	**0.0530**
GCC-PHAT + SCHC	0.7625	0.0236	0.0525
MVDR + SCHC	0.7356	0.0206	0.0479
MVDRW + SCHC	0.6931	0.0190	0.0424
DOANet + CS0145	0.6875	0.0150	0.0509
DOANet + CS2367	**0.7941**	**0.0225**	**0.0653**
DOANet + ACU	0.7806	0.0178	0.0549

Table 5 AUC of static elevation angle deviation

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.8845	0.0206	0.0537
DS	0.8662	0.0208	0.0535
DSW	0.8731	0.0206	0.0507
GCC-NONLIN	0.8931	0.0225	0.0588
GCC-PHAT	**0.8970**	**0.0227**	**0.0600**
MVDR	0.8593	0.0185	0.0507
MVDRW	0.8782	0.0189	0.0528
DNM + SCHC	0.8780	0.0199	0.0507
DS + SCHC	0.8630	0.0201	0.0514
DSW + SCHC	0.8755	0.0208	0.0521
GCC-NONLIN + SCHC	0.8914	0.0222	0.0579
GCC-PHAT + SCHC	**0.8988**	**0.0227**	**0.0611**
MVDR + SCHC	0.8657	0.0178	0.0500
MVDRW + SCHC	0.8792	0.0197	0.0542
DOANet + CS0145	0.9162	0.0116	0.0551
DOANet + CS2367	**0.9192**	0.0116	**0.0581**
DOANet + ACU	0.9169	**0.0123**	0.0558

Table 6 AUC of static great-circle angular distance

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.7102	0.0157	0.0389
DS	**0.7877**	0.0169	0.0426
DSW	0.7428	0.0171	0.0407
GCC-NONLIN	0.7785	**0.0190**	**0.0456**
GCC-PHAT	0.7845	0.0181	0.0447
MVDR	0.7819	0.0155	0.0405
MVDRW	0.7435	0.0146	0.0368
DNM + SCHC	0.6935	0.0146	0.0359
DS + SCHC	0.7819	0.0150	0.0394
DSW + SCHC	0.7470	0.0171	0.0412
GCC-NONLIN + SCHC	0.7803	**0.0190**	**0.0456**
GCC-PHAT + SCHC	**0.7910**	0.0181	0.0449
MVDR + SCHC	0.7731	0.0139	0.0389
MVDRW + SCHC	0.7440	0.0148	0.0375
DOANet + CS0145(A) + CS0145(E)	0.9014	0.0032	0.0269
DOANet + CS0145(A) + CS2367(E)	0.9028	0.0035	0.0278
DOANet + CS0145(A) + ACU(E)	0.9028	0.0037	0.0285
DOANet + CS2367(A) + CS0145(E)	0.9169	0.0063	0.0352
DOANet + CS2367(A) + CS2367(E)	**0.9194**	**0.0067**	**0.0368**
DOANet + CS2367(A) + ACU(E)	0.9181	0.0065	0.0361
DOANet + ACU(A) + CS0145(E)	0.9113	0.0051	0.0292
DOANet + ACU(A) + CS2367(E)	0.9150	0.0049	0.0313
DOANet + ACU(A) + ACU(E)	0.9127	0.0053	0.0303

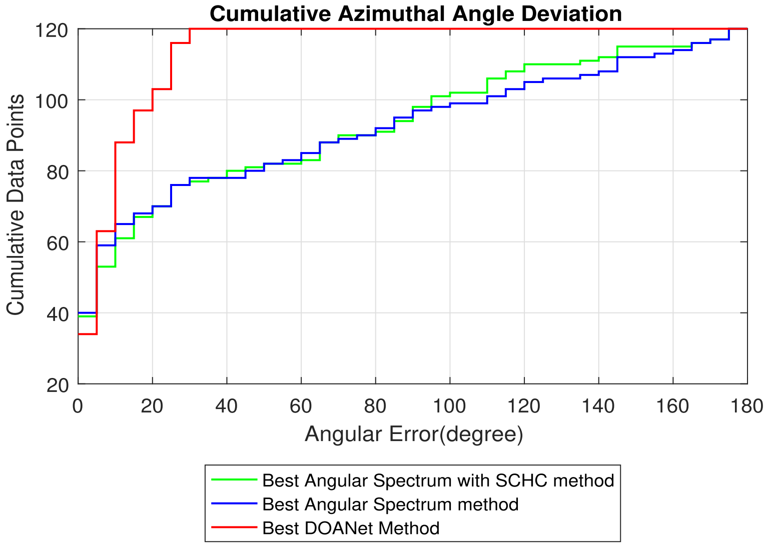

Fig. 8 AUCs for cumulative static azimuth angle deviation

6.1 Static task performance analysis

The AUC for azimuth angle deviation, elevation angle deviation, and great-circle angular distance deviation for the different systems are presented in Tables 4, 5, and 6, respectively. The AUCs for the best scheme from each system are also provided in Figs. 8, 9, and 10.

6.1.1 Azimuth angle deviation

Table 4 shows that for the baseline system, using the delay-and-sum (DS) technique gave the best results. Overall, the best performing system was DOANet using microphone channels 2, 3, 6, and 7. From Fig. 8, it is clear that the range of angular deviations for DOANet is much lower than for the baseline systems. This indicates DOANet is more consistent with its predictions than the other systems.

6.1.2 Elevation angle deviation

Table 5 shows that for the baseline system, using the generalized cross-correlation phase-transform (GCC-PHAT) technique gave the best results. Overall, the best performing system was still DOANet using microphone channels 2, 3, 6, and 7. From Fig. 9, we again see that the range of angular deviation for DOANet is more restricted compared to the baseline systems.

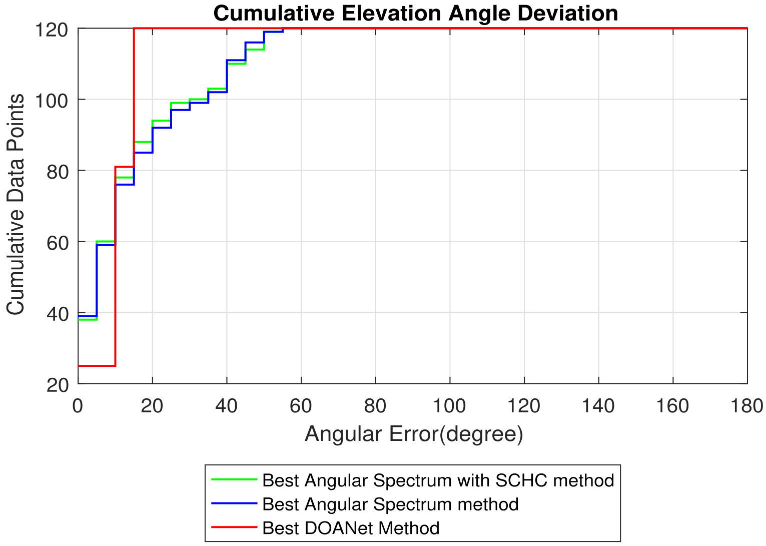

Fig. 9 AUCs for cumulative static elevation angle deviation

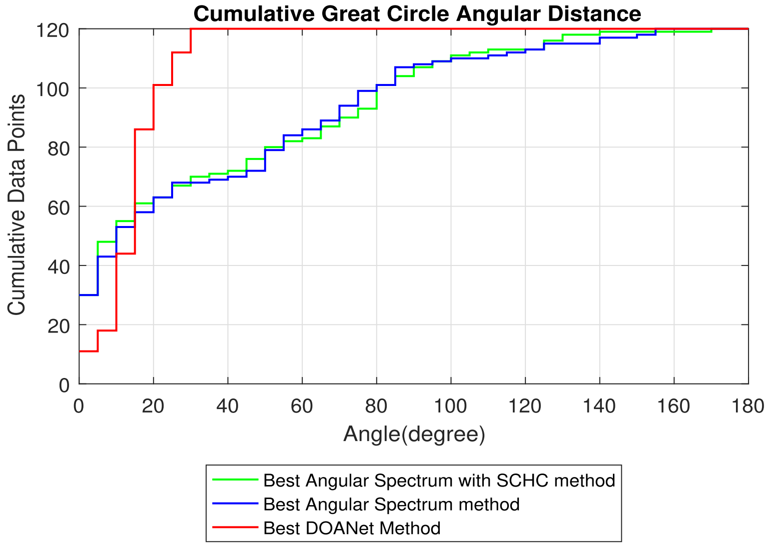

Fig. 10 AUCs for cumulative static great-circle angular distance

6.1.3 Great-circle angular distance

Table 6 shows that when comparing the great-circle angular distance, which is a measure that combines both the azimuth and elevation angles, all configurations of DOANet are better than the two baseline systems by a significant margin. This is also evident in Fig. 10, where we can see that the angular deviations for the baseline systems cover a wider range and therefore are less consistent. We also see that using microphone channels 2, 3, 6, and 7 gave the best results for DOANet. If we consider AUC(<10) and AUC(<20), however, we do see that DOANet falls a little short. This indicates that the baseline systems have a better angular resolution for these samples with low angular deviation.

6.2 In-flight task performance analysis

The AUC for azimuth angle deviation, elevation angle deviation, and great-circle angular distance deviation for the different systems are presented in Tables 7, 8, and 9, respectively. The AUCs for the best scheme from each system are also provided in Figs. 11, 12, and 13. It is worth noting that for all the metrics considered, both DOANet and modified baseline system outperformed the baseline system by a significant margin.

6.2.1 Azimuth angle deviation

Table 7 shows that for the baseline system, using the weighted delay-and-sum (DSW) technique gave the best results. Overall, the best performing system was DOANet using microphone channels 2, 3, 6, and 7. From Fig. 11, we can see that compared to the baseline system, both the modified baseline system and DOANet perform significantly better. For smaller angle deviations, the modified baseline system has a slight edge over DOANet.

6.2.2 Elevation angle deviation

Table 8 shows that for the baseline and modified baseline systems, using the delay-and-sum (DS) and weighted delay-and-sum (DSW) techniques gave the best results, respectively. Overall, the best performing system was DOANet using microphone channels 0, 1, 4, and 5. From Fig. 12, we can see that the performance of both the modified baseline system and DOANet is better than the

Table 7 AUC of in-flight azimuth angle deviation

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.3378	0.0045	**0.0101**
DS	0.3392	0.0031	0.0073
DSW	**0.3587**	0.0045	**0.0101**
GCC-NONLIN	0.3462	0.0042	0.0090
GCC-PHAT	0.3285	0.0031	0.0066
MVDR	0.3382	0.0042	0.0083
MVDRW	0.3583	**0.0049**	0.0097
DNM + SCHC	0.6785	0.0089	0.0253
DS + SCHC	**0.6965**	**0.0097**	**0.0306**
DSW + SCHC	0.6281	0.0066	0.0191
GCC-NONLIN + SCHC	0.6576	0.0063	0.0198
GCC-PHAT + SCHC	0.6313	0.0063	0.0170
MVDR + SCHC	0.6788	0.0089	0.0250
MVDRW + SCHC	0.6073	0.0028	0.0101
DOANet + CS0145	0.6875	0.0045	**0.0153**
DOANet + CS2367	**0.7941**	00.0035	0.0146
DOANet + ACU	0.7806	**0.0049**	0.0139

baseline. Unlike previous scenarios, DOANet obtained a better AUC(<10) score than the other systems.

Table 8 AUC of in-flight elevation angle deviation

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.5566	0	0
DS	**0.5691**	**0.0014**	**0.0035**
DSW	0.5642	0	0
GCC-NONLIN	0.5545	0	0
GCC-PHAT	0.5559	0	0
MVDR	0.5552	0	0
MVDRW	0.5618	0	0
DNM + SCHC	0.9156	**0.0163**	**0.0510**
DS + SCHC	0.8837	0.0128	0.0392
DSW + SCHC	**0.9233**	0.0139	0.0493
GCC-NONLIN + SCHC	0.8958	0.0087	0.0337
GCC-PHAT + SCHC	0.8917	0.0097	0.0337
MVDR + SCHC	0.9097	0.0153	0.0451
MVDRW + SCHC	0.9153	0.0146	0.0431
DOANet + CS0145	**0.9740**	**0.0330**	**0.0858**
DOANet + CS2367	0.9726	0.0326	0.0847
DOANet + ACU	0.9653	0.0250	0.0771

Table 9 AUC of in-flight great-circle angular distance

Technique	AUC	AUC(<10)	AUC(<20)
DNM	0.3091	0	0
DS	0.3066	0	0
DSW	0.3108	0	0
GCC-NONLIN	0.3281	0	0
GCC-PHAT	**0.3302**	0	0
MVDR	0.3076	0	0
MVDRW	0.3229	0	0
DNM + SCHC	0.7128	0.0045	0.0188
DS + SCHC	0.7079	**0.0052**	**0.0191**
DSW + SCHC	0.7020	0.0028	0.0115
GCC-NONLIN + SCHC	0.6975	0.0024	0.0115
GCC-PHAT + SCHC	0.6715	0.0017	0.0083
MVDR + SCHC	**0.7256**	0.0038	0.0156
MVDRW + SCHC	0.6857	0.0024	0.0073
DOANet + CS0145(A) + CS0145(E)	0.7344	0.0031	0.0125
DOANet + CS0145(A) + CS2367(E)	0.7382	**0.0038**	0.0135
DOANet + CS0145(A) + ACU(E)	0.7285	0.0035	0.0132
DOANet + CS2367(A) + CS0145(E)	0.8139	0.0031	0.0135
DOANet + CS2367(A) + CS2367(E)	**0.8160**	0.0028	**0.0139**
DOANet + CS2367(A) + ACU(E)	0.8101	0.0031	0.0132
DOANet + ACU(A) + CS0145(E)	0.7976	0.0014	0.0083
DOANet + ACU(A) + CS2367(E)	0.8000	0.0017	0.0087
DOANet + ACU(A) + ACU(E)	0.7944	0.0014	0.0076

6.2.3 Great-circle angular distance

Table 9 shows that in terms of the great-circle angular distance, DOANet using microphone channels 2, 3, 6, and 7 performed better than both baseline and modified baseline systems. From Fig. 13, we can see that for angular deviations less than 20, the performance of DOANet is very similar to the modified baseline system. It should be mentioned that for all angular spectrum techniques available in the baseline system, all angular distances were greater than 40°.

6.3 Summary

DOANet is seen to outperform both the baseline and modified baseline techniques while comparing the AUC values. However, for the fractional AUC values, AUC(< 10) and AUC(< 20), DOANet falls behind the modified baseline techniques in most cases. To explore the results further, we performed statistical significance tests (p value of two-sample t test at 0.05 significance level) using the deviation of predicted azimuth, elevation, and the great-circle angular distance values from ground truth. The p values obtained when comparing the technique with the highest AUC (DOANet or modified baseline) with the best baseline method are summarized in Table 10. A p value less than 0.05 indicates that the technique with higher AUC value is indeed better, whereas a p value greater than or equal to 0.05 indicates the higher AUC value has no statistical significance.

To illustrate how the p values were calculated in Table 10, let us consider the task of static azimuth angle deviation (first row). Comparing AUC, DOANet + CS2367 had the overall highest AUC value and DS had the highest AUC among baseline techniques (see Table 4). So we conducted statistical tests between DOANet + CS2367 and DS. Similarly, the pairs compared for AUC(< 10) and AUC(< 20) were GCC-NONLIN + SCHC vs. GCC-NONLIN and DOANet + CS2367 vs. GCC-NONLIN, respectively. When comparing results for AUC(< 10) and AUC(< 20), we did not consider all the data points; out of the total 120 static test data points, we included only those data points where the angular deviation was less than 10 and 20° for AUC(< 10) and AUC(< 20), respectively.

Observing Table 10, for static tasks, DOANet was always statistically better compared to its best baseline counterpart wherever it had the highest AUC and fractional AUC values. However, techniques involving SCHC were not always statistically better despite having higher AUC values (p value was greater than 0.05). From this, we can conclude that DOANet provides a statistically significant improvement over baseline methods for static tasks.

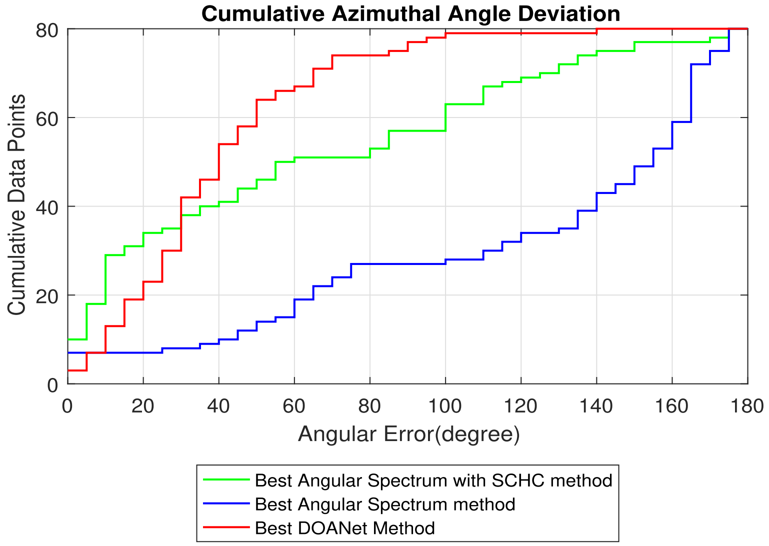

Fig. 11 AUCs for cumulative in-flight azimuth angle deviation

In in-flight tasks, both DOANet and techniques with SCHC were statistically better than the best baseline methods. To analyze the results further, we performed statistical tests between the best DOANet schemes and SCHC techniques for the in-flight cases. The *p* values obtained are provided in Table 11.

By looking at Tables 10 and 11 together, we can see that whenever DOANet had the higher AUC values, the difference was always statistically significant (*p* value less than 0.05). Conversely, when SCHC techniques had higher AUC values than DOANet, the difference was never statistically significant, with the *p* values being much larger than 0.05. From this, we conclude that, in most of the cases, our proposed model does indeed provide an improvement over baseline methods; at worst, it is never statistically worse than modified baseline methods, and always better than the original baseline methods.

7 Conclusion

In this article, we explored the challenge of sound source localization (SSL) from UAVs in the context of detecting human speech sounds for search and rescue operations. We proposed an end-to-end one-dimensional dilated convolutional neural network called DOANet for tackling this

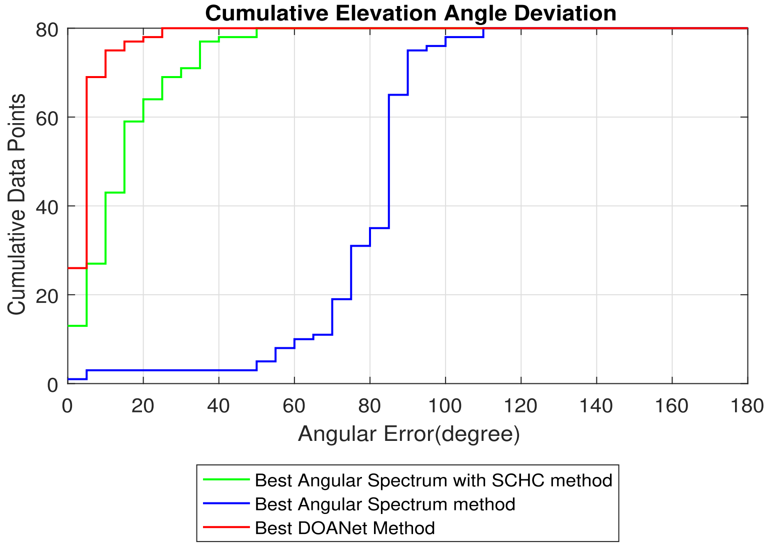

Fig. 12 AUCs for cumulative in-flight elevation angle deviation

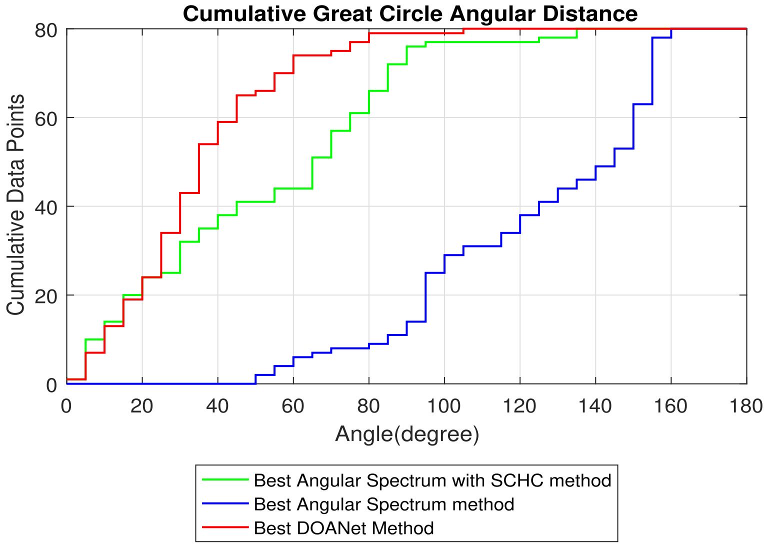

Fig. 13 AUCs for cumulative in-flight great-circle angular distance

challenge. To train our network, we used the DREGON dataset along with a synthetic dataset that we generated using computer simulation. We compared our system with a baseline that utilized traditional angular spectrum methods for SSL. We also augmented the baseline system with an algorithm for reducing the ego-noise of the UAV which utilized the UAV's rotor speed information and compared the modified system with our proposed DOANet. The results we obtained demonstrated that DOANet was able to achieve a statistically significant improvement over the baseline methods in most of the metrics considered and at worst was still statistically comparable to the modified baseline methods. Our proposed model was able to achieve this result directly from raw audio input without needing any prior filtering of ego-noise or hand-crafted techniques. We believe this makes our method more flexible—in that it can be improved simply by training it with more real data collected from practical outdoor scenarios. We also observed that while our model was more accurate overall, it scored lower in terms of fractional AUC values—AUC($<$ 10) and AUC($<$ 20)—compared to the modified baseline methods. This indicates our model is less accurate at fine grain resolution of the elevation and azimuthal angles. In practical search and rescue scenarios, the UAV would need to "home in" on the target sound source. A combination of DOANet and the modified baseline methods may be used for better performance in such a case; DOANet would provide the initial rough direction of the sound, and the modified baseline methods would be used for finer

Table 10 Summary of the best techniques for different tasks along with *p* values of two-sample *t* test at 0.05 significance level when comparing against best baseline method. *AD* azimuthal angle deviation, *ED* elevation angle deviation, *GCAD* great-circle angular distance

Task	Metric	AUC	AUC(<10)	AUC(<20)
Static	AD	DOANet + CS2367	GCC-NONLIN + SCHC	DOANet + CS2367
		0.00	0.63	0.00
	ED	DOANet + CS2367	GCC-PHAT + SCHC	GCC-PHAT + SCHC
		0.01	0.87	0.01
	GCAD	DOANet +CS2367(A) + CS2367(E)	GCC-NONLIN + SCHC	GCC-NONLIN + SCHC
		0.00	0.26	0.00
In-flight	AD	DOANet + CS2367	DS + SCHC	DS + SCHC
		0.00	0.01	0.00
	ED	DOANet + CS0145	DOANet + CS0145	DOANet + CS0145
		0.00	0.00	0.00
	GCAD	DOANet + CS2367(A) + CS2367(E)	DS + SCHC	DS + SCHC
		0.00	0.00	0.00

Table 11 Computed *p* values of two-sample *t* test at 0.05 significance level when comparing the best DOANet scheme against the best SCHC technique for in-flight tasks. *GC* great circle

Metric	*p* values of two-sample *t* test ($\alpha = 0.05$)		
	AUC	AUC(<10)	AUC(<20)
Azimuthal deviation	0.01	0.78	0.80
Elevation deviation	0.00	0.00	0.00
GC angular distance	0.00	0.47	0.80

estimation once the UAV is closer to the target. We hope to expand the scope of our work to include tracking the dynamic performance of DOANet in real time to see if it is able to gradually lead the UAV to the actual source of the sound as well as collect more data from outdoor environments to improve DOANet further.

Acknowledgements
We would like to thank the organizing team of *IEEE Signal Processing Cup 2019: Search & Rescue with Drone-Embedded Sound Source Localization* for creating a competitive platform and inspiring to work on the problem.

Authors' contributions
ABAQ conducted the research to develop and prepare the results on the DOANet approach. KMNH worked with ABAQ for generating synthetic data and developing the method. MFS has worked on preparing the graphical representations of the results of DOANet approach. SAI provided his hardware support to train the models and prepare the results. AA has conducted the research on angular spectrum with SCHC methods. MMR and MTI also worked with her to develop the method and generate the results. MAH and SH supervised the whole work. All the authors have contributed to write the manuscript. ABAQ, KMNH, and SH finally coordinated and revised the manuscript. All authors read and approved the manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 4 January 2020 Accepted: 14 October 2020
Published online: 05 November 2020

References

1. D. Gilman, M. Easton, Unmanned aerial vehicles in humanitarian response. U. N. Off. Coord. Humanitarian Aff. https://www.unocha.org/fr/publication/policy-briefs-studies/unmanned-aerial-vehicles-humanitarian-response. Accessed 22 June 2014
2. G. Sharma, Armed with drones, aid workers seek faster response to earthquakes, floods. Reuters. Accessed 15 May 2016
3. M. Basiri, F. Schill, P. U. Lima, D. Floreano, in *IEEE International Conference on Intelligent Robots and Systems*. Robust acoustic source localization of emergency signals from Micro Air Vehicles (Institute of Electrical and Electronics Engineers (IEEE), Vilamoura, 2012), pp. 4737–4742. https://doi.org/10.1109/IROS.2012.6385608
4. T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, K. Nakadai, in *IEEE International Conference on Intelligent Robots and Systems*. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array (Institute of Electrical and Electronics Engineers(IEEE), Chicago, Illinois, 2014), pp. 1902–1907. https://doi.org/10.1109/IROS.2014.6942813
5. K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, H. G. Okuno, Design of UAV-embedded microphone array system for sound source localization in outdoor environments. Sensors (Switzerland) (2017). https://doi.org/10.3390/s17112535
6. L. Wang, R. Sanchez-Matilla, A. Cavallaro, in *IEEE International Conference on Intelligent Robots and Systems*. Tracking a moving sound source from a multi-rotor drone (Institute of Electrical and Electronics Engineers (IEEE), Madrid, 2018), pp. 2511–2516. https://doi.org/10.1109/IROS.2018.8594483
7. M. Strauss, P. Mordel, V. Miguet, A. Deleforge, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)*. DREGON: dataset and methods for UAV-embedded sound source localization (IEEE, Madrid, Spain, 2018), pp. 5735–5742. https://doi.org/10.1109/IROS.2018.8593581. https://hal.inria.fr/hal-01854878
8. K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, H. G. Okuno, in *IEEE International Conference on Intelligent Robots and Systems*. Noise correlation matrix estimation for improving sound source localization by multirotor UAV (Institute of Electrical and Electronics Engineers (IEEE), Tokyo, 2013), pp. 3943–3948. https://doi.org/10.1109/IROS.2013.6696920
9. A. Schmidt, A. Deleforge, W. Kellermann, in *IEEE International Conference on Intelligent Robots and Systems*. Ego-noise reduction using a motor data-guided multichannel dictionary (Institute of Electrical and Electronics Engineers (IEEE), Daejeon, 2016), pp. 1281–1286. https://doi.org/10.1109/IROS.2016.7759212
10. L. Wang, A. Cavallaro, Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles. IEEE Sensors J. **17**(8), 2447–2455 (2017). https://doi.org/10.1109/jsen.2017.2669262
11. P. Marmaroli, X. Falourd, H. Lissek, in *Acoustics 2012*. A UAV motor denoising technique to improve localization of surrounding noisy aircrafts: proof of concept for anti-collision systems, (Nantes, 2012). https://hal.archives-ouvertes.fr/hal-00811003
12. S. Yoon, S. Park, S. Yoo, in *2016 IEEE International Conference on Consumer Electronics, ICCE 2016*. Two-stage adaptive noise reduction system for broadcasting multicopters (Institute of Electrical and Electronics Engineers (IEEE), Las Vegas, 2016), pp. 219–222. https://doi.org/10.1109/ICCE.2016.7430588
13. T. Morito, O. Sugiyama, R. Kojima, K. Nakadai, in *IEEE International Conference on Intelligent Robots and Systems*. Partially shared deep neural network in sound source separation and identification using a uav-embedded microphone array (Institute of Electrical and Electronics Engineers (IEEE), Daejeon, 2016), pp. 1299–1304. https://doi.org/10.1109/IROS.2016.7759215
14. B. Yen, Y. Hioka, B. Mace, in *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings*. Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information (Institute of Electrical and Electronics Engineers (IEEE), Tokyo, 2018), pp. 1–5. https://doi.org/10.1109/IWAENC.2018.8521324
15. J. M. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards end-to-end acoustic localization using deep learning: from audio signals to source position coordinates. Sensors (Switzerland). **18**(10), 3418 (2018). https://doi.org/10.3390/s18103418
16. N. Yalta, K. Nakadai, T. Ogata, Sound source localization using deep learning models. J. Robot. Mechatron. **29**(1), 37-48 (2017). https://doi.org/10.20965/jrm.2017.p0037
17. F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2016)
18. M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, S. Mougiakakou, Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE J. Biomed. Health Inform. **23**, 714–722 (2019). https://doi.org/10.1109/jbhi.2018.2818620
19. S. Hossain, S. Najeeb, A. Shahriyar, Z. Abdullah, M. Haque, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A pipeline for lung tumor detection and segmentation from ct scans using dilated convolutional neural networks (Institute of Electrical and Electronics Engineers (IEEE), Brighton, 2019), pp. 1348–1352. https://doi.org/10.1109/ICASSP.2019.8683802
20. A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, L. Marcenaro, Audio-based search and rescue with a drone: highlights from the ieee signal processing cup 2019 student competition [sp competitions]. IEEE Signal Proc. Mag. **36**(5), 138–144 (2019). https://doi.org/10.1109/msp.2019.2924687
21. C. Blandin, A. Ozerov, E. Vincent, Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Sig. Process. **92**(8), 1950–1960 (2012). https://doi.org/10.1016/j.sigpro.2011.09.032
22. J. Capon, High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE. **57**(8), 1408–1418 (1969). https://doi.org/10.1109/IWAENC.2018.8521324
23. M. S. Bartlett, Smoothing periodograms from time-series with continuous spectra. Nature. **161**(4096), 686–687 (1948)

24. H. Krim, M. Viberg, Two decades of array signal processing research: the parametric approach. IEEE Signal Proc. Mag. **13**, 67–94 (1996)
25. A.vd. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
26. S. Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. Van Den Oord, O. Vinyals, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Temporal modeling using dilated convolution and gating for voice-activity-detection (Institute of Electrical and Electronics Engineers (IEEE), Calgary, 2018), pp. 5549–5553. https://doi.org/10.1109/ICASSP.2018.8461921
27. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
28. S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, in *Advances in Neural Information Processing Systems*, ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. How does batch normalization help optimization? (Curran Associates, Inc., 2018), pp. 2483–2493
29. IEEE Signal Processing Cup 2019. http://dregon.inria.fr/datasets/signal-processing-cup-2019. Accessed 22 Oct 2020
30. J. S. Garofolo, Timit acoustic phonetic continuous speech corpus. Web Download. Linguist. Data Consortium, 1993 (1993)
31. R. Scheibler, E. Bezzam, I. Dokmanic, *Pyroomacoustics: a python package for audio room simulation and array processing algorithms* (Institute of Electrical and Electronics Engineers (IEEE), Calgary, 2018). https://doi.org/10.1109/icassp.2018.8461310
32. L. Griffiths, C. Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 27–34 (1982). https://doi.org/10.1109/TAP.1982.1142739
33. F. Chollet, *Deep Learning with Python*, 1st. (Manning Publications Co., New York, 2018)
34. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
35. Y. Bengio, J. Louradour, R. Collobert, J. Weston, in *Proceedings of the 26th Annual International Conference on Machine Learning ICML '09*. Curriculum learning (Association for Computing Machinery, New York, NY, USA, 2009), pp. 41–48. https://doi.org/10.1145/1553374.1553380. https://doi.org/10.1145/1553374.1553380
36. D. P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2017)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions