Yen and Hioka *EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:13
https://doi.org/10.1186/s13636-020-00181-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Noise power spectral density scaled SNR response estimation with restricted range search for sound source localisation using unmanned aerial vehicles

Benjamin Yen* and Yusuke Hioka

Abstract

A method to locate sound sources using an audio recording system mounted on an unmanned aerial vehicle (UAV) is proposed. The method introduces extension algorithms to apply on top of a baseline approach, which performs localisation by estimating the peak signal-to-noise ratio (SNR) response in the time-frequency and angular spectra with the time difference of arrival information. The proposed extensions include a noise reduction and a post-processing algorithm to address the challenges in a UAV setting. The noise reduction algorithm reduces influences of UAV rotor noise on localisation performance, by scaling the SNR response using power spectral density of the UAV rotor noise, estimated using a denoising autoencoder. For the source tracking problem, an angular spectral range restricted peak search and link post-processing algorithm is also proposed to filter out incorrect location estimates along the localisation path. Experimental results show the proposed extensions yielded improvements in locating the target sound source correctly, with a 0.0064–0.175 decrease in mean haversine distance error across various UAV operating scenarios. The proposed method also shows a reduction in unexpected location estimations, with a 0.0037–0.185 decrease in the 0.75 quartile haversine distance error.

Keywords: Microphone array, Unmanned aerial vehicle, Rotor noise, Source localisation, Denoising autoencoder, Power spectral density, Restricted peak search

1 Introduction

Unmanned aerial vehicles (UAVs) have recently gained huge popularity over a wide range of applications, such as filming [2], search and rescue [3], or security and surveillance [4]. One of the significant advantages of UAVs is its flexibility in manoeuvrability, allowing ease of navigation through environments that are difficult or dangerous for human access. In the application of search and rescue, there are already several reports of successful rescue missions where victims were found stranded in environments that are difficult to navigate through [5–7]. The key to its success is the use of localisation technologies to track down the whereabouts of the stranded victims effectively. To this day, various sensing technologies were utilised for search and rescue purposes such as high-resolution cameras or thermal imaging. While such sensing technologies are well-proven and highly effective under many types of environments, information from sound is also one that should not be overlooked, for it is common to encounter scenarios where the environment renders visual information as unusable. For example, for a UAV hovering over a mountain range, where vegetation could hinder the visibility of a rescue target, the target could be detected and located by sound. With localisation being the key objective to perform the search and rescue task properly, it

*Correspondence: benjamin.yen@ieee.org
This study is based on a contribution to the 2019 IEEE Signal Processing Cup (SPCup): Audio-Based Search and Rescue with a Drone [1].
Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, 20 Symonds Street, Auckland, 1010, New Zealand

is vital that the utilised sensing technologies are effective under a wide range of environments [8], including adverse environments such as those where visual information is severely impaired. In turn, when one method is rendered unusable, others still remain effective. However, audio recording using UAVs has shown to be challenging due to the high noise levels radiated from the UAV rotors. This significantly affects the quality of the audio signals to aid not only with search and rescue, but also with any applications [9–12].

In recent years, numerous studies attempt to perform localisation of sound sources using UAVs. Many achieve this by utilising signal processing techniques that revolve around the usage of an array of microphones [13]. With the significant contamination of recordings caused by rotor noise being a problem, numerous studies attempt to eliminate the effects of rotor noise itself. Examples include denoising the input signals by forming a reference rotor noise profile based on its tonal components [14], or capturing the noise correlation matrix in a supervised manner [15]. Other approaches include spatial filtering of the rotor noise, such as the study carried out in [16], given that the rotor positions are fixed relative to the microphones. While rotor noise is nearly omnidirectional along the rotor plane, there are sweet-spots above or below the rotors where radiation could be less intensive. Authors from [17] exploit this by placing microphones above the UAV rotors and employ a spatial likelihood function based on the direction of the arrival of the target sound source. The study has shown promising results when the target sound is located in the direction where rotor noise radiation is least apparent. However, such an approach is only effective in locations where such conditions can be met.

Many studies also set to address the challenges via further developing existing localisation techniques. For example, authors in [15, 18, 19] extended the multiple signal classification (MUSIC) method [13], namely modifying the noise correlation matrix to combat the challenges encountered with the high levels of rotor noise. However, these were carried out under a fixed UAV with a fixed target sound source position. Works from [20] carried the extended MUSIC approach for a flying UAV. However, the target sound source was limited to whistle sounds, which would be unrealistic in many practical scenarios. Approaches based on the steered response power with phase transform (SRP-PHAT) were used by [21] with Doppler shift for a fixed-wing UAV. This was also extended in [22] by detecting and localising chirp signals emitted from nearby UAVs to avoid potential collisions between each other. However, in both studies, the target sound was limited to narrowband signals with a known frequency. Optimising microphone placements has also shown improvement in localisation performance [23–25]. However, the localisation performance starts to degrade when the movement of the UAV increases. Recent studies also showed approaches using convolutional neural networks (CNNs) for source localisation, such as [26]. A comprehensive list of related studies can be found in [17].

While most of the studies mentioned above were able to present improved accuracy and precision using their highly responsive algorithms, they usually require certain assumptions to be imposed. In particular, most of the studies mentioned above assume that the UAV rotor noise has good continuity in the time-frequency (T-F) spectrum in order to reduce the influence of UAV rotor noise effectively [15, 17–19]. An instance includes assuming the tonal components of the rotor noise do not vary in a highly random manner. While this assumption is valid for most cases, it depends highly on the placement of the microphone array. Often, however, the microphone array is restricted to be placed below the rotor plane, of which the noise becomes dominated by the flow generated from the propeller's thrust. This presents an additional layer of challenge to the already low signal-to-noise ratio (SNR) of the audio signals since flow noise is highly random and nonlinear, and thus, the correlation between time frames is less likely to hold. Coupled with the high responsiveness of the methods itself, it could potentially lead to highly unstable performance.

Authors in [27] developed the *multi-source time difference of arrival* (*TDOA*) *estimation in reverberant audio using angular spectra* framework that is more robust to the practical challenges addressed. This was the baseline method provided in the 2019 IEEE Signal Processing Cup (SPCup) [1]. The method aims to perform robust sound source localisation even in reverberant environments effectively. While the localisation response is not as precise as the studies mentioned prior, the performance is consistent and generally stable, even under moderate levels of reverberation. Although the study was not targeted for a UAV scenario, this aspect can be addressed by reducing the influences coming from the rotor noise, as demonstrated from the aforementioned existing studies. For example, all winning teams that participated in the finals of the SPCup utilised the method in [27] along with multichannel Wiener postfiltering for UAV noise reduction. In addition, various approaches were utilised to address the challenges faced in the operating UAV scenario [1]. For example, Team AGH utilised a Kalman filter to improve continuity of the estimated source location paths. Meanwhile, Team SHOUT COOEE! improved the continuity of the paths using a heuristic method inspired by the Viterbi algorithm. On the other hand, Team Idea!_SSU estimated the paths via a two-step procedure by first estimating a global source path, followed by a refined estimation using a restricted

angular search range around the global estimated direction [1].

While noise reduction using T-F masking such as Wiener postfiltering is not uncommon, it has shown to be useful across many acoustic scenarios. For instance, the study from [28] presented and compared various T-F masks which most showed localisation performance improvement, as well as methods based on CNNs [29]. T-F masks specific towards noise that has long time-dependence/continuation, a property present in UAV rotor noise, have also been studied, such as those from [30, 31]. A UAV-specific study for source enhancement using CNNs was carried out in [32]. Studies from [33] also showed that by accurately estimating the power spectral densities (PSDs) of the individual sound sources, source enhancement and rotor noise denoising could be effectively carried out via beamforming with Wiener postfiltering. Building on this idea, this study proposes a rotor noise reduction algorithm based on accurate estimation of the rotor noise PSD, which is incorporated into existing robust source localisation techniques. In addition, the study also proposes a post-processing algorithm to smooth the estimated source location paths. The proposed method sets to extend the baseline method from [27] for the UAV problem. As the method is developed for the participation of the SPCup, it is designed around the competition dataset, containing audio recordings corresponding to its microphone array mounted UAV system.

The rest of the paper is organised as follows. A description of the UAV, microphone array, and problem setup is given in Section 2, followed by details of the proposed method in Section 3. Experimental setup and parameters are described in Section 4, followed by the performance evaluation of the proposed method in Section 5. Finally, the paper is concluded with some remarks in Section 6.

2 UAV system and problem setup

As mentioned in Section 1, the baseline method from [27] was able to deliver consistent and stable localisation performance in a range of levels of reverberation, making it effective in practical scenarios. Hence, this study aims to extend the baseline method for the UAV problem. This section presents the problem setup, including the definition of the sound sources and input signal, before discussing constraints specific to the UAV setting.

An overview of the audio recording UAV including the microphone array setup used in this study is shown in Fig. 1. The problem assumes a UAV system with a M-sensor microphone array embedded, receiving a target sound source, L interfering *spatially coherent* noise sources (including those generated by U UAV rotors), and ambient *spatially incoherent* noise. The objective of the system is to accurately locate the target sound source using the M-channel noisy recordings. The short-time Fourier transform (STFT) of the microphone array's input signals is expressed in vector form as:

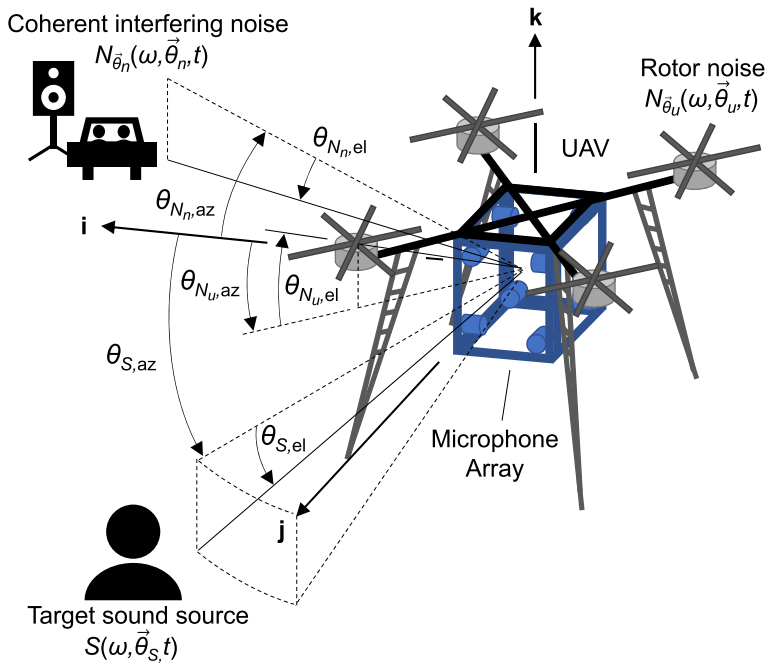

Fig. 1 Overview of problem setup for sound source localisation using UAV

$$\begin{aligned}\mathbf{x}(\omega,t) &:= \left[\, X_1(\omega,t),\ \cdots,\ X_M(\omega,t) \,\right]^T \\ &= \mathbf{a}(\omega,\vec{\theta}_S)S(\omega,\vec{\theta}_S,t) \\ &+ \sum_{u=1}^{U} \mathbf{a}(\omega,\vec{\theta}_{N_u})N(\omega,\vec{\theta}_{N_u},t) \\ &+ \sum_{n=U+1}^{L} \mathbf{a}(\omega,\vec{\theta}_{N_n})N(\omega,\vec{\theta}_{N_n},t) + \mathbf{v}(\omega,t), \end{aligned} \tag{1}$$

$$\mathbf{a}(\omega,\vec{\theta}) = \left[A_1(\omega,\vec{\theta}),\cdots,A_M(\omega,\vec{\theta})\right]^T, \tag{2}$$

$$\mathbf{v}(\omega,t) = [V_1(\omega,t),\cdots,V_M(\omega,t)]^T, \tag{3}$$

where ^{T} denotes the transpose, $X_m(\omega,t)$ is the STFT of the mth microphone's input signal, $\mathbf{a}_\theta(\omega)$ and $\mathbf{v}(\omega,t)$ are the vector of transfer functions between the source $\vec{\theta} = [\theta_{\text{el}},\theta_{\text{az}}]^T$ (where el and az indicate the elevation and azimuth directions, respectively) and each microphone m, and the incoherent noise vector observed by the microphone array, respectively. $S(\omega,\vec{\theta}_S,t)$, $N(\omega,\vec{\theta}_{N_u},t)$, and $N(\omega,\vec{\theta}_{N_n},t)$ are the STFT of the target sound source at angle $\vec{\theta}_S$, the noise source coming from the uth rotor at angle $\vec{\theta}_{N_u}$, and the nth *spatially coherent interfering* noise source at angle $\vec{\theta}_{N_n}$, respectively. ω and t denote the angular frequency (of F frequency bins) and the time frame index. $\vec{\theta}_S$, $\vec{\theta}_{N_u}$, and $\vec{\theta}_{N_n}$ are expressed as follows for the 3D problem in spherical coordinates:

$$\vec{\theta}_S = \left[\theta_{S,\text{el}},\theta_{S,\text{az}}\right]^T, \tag{4}$$

$$\vec{\theta}_{N_u} = \left[\theta_{N_u,\text{el}},\theta_{N_u,\text{az}}\right]^T, \tag{5}$$

$$\vec{\theta}_{N_n} = \left[\theta_{N_n,\text{el}},\theta_{N_n,\text{az}}\right]^T. \tag{6}$$

Several assumptions are imposed on the setup. Given the difference in characteristics between the sound sources, the problem assumes the target sound source and rotor noise sources to be mutually uncorrelated. For the source localisation task, the main objective is to identify the directions of the target sound source. This usually requires knowing the transfer function of the audio sources with respect to the microphone array, in order to capture the true characteristics of $\mathbf{a}(\omega,\vec{\theta})$ correctly. This includes knowing the acoustical characteristics of the environment (i.e. impulse response). Unfortunately, such information is generally unavailable. As such, we impose an assumption that the UAV is operating at some height above ground, regardless of the environment beneath, and is thus mostly open air. Therefore, the environment is approximately of a free field, and that $\mathbf{a}(\omega,\vec{\theta})$ is assumed as the steering vector of a plane wave [33], described as:

$$\mathbf{a}(\omega,\vec{\theta}) = \left[e^{-j\omega\tau_{\vec{\theta},1}},\cdots,e^{-j\omega\tau_{\vec{\theta},M}}\right]^T, \tag{7}$$

where $\tau_{\vec{\theta},m}$ is the time difference of arrival (TDOA) at the mth microphone with respect to the reference microphone typically placed at the origin of the coordinate. It should be noted that this assumption is merely made for modelling the transfer function between the microphones and the sound source. In practice, such as that from the database provided by the SPCup (see Section 2), some level of reverberation is expected.

The problem, as setup by the SPCup requirements, assumes three distinct tasks for the UAV and the target sound source:

1. Hovering UAV—In this scenario, the target sound source and UAV are assumed as fixed in position throughout the audio recording.
2. Flying (i.e. moving) UAV, broadband sound source—In this scenario, the target sound source is assumed fixed. However, the UAV is assumed to be moving relative to the target sound source. The target source is a continuous broadband signal.
3. Flying (i.e. moving) UAV, speech sound source—Like task 2, the target sound source is assumed fixed, and the UAV is assumed to be moving relative to the target sound source. The target source is speech.

For tasks 2 and 3, the UAV is assumed to be moving gradually, such that there are no erratic variations in the tonal components in the rotor noise build-up. In addition, due to the dataset used for the study (see Section 4.1) only containing the target sound source and UAV rotor noise, it is assumed that no additional coherent interfering noise sources exist (i.e. $L = U$), such that the adversity of the environment for the audio recording UAV is only based on the levels of the UAV rotor noise relative to the target source. Finally, the problem is limited to overdetermined cases, where $M \geq L + 1$.

3 Proposed method

Figure 2 shows a block diagram of the proposed localisation method. The method follows the general structure of the SPCup baseline method (see Section 1), as the method gave decent results over a range of input noise conditions in a preliminary study. However, significant performance degradation was found under lower input SNR cases, where rotor noise begins to dominate the recorded signal. Naturally, like many other studies mentioned in Section 1, developing a means of reducing the effects of the UAV rotor noise would directly benefit in preventing false detection of the target sound source location.

Different to the methods discussed in Section 1, where noise reduction is generally carried out in the noise correlation matrix design, this study proposes a PSD-based weighting function to reduce the UAV rotor noise effects. Given the complex nature of rotor noise, which is dominated by the flow coming from the thrust of the

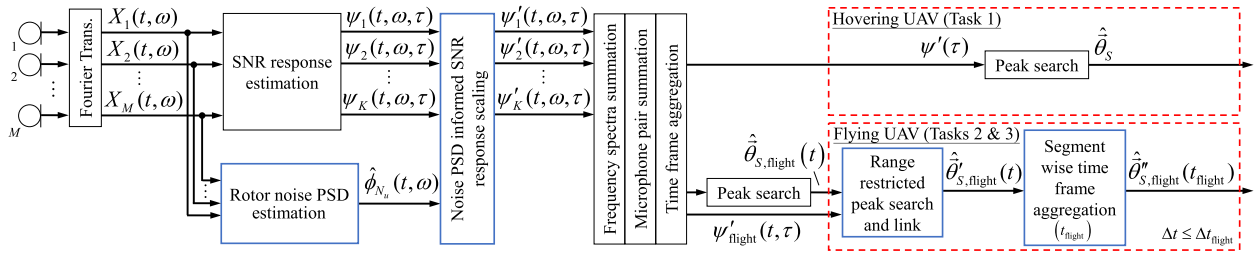

Fig. 2 Block diagram of the proposed method. Blue boxes indicate modifications to [27] introduced by the proposed method. Processes in the boxes with red dashed line are selected based on the scenario (i.e. hovering or flying UAV)

UAV's propellers, a machine learning approach is proposed. The approach sets to estimate the rotor noise PSD using the PSDs of the microphone input signals, removing any irrelevant sources present (i.e. target sound source), before using this information to design a noise filter specifically removing UAV rotor noise.

In the case of a flying UAV (i.e. tasks 2 and 3), due to the constant change in location between the UAV and the target sound source, localisation has to be carried out in shorter time periods. This results in less input information available to accurately estimate the source direction, which also becomes a factor in performance degradation. However, given (as mentioned in Section 2) that the UAV is assumed to move gradually, estimated locations in each time period should not vary erratically. Therefore, a post-processing algorithm designed explicitly for tasks 2 and 3 is also proposed. The algorithm takes into account of the assumption as mentioned earlier and filters out location estimates deemed erratic, from which in-depth location search at these problematic estimates is carried out to improve continuity if the overall estimated location path.

This section first introduces the baseline method from [27] in Section 3.1, followed by the extensions and modifications made to the baseline method, as shown by the blue boxes in Fig. 2. These extensions are the UAV rotor noise PSD estimation algorithm used to reduce the rotor noise effects (see Section 3.2), and the post-processing algorithm (see Section 3.3) for tasks 2 and 3, respectively.

3.1 Multi-source TDOA estimation in reverberant audio using angular spectra

This section outlines the baseline method [27] that is utilised in this study. Although the method is capable of localising multiple sources, for this study, the problem is limited to the single target sound source (i.e. $N(\omega, \vec{\theta}_{N_n}, t)$ is not considered in this study). The method is similar to the SRP technique, where SNR is calculated in the angular (TDOA) and T-F spectrum using pairs of microphones within the array, giving $K = {}_M C_2$ unique spectrum. For this study, this will be referred as the *SNR response*. An overall SNR response in terms of $\vec{\theta}$ (i.e. an angular spectrum) is then obtained by aggregating the K individual SNR responses together. Details of the aggregation process are given later in this section. Many conventional localisation techniques such as generalised cross-correlation-phase transform (GCC-PHAT) [34], delay-and-sum (DS) [35], and minimum variance distortionless response (MVDR) [36] beamforming, or even MUSIC, can be utilised to calculate the SNR response. The study [27] also developed the diffuse noise model (DNM), a modified MVDR approach, which uses a noise model to improve robustness against ambient noise, assuming the noise is diffuse in nature.

Prior to calculation of the SNR response, a grid of TDOAs τ covering the relevant range of $\vec{\theta}$ in the elevation and azimuth plane (i.e. the angular spectra), where the target sound source is assumed to be located for each kth microphone pair, is established as follows:

$$\tau_k(\theta_{\text{el}}, \theta_{\text{az}}) = \frac{p_k \sin\left(\alpha_k\left(\theta_{\text{el}}, \theta_{\text{az}}\right)\right)}{c_0}, \tag{8}$$

$$\alpha_k(\theta_{\text{el}}, \theta_{\text{az}}) = \cos^{-1}\left(\frac{\mathbf{d}_k\left(\theta_{\text{el}}, \theta_{\text{az}}\right) \cdot \Delta\mathbf{p}_k}{p_k}\right), \tag{9}$$

where $\mathbf{d}_k$ is the directional vector associated with angle $\vec{\theta}$ and c_0 is the speed of sound. $\Delta\mathbf{p}_k$ is the separation between the kth pair of microphones in Cartesian coordinates, and p_k is the magnitude of the separating distance. This is used to map the TDOAs coming from the angular range of interest τ towards their respective angles $\vec{\theta}$ (i.e. the basis of the angular spectra).

The baseline method from [27] provides several localisation techniques to calculate the SNR response for localisation. For instance, the SNR response for DS [37] and MVDR beamforming is calculated as:

$$\psi_k^{\text{DS}}(\tau_k) = \frac{\mathbf{a}^H(\tau_k)\hat{\mathbf{R}}_{\mathbf{xx},k}\mathbf{a}(\tau_k)}{2\text{tr}\left(\hat{\mathbf{R}}_{\mathbf{xx},k}\right) - \mathbf{a}^H(\tau_k)\hat{\mathbf{R}}_{\mathbf{xx},k}\mathbf{a}(\tau_k)}, \tag{10}$$

$$\psi_k^{\text{MVDR}}(\tau_k) = \frac{\left(\mathbf{a}^H(\tau_k)\hat{\mathbf{R}}_{\mathbf{xx},k}^{-1}\mathbf{a}(\tau_k)\right)^{-1}}{\frac{1}{2}\text{tr}\left(\hat{\mathbf{R}}_{\mathbf{xx},k}\right) - \left(\mathbf{a}^H(\tau_k)\hat{\mathbf{R}}_{\mathbf{xx},k}^{-1}\mathbf{a}(\tau_k)\right)^{-1}}, \tag{11}$$

respectively, where $\hat{\mathbf{R}}_{\mathbf{xx},k}(t,\omega)$ is the *empirical covariance matrix* [27] of the input signals in all T-F bins from the kth microphone pair, and $\hat{\cdot}$ denotes an estimate.

On the other hand, the SNR response for the GCC-PHAT approach is calculated as:

$$\psi_k^{\text{GCC}}(t,\omega,\tau_k) = \mathbb{R}\left(\frac{\hat{R}_{12,k}(t,\omega)}{\left|\hat{R}_{12,k}(t,\omega)\right|}e^{-i\omega\tau_k}\right), \tag{12}$$

where $\hat{R}_{12,k}(t,\omega)$ is the cross-correlation between microphone input channels 1 and 2 from the kth microphone pair, and $\mathbb{R}(\cdot)$ denotes the real part of a complex number.

Note that t and ω in $\hat{\mathbf{R}}_{\mathbf{xx},k}(t,\omega)$ of (10) and (11) are omitted for brevity. In addition, $\theta_{S,\text{el}}$ and $\theta_{S,\text{az}}$ of τ are also omitted for brevity in (10)–(12) as well as the rest of this paper unless otherwise specified. A nonlinear extension of GCC-PHAT (GCC-NONLIN) proposed in [38], as well as DNM, is also provided by the SPCup baseline, and their respective SNR response calculation ($\psi_k^{\text{GCC-NONLIN}}$ and ψ_k^{DNM}) can be found in [27].

Following the calculation of $\psi_k(t,\omega,\tau_k)$, the SNR responses are aggregated together across the frequency bins, time frames, and microphone pairs, to deliver the overall angular spectrum. Subsequently, the peak response is identified as the sound source location. Aggregation across the frequency bins and the microphone pairs is carried out via summing while time frames can be summed or taken the maximum as shown respectively in (13) and (14):

$$\psi'^{\text{sum}}(\tau) = \sum_{t=1}^{T_{\text{hover}}}\sum_{k=1}^{K}\sum_{\omega=1}^{F}\psi_k(t,\omega,\tau_k), \tag{13}$$

$$\psi'^{\text{max}}(\tau) = \max_{t}^{T_{\text{hover}}}\sum_{k=1}^{K}\sum_{\omega=1}^{F}\psi_k(t,\omega,\tau_k). \tag{14}$$

In task 1 (i.e. hovering UAV), the relative location between the microphone array and the target sound source remains fixed. Therefore, all T_{hover} time frames are aggregated to give a single location estimate. For tasks 2 and 3 (i.e. flying UAV), aggregation cannot be carried out across all time frames and is thus instead carried out in segments of the input audio. This results in a smaller group of time frames T_{flight} used for localising the target sound source during each audio segment. These are calculated as:

$$\psi'^{\text{sum}}_{\text{flight}}(t,\tau) = \sum_{t=1}^{T_{\text{flight}}}\sum_{k=1}^{K}\sum_{\omega=1}^{F}\psi_k(t,\omega,\tau_k), \tag{15}$$

$$\psi'^{\text{max}}_{\text{flight}}(t,\tau) = \max_{t=1}^{T_{\text{flight}}}\sum_{k=1}^{K}\sum_{\omega=1}^{F}\psi_k(t,\omega,\tau_k). \tag{16}$$

The estimated target sound source TDOA $\hat{\tau}_S$ corresponds to the TDOA τ that gives the maximum overall SNR response from $\psi'(\tau)$. These are obtained as:

$$\hat{\tau}_S = \arg\max_{\tau}\left(\psi'(\tau)\right), \tag{17}$$

$$\hat{\tau}_{S,\text{flight}}(t) = \arg\max_{\tau}\left(\psi'_{\text{flight}}(t,\tau)\right). \tag{18}$$

As mentioned earlier, the grid or spectra of TDOAs τ directly map towards the angular spectra $\vec{\theta}$. This mapping relationship does not change even after the aggregation process. Therefore, the source location in terms of angle for tasks 1, 2 ($\hat{\vec{\theta}}_S$), and 3 ($\hat{\vec{\theta}}_{S,\text{flight}}(t_{\text{flight}})$) is obtained using the angular spectra derived from (8) and (9).

3.2 Noise PSD informed SNR response scaling

This section introduces the UAV rotor noise PSD-based weighting envelope to scale and denoise the SNR response $\psi(t,\omega,\tau)$. We refer to this process as *SNR response scaling*.

Given the relatively structured and time-continuous nature of UAV rotor noise PSDs, conventional neural network (NN) architectures such as multilayer perception, or other supervised NNs could be an adequate mapping function to model the noise PSDs under different input conditions [10], given sufficient training data is provided. However, in this study, the rotor noise data available for training is limited (see Section 4.2), and therefore, conventional NNs would not suit this condition.

On the other hand, denoising autoencoders (DAEs) learn a compressed representation of the uncorrupted input, rather than a full mapping of the training data in an unsupervised manner, and thus can be used for feature extraction and denoising [39]. This could help relax the requirement for a large number of hard-coded labels and simply let the DAE act as a denoising tool. Therefore, we propose a DAE to produce the required PSD data for the localisation task. DAE is an extension to the classical AE, where it attempts to clean the noisy input such that only the target output signal remains [39]. Since the objective of this algorithm is to create a PSD-based envelope to scale and denoise the SNR response $\psi(t,\omega,\tau)$, the target output signal of the DAE is the rotor noise PSD $\phi_{N_u}(\omega)$ with the inputs being the PSDs $\phi_m(t,\omega)$ from the microphone recordings. Therefore, different from the conventional use of a DAE, this process achieves a "de-targeting" effect, through recognising the target sound source as the equivalent "noise corruption" to remove.

The input audio PSD $\phi_m(t,\omega)$ is calculated by using the Welch method [40] given by $\phi_{\mathcal{X}}(t,\omega) = \lambda\phi_{\mathcal{X}}(t-1,\omega) + (1-\lambda)|\mathcal{X}(t,\omega)|^2$, where λ is the forgetting factor, and $\mathcal{X}$ represents an arbitrary signal. This is achieved by first feeding the input audio PSDs $\phi_m(t,\omega)$ to map towards the hidden representations z, forming the encoder component of the DAE. Subsequently, the rotor noise PSD

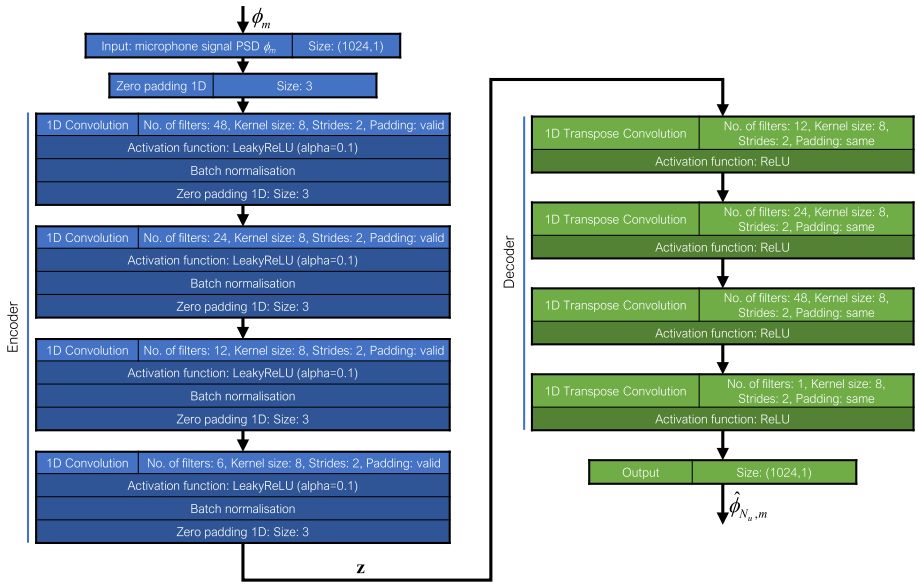

Fig. 3 DAE architecture of the SNR response scaling algorithm

$\hat{\phi}_{N_u}(\omega)$ is reconstructed from z, which forms the decoder component of the DAE.

Since the task is to perform denoising of the original input audio PSD (i.e. "de-targeting" the target sound source), it is essentially a regression task. The size of the input audio PSD data is $T_{\mathrm{DAE}} \times F$, where $T_{\mathrm{DAE}} = 1$ corresponds to the number of PSD frames taken per observation. For the regression task, the decoder of the DAE uses the rectified linear units (ReLU) activation function. Given that the DAE consists of several layers overall, the encoder of the DAE uses the leaky ReLU (LeakyReLU) [41] activation function, as a means of preventing possibilities of vanishing gradients, which has found in this study to slightly reduce training loss over ReLU. The DAE architecture is shown in Fig. 3.

The DAE is optimised with respect to the mean square error (MSE) between the output PSD $\hat{\phi}_{N_u}(t,\omega)$ and the true rotor noise PSD $\phi_{N_u}(t,\omega)$. To optimise MSE loss, the Adam optimiser is used [42]. The DAE is trained for each m microphone channels, giving a total of M DAEs for producing the SNR response scaling weighting envelope. However, since the task is to perform localisation, there is no requirement to achieve pinpoint accuracy in the PSD estimation for each k microphone pair, which is usually required for, for example, source enhancement [10, 33]. Furthermore, given the microphones used in this study are of identical build and omnidirectional, it is assumed that the estimated PSDs would not change drastically across microphones. Therefore, the estimated PSD with the most prominent amplitude response out of the M microphones for each frequency bin ω is selected and applied to scale the SNR responses for all K microphone pairs, with the prospect of maximising effectiveness in noise removal. In addition, the estimated PSD frames are grouped and averaged to match the time frames for the localisation process (see Table 1).

Finally, the rotor noise PSD scaled SNR response is obtained as:

$$\psi'_k(t,\omega,\tau) = \frac{\psi_k(t,\omega,\tau)}{\hat{\phi}_{N_u}(t,\omega)}. \tag{19}$$

After scaling the SNR response with the UAV rotor noise PSD weighting envelope, to obtain the final angular spectrum of the sound source, the aggregation process previously mentioned in Section 3.1 ((13)–(16)) is applied on $\psi'_k(t,\omega,\tau)$, before obtaining $\hat{\tau}_S$ leading to $\hat{\hat{\theta}}_S$ using (8) and (9).

Table 1 Experimental problem setup specifications

UAV scenario	**Hovering**	**Flying**
No. of recordings	**300**	**36**
Duration (s)	**∼ 3**	**4**
Sampling rate (kHz)	44.1	44.1
Target sound types	Speech	Speech and broadband
STFT time frames T	∼ 128	45 (0.0833 s intervals)
STFT time frames post (13)–(16)	1	15 (0.25 s intervals)

For the flying UAV scenario (i.e. tasks 2 and 3), azimuth and elevation angles of the source are taken as a mean value within a 500-ms window centred on each of its given time-stamps

3.3 Angular spectral range restricted peak search and link

As discussed in Section 3.1, for tasks 2 and 3, the shorter audio signal length for each location estimate means time frame aggregation is carried out in smaller groups of frames T_{flight}, which potentially causes a loss in angular spectral resolution. In addition, higher speed variations in the individual UAV rotors would also increase the complexity of the PSD for the DAE to estimate, potentially leading to further performance degradation. This section introduces an angular spectral range restricted peak search and link post-processing algorithm, for which we refer to as the *restricted peak search and link* (*RPSL*). The algorithm is applied towards the localisation output $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ before time frame aggregation is carried out (see Fig. 2), as a mean to compensate this problem.

The flowchart describing the algorithm is shown in Fig. 4. The algorithm makes use of several iterations of SNR response peak searching in the angular spectrum to obtain the correct sound source travel path, which generally follows these main processes:

1. Using localisation output $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ as the reference path of locations, for each time frame t, check the degree of separation $\Delta\hat{\tilde{\theta}}'_{S,\text{flight}}$ between the corresponding location with respect to the location of the preceding and succeeding time frames $t \pm 1$.
2. Perform restricted peak search using (18) with the SNR response $\psi'_{\text{flight}}(t,\tau)$ (see Fig. 2) and $\vec{\theta}_{\text{res}}(c,t_c)$ (see Fig. 4) around time frames giving unexpected locations (i.e. exceeding the nominal degree of separation $\Delta\vec{\theta}_{\text{res}}$), and obtaining the correct locations.
3. The above steps are repeated until valid locations can no longer be found, or if the start/end of the localisation path has been reached (i.e. $t_c \pm 1 \notin [t_{\text{start}}, t_{\text{end}}]$). This forms a "chain" of locations, or a local path (denoted as the cth chain in Fig. 4), to later to be compared against when forming the final global path of locations.
4. After obtaining all C chains of local paths, a final path of locations $\hat{\tilde{\theta}}'_{S,\text{flight}}(t)$ is formed by finding locations that appear most frequently amongst the C chains at the given time frame. Ideally, this would improve the consistency and smoothness compared to the original $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$.

Finally, the T_{flight} time frames in $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ are aggregated together to obtain $\hat{\tilde{\theta}}''_{S,\text{flight}}(t_{\text{flight}})$ (see Fig. 2). Details of this process are discussed later in Section 4.1.

The selection of the search range parameter $\Delta\hat{\tilde{\theta}}_{\text{res}}$ is heuristically tuned based on whether the estimated path of locations was the most sensible overall (i.e. no aggressive jumps or unnatural changes in direction).

To enable online-processing capabilities to the RPSL post-processing algorithm, this process is carried out in batches of frames of $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ that corresponds to 2 second blocks of audio, with the exception of the last batch, which would depend on the number of frames remaining. Such an approach is not uncommon, where online-processing is done via blocks of time frames, rather than individual frames alone [43].

Restricting the angular range for peak search reduces the risk of picking up disturbances with SNR angular spectral response more prominent than that of the target sound source (since they are excluded from the restricted search range). However, this assumes the original localisation path $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ is correct in a reasonable portion of the time frames. Conceptually, the method presented here is somewhat similar to the two-step flight path approach from Team Idea!_SSU [1].

Measures are also developed if a particular local path fails to find a peak with high enough SNR response to link towards. For example, the algorithm skips time frames (and proceeds to the next) where the restricted peak search fails to obtain a valid location until one with a valid location is found. Following this, the skipped locations in-between the two valid time frames are obtained via interpolation. Figure 5 shows an example of the improvement in localisation path with each stage (SNR response scaling and RPSL post-processing) of the proposed algorithm applied.

4 Experiments

As a team participating in the SPCup, the performance of the proposed method is evaluated against the competition dataset provided by the organiser of the SPCup [1]. Therefore, the proposed algorithm is tuned towards the UAV and microphone array system to develop the dataset. This section presents the details of the experimental setup of the given dataset, including the description of the UAV system, and various constraints found in the dataset. This is followed by an overview of the experimental parameters and additional information used for the proposed method, such as details of the training dataset for the UAV rotor noise PSD estimation process.

4.1 Experimental setup

The proposed method is evaluated using the DREGON database from [44], which makes use of the UAV system shown in Fig. 6. Details of the microphone array and rotor positions are shown in Fig. 7. The UAV is from MicroKopter®, utilising the 8 Sounds USB and Many Ears audio processing framework [45]. The UAV system utilises an array of 8 omnidirectional electret condenser microphones, located directly below the centre of the UAV, as shown in Fig. 7. All positions shown in Fig. 7 are in reference to the microphone array's baricenter.

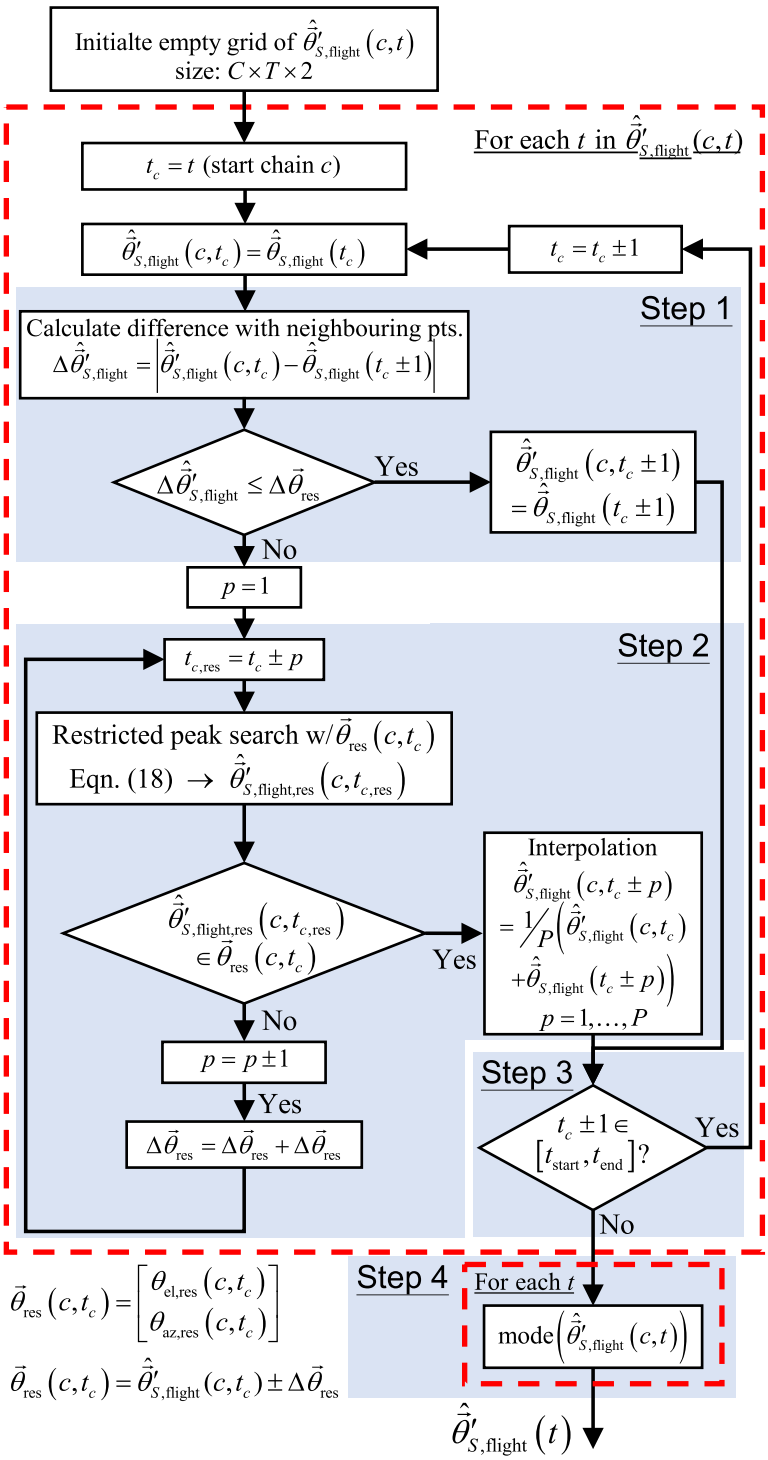

Fig. 4 Flowchart of the angular spectral range restricted peak search and link algorithm. The steps highlighted correspond to the steps described in Section 3.3

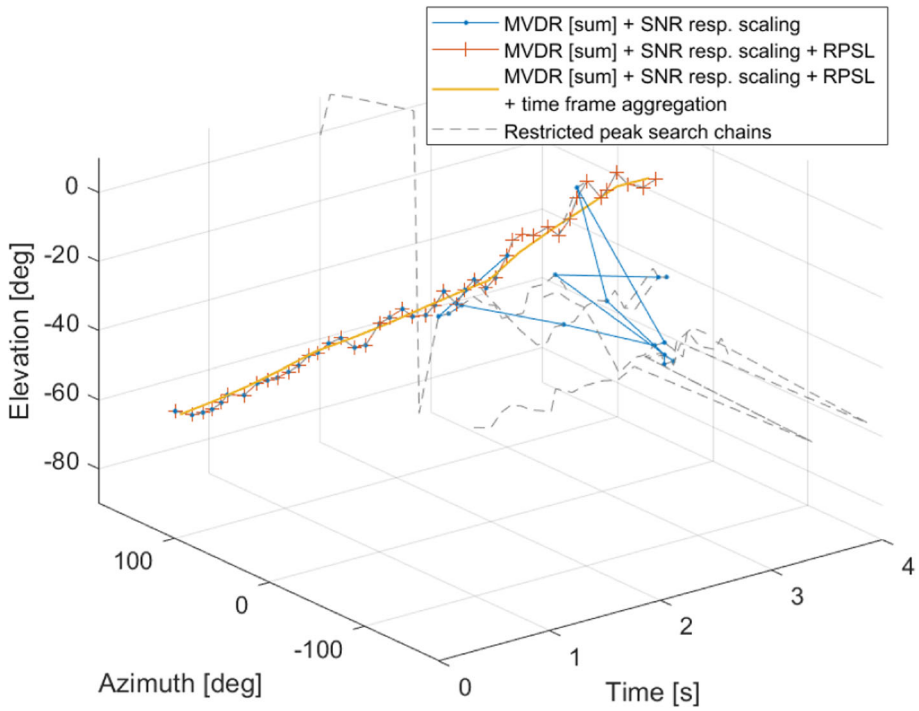

Fig. 5 Example of localisation performance with SNR response scaling (Section 3.2) and RPSL post-processing (Section 3.3) applied. (SPCup flying UAV broadband sound source (task 2) case 11 [1])

Table 1 shows the specifications of the evaluation dataset used for this study. As mentioned in Sections 2 and 3, the proposed method in this study is evaluated against the three tasks. Task 1 contains 300 individual cases; each consists of a $\sim$ 3-s microphone array recording for estimating a single location of the target sound source. On the other hand, tasks 2 and 3 consist of 20 cases with broadband sound source and 16 cases with speech sound source used as the target sound source. Each case in tasks 2 and 3 consists of 15 location points to be estimated in 0.25-s intervals along the duration of the 4-s microphone array recording. Therefore, the time frames from the STFT of the recordings are grouped into sections of 6 frames centering around each time-stamp

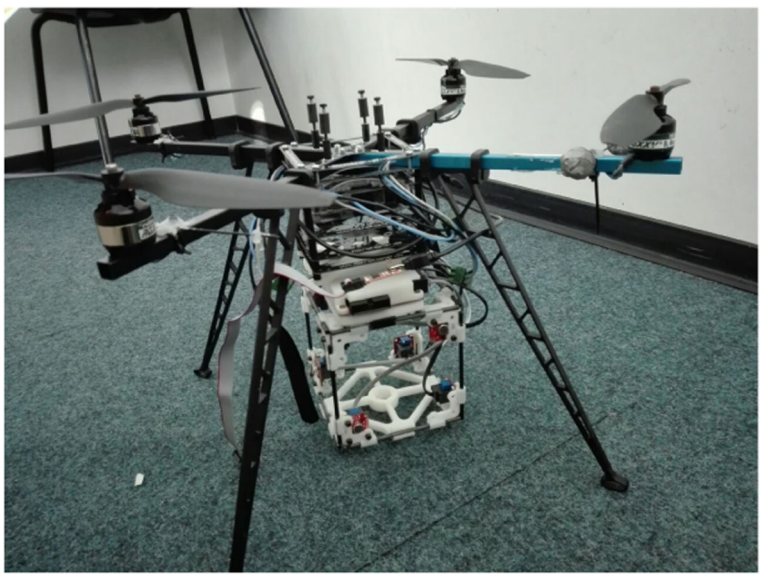

Fig. 6 Audio recording UAV overview. Image provided by SPCup syllabus [1]

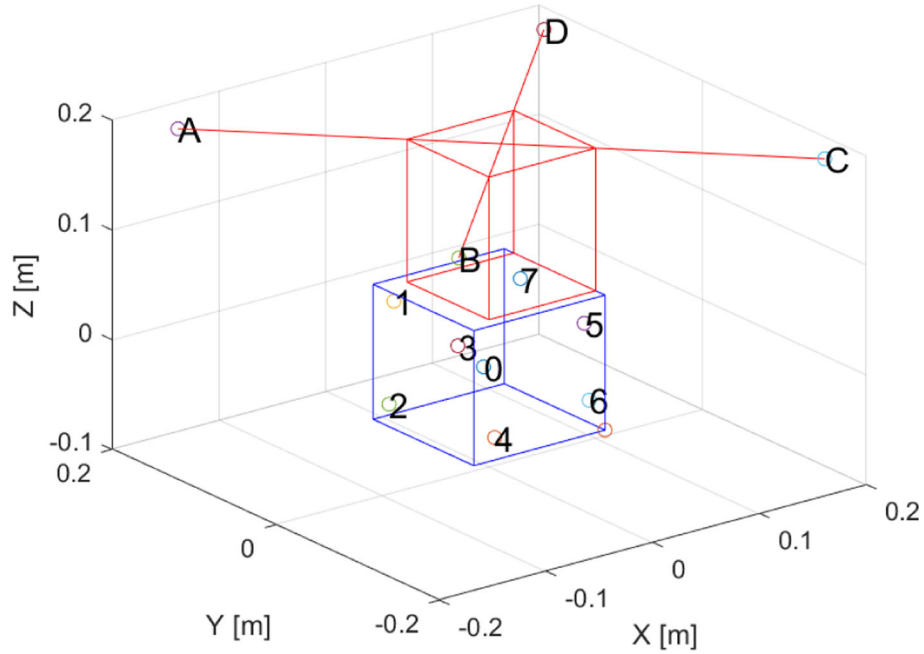

Fig. 7 Microphone array (0–7) and rotor (A–D) geometry overview. Figure provided by SPCup syllabus [1]

(i.e. $t_{\text{flight}} = [0.25\text{s}, 0.5\text{s}, ..., 3.75\text{s}]$), with aggregation performed on each of these groups of time frames. Further information regarding the evaluation dataset can be found in Table 1.

4.2 Noise PSD informed SNR response scaling: experimental parameters

To obtain the training and validation dataset required for the DAE from the SNR response scaling algorithm, labelled data containing rotor noise with and without the target source signal is required. Given the dataset is constrained to what was available in the DREGON database, data augmentation had to be performed in order to obtain a sufficient amount of data for training and validation. Initially, individual rotor recordings from the *development dataset* were used to provide the rotor noise data. However, due to the limited speed range coverage, extracts of rotor noise-only sections from the microphone recordings in the competition dataset were also used as part of the training process. This is achieved by manually editing out sections of the competition audio recordings containing traces of target sound source signals. The rotor noises are then mixed with a corpus of speech recordings to form the input observations, as part of the data augmentation process. For the corpus, the *REpeated HArvard Sentence Prompts* (*REHASP*) corpus [46] was used. The sentences were randomly selected with a balanced mixture of male and female speech. Since the individual cases contained in the competition dataset are normalised with respect to its signal amplitude, the rotor noise present in each case would have varying loudness depending on the input SNR. Therefore, one cannot simply train a mapping function by assuming rotor noise is consistent in power, meaning that the training dataset would contain repetitions of the rotor noise audio extracts with different amplitude scaling to compensate for this variation. For broadband sound source, since its acoustical characteristics are unknown, the only obtainable labels were from the *development dataset*. Thus, these recordings are mixed with the collected UAV rotor noise dataset and included for training.

Table 2 outlines the dataset specifications. To provide generalisation towards the DAE with the available data, the entire training dataset described in Table 2 is used to train a single DAE for each microphone. Given the lack of unique UAV rotor noise data, only 4% of the data from the training dataset described in Table 2 is used to obtain the validation dataset, to preserve as many training data as possible. The observations in the dataset are randomly shuffled prior to the split. For testing, the competition dataset described in Table 1 was used, giving a total of 45,338 observations.

It should be noted that the data constraint workaround described in this study was driven by the limited time and resources in the time of the competition. Ideally, rotor noise recordings would be obtained via independent noise recordings by using the exact UAV system per described in

Table 2 Specifications and parameters for the rotor noise PSD estimation DAE and RPSL post-processing algorithm

STFT length (overlap shift)	**2048 [46.4 ms] (1024 [23.2 ms])**
Forgetting factor λ	**0.3**
Dimension of each frame	**1 × 1024 (T_{DAE} × F)**
Training dataset	
Total no. of frames (from hovering UAV dataset)	417,180 [9687 s]
Total no. of frames (from flying UAV dataset, broadband)	4675 [108.6 s]
Total no. of frames (from flying UAV dataset, speech)	9435 [219.1 s]
Learning rate	5×10^{-5}
No. of epochs	2000
Testing dataset	
Total no. of frames (from hovering UAV dataset)	39,182 [920.1 s]
Total no. of frames (from flying UAV dataset, broadband)	3420 [80 s]
Total no. of frames (from flying UAV dataset, speech)	2736 [64 s]
RPSL parameters	
$\Delta\vec{\theta}_{\text{res}}$ (deg)	50 (task 2), 35 (task 3)

the SPCup syllabus [1], for which a DAE with much higher performance is expected.

4.3 Evaluation metric

The performance of the proposed method is evaluated against the baseline method [27], with GCC-PHAT, GCC-NONLIN, MVDR, DS, and DNM as the localisation techniques (as provided by the SPCup), using both sum and max aggregation (i.e. (13)–(16)). This results in 10 baseline methods to compare against the proposed method. For tasks 2 and 3, due to the proposed method containing two distinct components (SNR response scaling and RPSL post-processing), results with both components applied, as well as those with only one of the components applied, are presented for comparison.

Since the true locations for each test case are unique regardless of the chosen task, instead of directly comparing the estimated locations against the ground truth, the error between the estimated locations and the true locations is calculated and used as the evaluation metric. For the 3D problem presented in this study, the accuracy of the localisation performance is measured by comparing the haversine distance error D [47] (for which will be referred as *distance error*) between the estimated location $\hat{\vec{\theta}}_S$ and the true location $\vec{\theta}_S$, assuming a unit sphere (i.e. $r = 1$, such that $0 \leq D \leq \pi$). This is obtained as follows:

$$\gamma = \sin^2\left(\frac{\hat{\theta}_{S,\text{el}} - \theta_{S,\text{el}}}{2}\right) + \cos\left(\hat{\theta}_{S,\text{el}}\right)\cos\left(\theta_{S,\text{el}}\right)\sin^2\left(\frac{\hat{\theta}_{S,\text{az}} - \theta_{S,\text{az}}}{2}\right), \tag{20}$$

$$D = 2r\sin^{-1}\sqrt{\gamma}. \tag{21}$$

Using the haversine distance error measure instead of directly comparing the difference between $\hat{\vec{\theta}}_S$ and $\vec{\theta}_S$ alleviates the varying sensitivity of $\hat{\theta}_{S,\text{az}}$ to the same amount of error at different $\hat{\theta}_{S,\text{el}}$ (e.g. same amount of angular error in $\hat{\theta}_{S,\text{el}}$ results in significantly different position errors between $\hat{\theta}_{S,\text{az}} = 0$ deg, and $\hat{\theta}_{S,\text{az}} = \pm 90$ deg).

After obtaining the distance error D, its mean, root mean square error (RMSE), maximum, minimum, and quartile measures are compared between the methods. Statistical analysis using the paired sample t test was conducted to evaluate and verify the difference of mean and median errors between the proposed method and each baseline method, respectively. To avoid erroneous inferences caused by multiple comparison, Bonferroni's correction was applied [48], i.e. for each test, the null hypothesis was rejected by $p < 0.005$ ($= 0.05/10$, given the performance is evaluated against the 10 baseline methods), using the best-performing combination from the proposed method as the benchmark for comparison.

5 Results and discussion

This section presents the experimental results for evaluating the performance of the proposed method against the baseline method as outlined in Section 4.3.

5.1 Task 1: Hovering UAV scenario

Figure 8 shows the distance errors of different localisation methods from task 1, with details of the statistical test results shown in Table 3. As shown, with all baseline techniques using max aggregation, the addition of SNR response scaling delivered improvements with respect to its unscaled baseline. In particular, the MVDR method

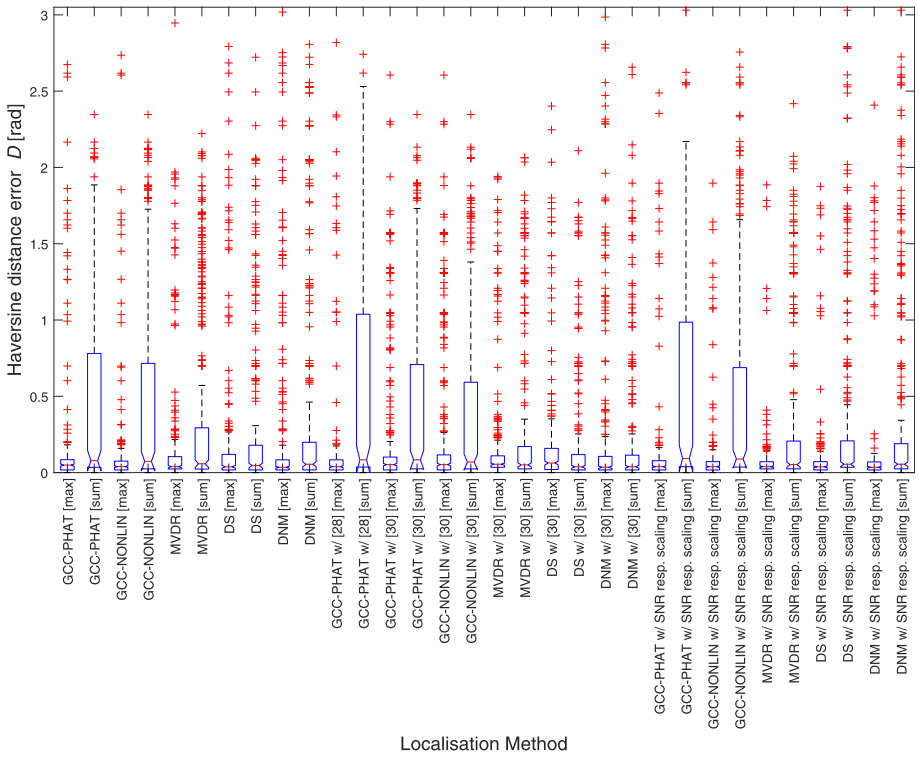

Fig. 8 Hovering UAV (task 1)—haversine distance error distribution. Red line indicates the median, upper and lower edges of the blue box indicate the 75% and 25% quartiles, upper and lower black bars indicate the maximum and minimum, and upper and lower corners of the trapezoidal indicate the 95% upper and lower confidence limits

with max aggregation and SNR response scaling is the best-performing combination. The method outperformed the baseline methods by delivering the lowest mean, maximum, and RMSE distance error measures. Results of the paired sample t test also indicate that the difference of the mean distance error being significantly different against the baseline methods, except GCC-NONLIN using max aggregation. The most apparent improvement is the significant reduction in outlier predictions, which can be seen in Fig. 8, as well as the reduction in the maximum and 0.75 quartile distance errors. This shows that SNR response scaling cleans up the rotor noise effects well, making the SNR response of the target sound source more apparent. Figure 9a, b shows an example of the SNR angular spectral response improvements brought upon using SNR response scaling. Here, influences from the UAV rotor noise are greatly reduced, revealing the peak response of the target sound source. As a result, it brought the estimated location closer to the ground truth. There are still a few cases (cases #59, 61, 70, 159, 160, 178, and 229 in the dataset [1]), which have very low input SNRs, and all methods evaluated (including the proposed method) are not able to give an accurate estimate. Nonetheless, the proposed method has shown a significant improvement in localising accuracy.

Apart from GCC-NONLIN, DNM method with max aggregation and SNR response scaling also delivered comparable results. The method delivered lower median and quartile distance error measures, with t test results showing that it is not significantly different to that of the MVDR method using max aggregation and SNR response scaling. However, due to the lower mean and RMSE distance errors, the MVDR method is considered the best-performing combination.

Since the SNR response scaling approach is essentially a T-F mask for filtering out effects of the UAV rotor noise, we compare its performance against other state-of-the-art T-F masks specialised for noisy and reverberant environments, using the study from [28] and [30]. Like SNR response scaling, the T-F mask from [28] is applied to the baseline method. However, given the method is designed for GCC-PHAT, results were only generated under this localisation method. As shown in Fig. 8 and Table 3, the T-F mask from [28] improved the localisation performance overall, reducing the distance error measures with respect to its corresponding baseline. Under the same GCC-PHAT localisation technique, the proposed SNR response scaling method overall outperformed [28] slightly, delivering lower mean, quartile, and RMSE distance error measures. However, with max aggregation,

Table 3 Haversine distance error performance comparison (task 1—hovering UAV)

	Haversine distance error *D* (rad)							***p* value: paired sample *t* test (ref. best case)**
	Mean	**Median**	**Min**	**Max**	**0.25 quartile**	**0.75 quartile**	**RMSE**	
Baseline								
GCC-PHAT (max)	0.1541	0.04908	**0**	2.674	**0.01745**	0.0849	0.4384	6.86×10^{-4}
GCC-PHAT (sum)	0.4637	0.07804	**0**	2.347	0.03491	0.7808	0.8073	3.42×10^{-21}
GCC-NONLIN (max)	0.1305	0.03903	**0**	2.736	**0.01745**	0.0752	0.3887	n.s.
GCC-NONLIN (sum)	0.4528	0.07376	**0**	2.347	0.02468	0.7160	0.8010	2.08×10^{-20}
MVDR (max)	0.1774	0.03903	**0**	2.947	0.02369	0.1047	0.4471	1.18×10^{-5}
MVDR (sum)	0.3738	0.05794	**0**	2.222	0.02314	0.2934	0.7081	3.91×10^{-17}
DS (max)	0.1976	0.03801	**0**	2.793	**0.01745**	0.1196	0.5143	4.81×10^{-6}
DS (sum)	0.2935	0.04637	**0**	2.722	**0.01745**	0.1795	0.6218	4.71×10^{-12}
DNM (max)	0.2535	**0.03491**	**0**	3.018	**0.01745**	0.0837	0.6435	3.14×10^{-7}
DNM (sum)	0.3811	0.05456	**0**	2.806	**0.01745**	0.1988	0.8074	1.87×10^{-13}
w/ [28] T-F mask								
GCC-PHAT (max)	0.1356	0.03903	**0**	2.818	**0.01745**	0.0837	0.4064	n.s.
GCC-PHAT (sum)	0.5164	0.08382	**0**	2.742	0.03654	1.0378	0.8776	1.48×10^{-23}
w/ [30] T-F mask								
GCC-PHAT (max)	0.2155	0.05236	**0**	2.605	**0.01745**	0.1018	0.4933	1.10×10^{-7}
GCC-PHAT (sum)	0.4162	0.08372	**0**	2.347	0.02468	0.7080	0.7424	1.16×10^{-19}
GCC-NONLIN (max)	0.2411	0.05236	**0**	2.605	**0.01745**	0.1162	0.5308	3.13×10^{-9}
GCC-NONLIN (sum)	0.3837	0.06981	**0**	2.347	0.02429	0.5923	0.7012	3.90×10^{-18}
MVDR (max)	0.1806	0.05504	**0**	1.941	0.03491	0.1101	0.4154	1.39×10^{-7}
MVDR (sum)	0.2702	0.05058	**0**	2.065	0.02427	0.1711	0.5633	4.44×10^{-12}
DS (max)	0.1897	0.06292	**0**	2.403	0.02030	0.1589	0.4344	2.11×10^{-7}
DS (sum)	0.1792	0.03796	**0**	2.110	0.01745	0.1180	0.4095	1.64×10^{-7}
DNM (max)	0.2586	0.03504	**0**	2.986	**0.01745**	0.1064	0.6273	2.35×10^{-8}
DNM (sum)	0.2113	0.03810	**0**	2.657	**0.01745**	0.1144	0.4941	4.24×10^{-9}
w/ SNR response scaling								
GCC-PHAT (max)	0.1278	0.03903	**0**	2.488	**0.01745**	0.0780	0.3732	n.s.
GCC-PHAT (sum)	0.4997	0.09259	**0**	3.031	0.03903	0.9861	0.8501	7.02×10^{-24}
GCC-NONLIN (max)	0.0961	0.03880	**0**	1.897	**0.01745**	0.0719	0.2596	n.s.
GCC-NONLIN (sum)	0.4575	0.08899	**0**	2.756	0.03654	0.6880	0.8133	1.16×10^{-20}
MVDR (max) (best case)	**0.0833**	0.03903	**0**	**1.886**	0.02424	0.0715	**0.2269**	N/A
MVDR (sum)	0.3263	0.05236	**0**	2.418	0.02468	0.2067	0.6420	1.10×10^{-14}
DS (max)	0.0975	0.03803	**0**	1.876	**0.01745**	0.0719	0.2790	n.s.
DS (sum)	0.3275	0.05504	**0**	3.031	0.03491	0.2080	0.7017	2.99×10^{-12}
DNM (max)	0.1323	**0.03491**	**0**	2.409	**0.01745**	**0.0698**	0.3893	n.s.
DNM (sum)	0.3515	0.05504	**0**	3.031	0.02468	0.1896	0.7498	2.16×10^{-12}

Results from the baseline method are first presented, followed by results using the T-F mask from [28] and [30] and the proposed method (SNR response scaling). Best-performing numericals for each category are highlighted in bold

the performance improvement is slight, as suggested by the *t* test. The T-F mask proposed by [30] is utilised via source enhancement using the minimum mean square error (MMSE) log-spectral amplitude estimator from [49] prior to source localisation using [27]. Contrary to the T-F mask from [28] and the proposed method, the T-F mask from [30] showed mixed performance. While there was general improvement in results with max aggregation, the

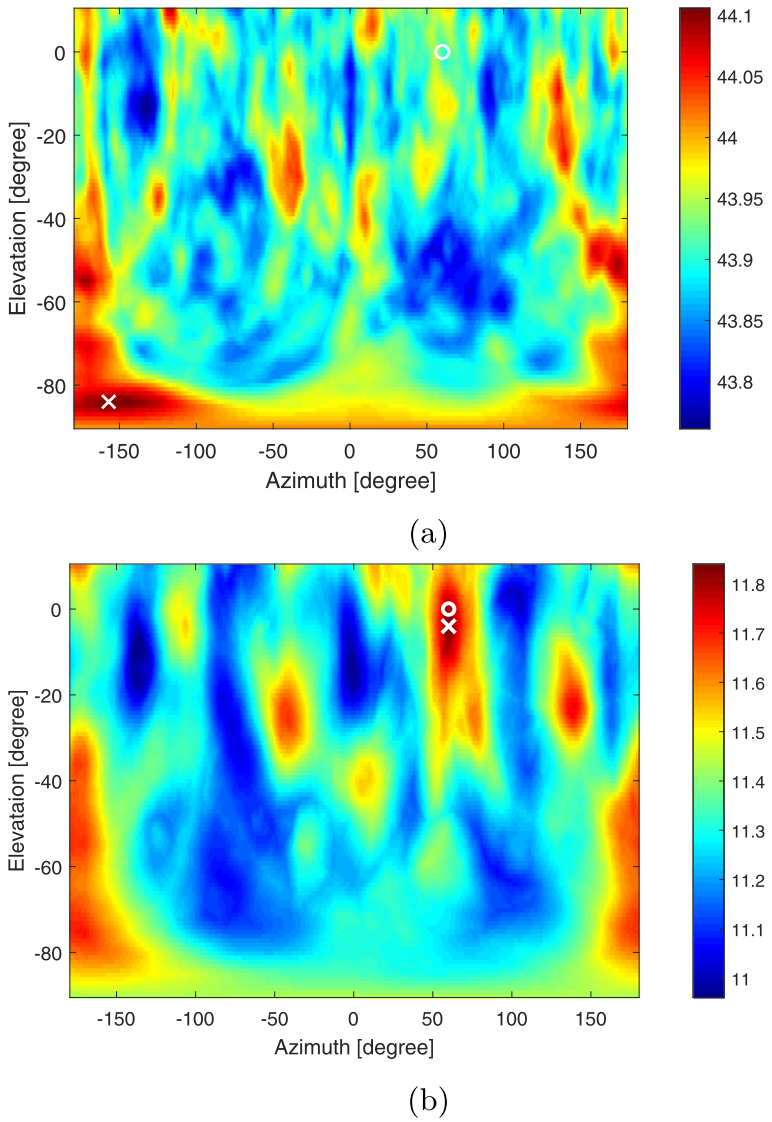

Fig. 9 SNR angular spectral response (dB) from SPCup hovering UAV (task 1) case 297 **a** w/o SNR response scaling and **b** w/ SNR response scaling, using MVDR with max aggregation. Circle (∘) and cross (×) in the diagram represent the ground truth and the algorithm's estimated peak response location, respectively

T-F mask performed worse than the baseline with sum aggregation, with the exception of DS, where both aggregation methods improved over the baseline. Since the T-F mask from [30] assumes continuity in the noise signals, which is a valid assumption, it could have been affected by the vast amount of wind/flow noise generated by the UAV rotors. With the nature of such noise being stochastic, the resultant enhanced signal could have introduced potential distortions. As such, the proposed SNR response scaling method overall outperformed [30] by a visible margin. This indicates that while a diffuse noise-based T-F mask is able to remove some aspects of the noise, a dedicated noise mask designed for UAV rotor noise would still be the desired option.

5.2 Task 2: Flying UAV scenario, broadband sound source

Figure 10 shows the localisation results for task 2, with details of the statistical test results shown in Table 4. Due to the lack of relevant UAV rotor noise data in the flying UAV cases for effective DAE training, the baseline method outperformed the SNR response scaled method under all localisation techniques. Observing the SNR response angular spectrum of the baseline and SNR response scaled method (see Fig. 11a, b), although SNR response scaling reduced noise surrounding the target source location, it came with a peak response for the target source less sharp than the baseline method, as shown in Fig. 11b. This could lead to increased variation in the location path estimations and thus decreasing overall accuracy. It is believed this is

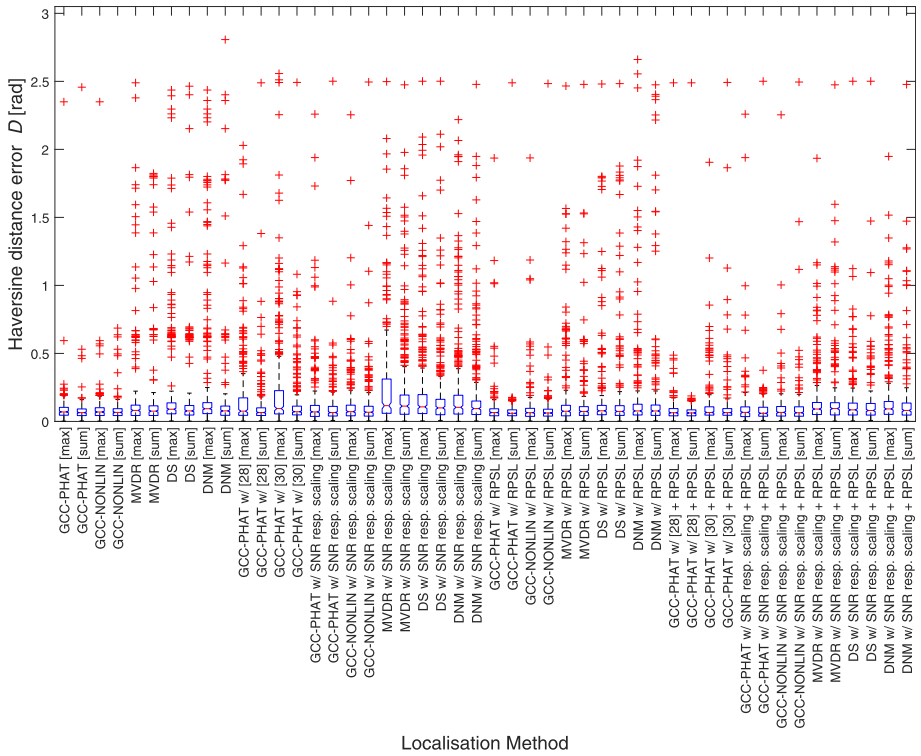

Fig. 10 Flying UAV—broadband sound source (task 2) haversine distance error distribution. Red line indicates the median, upper and lower edges of the blue box indicate the 75% and 25% quartiles, upper and lower black bars indicate the maximum and minimum, and upper and lower corners of the trapezoidal indicate the 95% upper and lower confidence limits

caused by the limited amount of available data from the DREGON dataset [44] for training the DAE: a limitation at the point of competition.

In contrast, the proposed method with only the RPSL post-processing algorithm applied outperformed the baseline with most of the localisation techniques, showing improvement in distance error measures, except being GCC-PHAT and GCC-NONLIN with max aggregation, where performance is similar. For this task, GCC-PHAT using sum aggregation is the best-performing combination, as evident in Table 4. In particular, the number of outlier cases significantly reduced, as evident in Fig. 10. This is also evident with localisation techniques other than GCC-PHAT. Since the primary function of GCC-PHAT (and GCC-NONLIN, another well-performing option) involves calculating the cross-correlation between pairs of microphone signals at different TDOAs, it is generally not influenced by spatial aliasing between the microphones and thus, under a certain input SNR, delivers consistent performance. However, as expected from the drawbacks from SNR response scaling, results with both SNR response scaling and RPSL post-processing algorithm applied are not the best-performing combinations. While the combinations reduced mean distance error overall, other metrics such as median and quartile measures delivered mixed results, compared to the baseline method.

An observation that should be noted is the exceptional performance from the baseline method using GCC-NONLIN with sum aggregation. While it is not the best-performing combination under mean and median distance errors, it delivered the lowest RMSE and maximum distance errors. From observing Fig. 10 and Table 4, it is apparent that GCC-NONLIN with sum aggregation is the only combination where there are no significant outliers, while the combination with RPSL post-processing has a single outlier. This could be the leading cause of this result. However, it should be noted that the RPSL post-processed variant presented lower distance error measures in the remaining categories, and thus, the benefits of RPSL should not be overlooked.

From observing the actual location estimates in Fig. 12, while the estimated location path has shown a slightly closer correlation with the ground truth, the RPSL post-processing method seems to have reduced some unstable variations along the location path. Therefore, in some respect, the RPSL algorithm seems to achieve a regularisation effect for the estimation of the path of locations, while bringing the overall location estimates closer to the ground truth. Another one of such example is shown in

Table 4 Haversine distance error performance comparison (task 2—flying UAV, broadband sound source)

	Haversine distance error *D* (rad)							***p* value: paired sample *t* test (ref. best case)**
	Mean	**Median**	**Min**	**Max**	**0.25 quartile**	**0.75 quartile**	**RMSE**	
Baseline								
GCC-PHAT (max)	0.0868	0.07004	0.002181	2.350	0.04567	0.1035	0.1662	n.s.
GCC-PHAT (sum)	0.0810	0.06459	0.004032	2.457	0.04299	0.0918	0.1690	n.s.
GCC-NONLIN (max)	0.0905	0.06951	0.002702	2.350	0.04345	0.1008	0.1746	n.s.
GCC-NONLIN (sum)	0.0811	0.06633	0.004032	**0.686**	0.04325	0.0936	0.1155	n.s.
MVDR (max)	0.1690	0.08072	0.004032	2.490	0.04379	0.1199	0.3743	7.88×10^{-6}
MVDR (sum)	0.1590	0.07550	0.003670	1.823	0.04377	0.1148	0.3633	5.15×10^{-5}
DS (max)	0.2293	0.08965	0.004032	2.436	0.05588	0.1365	0.4655	1.10×10^{-9}
DS (sum)	0.1648	0.07734	0.002618	2.465	0.04694	0.1169	0.3625	1.21×10^{-5}
DNM (max)	0.2619	0.09346	0.004032	2.436	0.05745	0.1381	0.5411	2.02×10^{-10}
DNM (sum)	0.1563	0.07733	0.003670	2.807	0.04490	0.1130	0.3910	2.62×10^{-4}
w/ [28] T-F mask								
GCC-PHAT (max)	0.1816	0.07557	0.002919	2.029	0.04480	0.1733	0.3429	1.78×10^{-11}
GCC-PHAT (sum)	0.1078	0.06546	0.005047	2.490	0.04372	0.1000	0.2203	5.32×10^{-6}
w/ [30] T-F mask								
GCC-PHAT (max)	0.2472	0.09462	0.002449	2.558	0.05509	0.2263	0.4670	2.35×10^{-14}
GCC-PHAT (sum)	0.1285	0.07141	0.004966	2.492	0.04797	0.1124	0.2464	4.34×10^{-9}
w/ SNR response scaling								
GCC-PHAT (max)	0.1335	0.07025	0.005124	2.259	0.03684	0.1176	0.2813	6.94×10^{-6}
GCC-PHAT (sum)	0.1010	0.06524	0.007256	2.501	0.03983	0.1105	0.1973	1.48×10^{-6}
GCC-NONLIN (max)	0.1227	0.07184	0.006535	2.254	0.04138	0.1196	0.2403	1.46×10^{-4}
GCC-NONLIN (sum)	0.1101	0.06766	0.004056	2.495	0.03898	0.1116	0.2240	6.54×10^{-6}
MVDR (max)	0.2669	0.11614	0.002071	2.498	0.05972	0.3117	0.4617	7.10×10^{-19}
MVDR (sum)	0.2468	0.11260	0.002633	2.475	0.05254	0.1940	0.4403	9.62×10^{-17}
DS (max)	0.2333	0.10711	0.004002	2.501	0.05783	0.1973	0.4350	1.37×10^{-14}
DS (sum)	0.1957	0.09952	0.007290	2.501	0.04939	0.1626	0.3760	2.86×10^{-12}
DNM (max)	0.2448	0.10371	**0.000610**	2.219	0.05417	0.1925	0.4523	1.69×10^{-14}
DNM (sum)	0.1967	0.09452	0.003334	2.478	0.05081	0.1501	0.3783	2.13×10^{-12}
w/ RPSL post-processing								
GCC-PHAT (max)	0.0922	0.06516	0.003577	1.936	0.04173	0.0930	0.1844	n.s.
GCC-PHAT (sum) (best case)	0.0746	0.05987	0.004076	2.490	0.04177	0.0852	0.1622	N/A
GCC-NONLIN (max)	0.0965	0.06428	0.003356	1.937	0.03727	0.0962	0.1927	3.34×10^{-3}
GCC-NONLIN (sum)	0.0805	0.06190	0.002988	2.484	0.03913	0.0900	0.1706	n.s.
MVDR (max)	0.1613	0.07477	0.001200	2.466	0.04130	0.1186	0.3334	7.74×10^{-9}
MVDR (sum)	0.1244	0.07330	0.003783	2.478	0.04414	0.1089	0.2646	3.56×10^{-6}
DS (max)	0.1619	0.07810	0.005783	2.481	0.04922	0.1179	0.3559	2.41×10^{-7}
DS (sum)	0.1689	0.07352	0.002433	2.484	0.04421	0.1159	0.3812	3.06×10^{-7}
DNM (max)	0.1810	0.07726	0.004760	2.661	0.04732	0.1266	0.4117	2.86×10^{-7}
DNM (sum)	0.1777	0.07683	0.004642	2.475	0.04434	0.1205	0.4284	1.12×10^{-6}

Table 4 Haversine distance error performance comparison (task 2—flying UAV, broadband sound source) (*Continued*)

	Haversine distance error D (rad)							p value: paired sample t test (ref. best case)
	Mean	**Median**	**Min**	**Max**	**0.25 quartile**	**0.75 quartile**	**RMSE**	
w/ [28] T-F mask + RPSL post-processing								
GCC-PHAT (max)	0.0845	0.06388	0.002919	2.490	0.04131	0.0944	0.1750	3.10×10^{-3}
GCC-PHAT (sum)	**0.0743**	0.06234	0.005047	2.490	0.04126	**0.0851**	0.1621	n.s.
w/ [30] T-F mask + RPSL post-processing								
GCC-PHAT (max)	0.1125	0.06586	0.002449	1.905	0.04632	0.1076	0.2055	4.39×10^{-6}
GCC-PHAT (sum)	0.1038	0.06491	0.004966	2.492	0.04532	0.0944	0.2311	1.03×10^{-3}
w/ SNR response scaling + RPSL post-processing								
GCC-PHAT (max)	0.0979	0.06657	0.005124	2.259	**0.03562**	0.1070	0.2145	n.s.
GCC-PHAT (sum)	0.0827	**0.05976**	0.007256	2.501	0.03822	0.1043	0.1704	2.03×10^{-3}
GCC-NONLIN (max)	0.0954	0.06629	0.006180	2.254	0.03947	0.1096	0.1854	n.s.
GCC-NONLIN (sum)	0.0971	0.06379	0.004056	2.495	0.03823	0.1076	0.2084	1.20×10^{-3}
MVDR (max)	0.1476	0.09110	0.002071	1.935	0.05083	0.1392	0.2560	2.79×10^{-11}
MVDR (sum)	0.1590	0.09449	0.002633	2.475	0.04783	0.1377	0.3062	2.02×10^{-10}
DS (max)	0.1321	0.08498	0.004002	2.501	0.04730	0.1347	0.2426	2.21×10^{-10}
DS (sum)	0.1208	0.08031	0.007290	2.501	0.04661	0.1321	0.2312	2.51×10^{-8}
DNM (max)	0.1693	0.09204	**0.000610**	1.948	0.05196	0.1463	0.3022	5.16×10^{-12}
DNM (sum)	0.1429	0.08071	0.003334	2.478	0.04672	0.1352	0.2823	4.35×10^{-9}

Results from the baseline method are first presented, followed by results using the T-F mask from [28] and [30] and the proposed method (SNR response scaling and RPSL). Best-performing numericals for each category are highlighted in bold

Fig. 13, the RPSL post-processing algorithm is able to limit the amount of fluctuation in location estimates relative to the baseline method, giving a more stable path.

Comparing the performance of GCC-PHAT using the proposed method against the T-F mask from [28] showed that both methods could not outperform the baseline. While SNR response scaling alone outperformed [28], when paired with the RPSL post-processing algorithm, [28] outperformed SNR response scaling. In fact, GCC-PHAT using sum aggregation with the T-F mask from [28] and RPSL post-processing is almost arguably the best-performing combination, delivering lowest mean and 0.25 quartile distance error measures, as shown in Table 4. However, due to the same combination without T-F masking giving near-identical performance, except median distance error, for which showed visibly better improvements, was considered the best-performing combination. Perhaps due to the T-F mask not being a data-driven solution, it is more stable against unfamiliar scenarios. Given that the T-F mask is primarily designed for speech signals, with the target source being broadband noise, could be the cause of the lack in performance. The T-F mask from [30] delivered the worst results compared to the other presented methods. Using the best-performing localisation technique (GCC-PHAT) showed that it was unable to outperform both [28] and the proposed method, with or without RPSL post-processing. This is likely driven by the broadband target source, where its diffuse characteristics rendered it difficult to distinguish the time-continuity in the UAV rotor noise from the target sound mixed signal. Overall, SNR response scaling and the T-F mask from [28] and [30] struggled to deliver noticeable improvements in localisation performance.

5.3 Task 3: Flying UAV scenario, speech sound source

Figure 14 shows the localisation results for task 3, with details of the statistical test results shown in Table 5. Similar to the results in task 2, the proposed method using only RPSL post-processing outperformed most baseline methods, delivering lower overall distance error measures under all aspects (mean, median, etc.). Furthermore, the proposed method with both SNR response scaling and RPSL post-processing using many of the localisation techniques also outperformed most baseline methods. In this task, the DS localisation technique with max aggregation and RPSL post-processing is the best-performing combination, delivering the lowest overall mean, median, and quartile distance error measures, with t test results indicating the improvement is distinct.

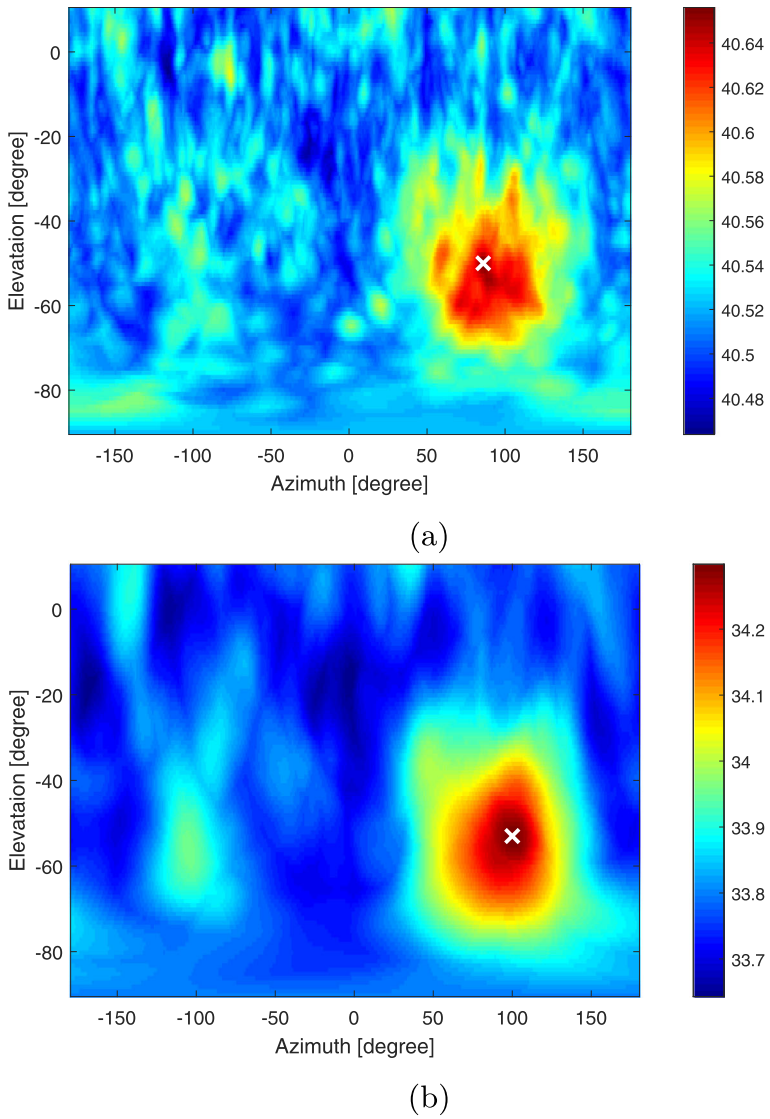

Fig. 11 SNR angular spectral response (dB) from SPCup flying UAV broadband sound source (task 2) case 11 **a** w/o SNR response scaling and **b** w/ SNR response scaling, using GCC-NONLIN with max aggregation. Cross (×) in the diagram represents the algorithm's estimated peak response location

One aspect to note is that the many of the localisation techniques using both SNR response scaling and RPSL post-processing delivered lower mean, median, 0.75 quartile, and RMSE measured compared to the same setup without SNR response scaling, as shown in Fig. 14 and Table 5. The only exceptions are MVDR with sum aggregation, DS with max aggregation, and DNM with max aggregation. This indicates that while SNR response scaling lowered the sensitivity in estimating peak response locations more accurately, its ability to reduce unwanted noise is still apparent. Furthermore, under the proposed method with only SNR response scaling, except MVDR with sum aggregation, showed a reduction in maximum distance errors relative to the baseline method. Given that SNR response scaling also improved the performance of MVDR with max aggregation for task 1 suggest that SNR response scaling is still able to deliver some benefits over the baseline method when paired with the RPSL post-processing algorithm. Another potential aspect could be driven by the temporally sparse nature of speech sources. This allows the distinguishing between target and UAV noise sources to be easier than, for example, broadband sources, which is much more continuous and diffuse. However, despite these indications of improvement, further tuning and proper DAE training are still required to bring out the true performance gains of SNR response scaling.

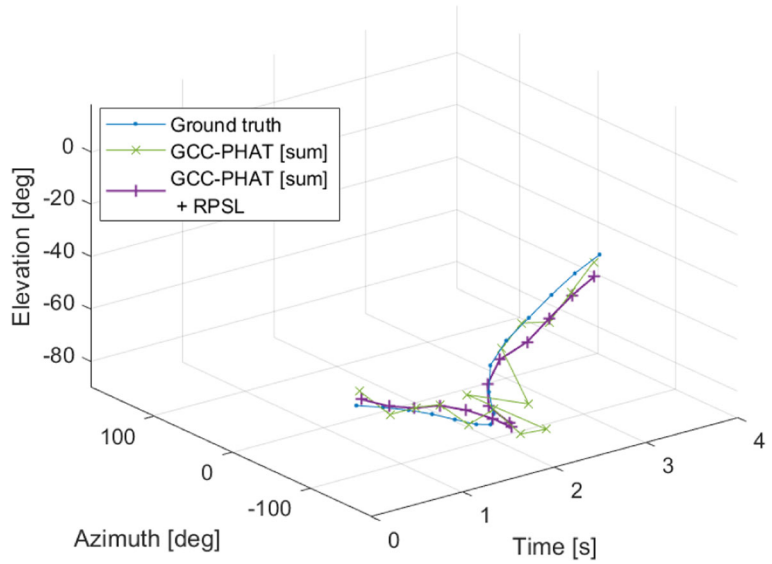

Fig. 12 Example of localisation path estimated for flying UAV broadband sound source (task 2) case 14 [1], using the baseline and proposed method (RPSL only)

Despite significant improvements in the overall distance error reduction, the accuracy of the predictions for most cases is not yet satisfactory for robust localisation path estimation. This is suspected to be caused by the non-stationary nature of speech (i.e. not all time frames contained the target sound source), which may be an issue with methods based on using the SNR response in angular spectra, according to the previous study [27]. Therefore, the fact that localisation can only be carried out in a limited number of time frames would have caused the proposed method to degrade significantly in performance when the source signal was speech. In addition, the input SNRs in task 3 seem to be much lower than that of other tasks in the DREGON database. This elevates the challenge in estimating the peak response of the target audio source.

Figures 15 and 16 demonstrate two of the more successful path estimates using the proposed method with only RPSL post-processing applied, compared against the baseline method (cases 2 and 3 [1]), using DS with max

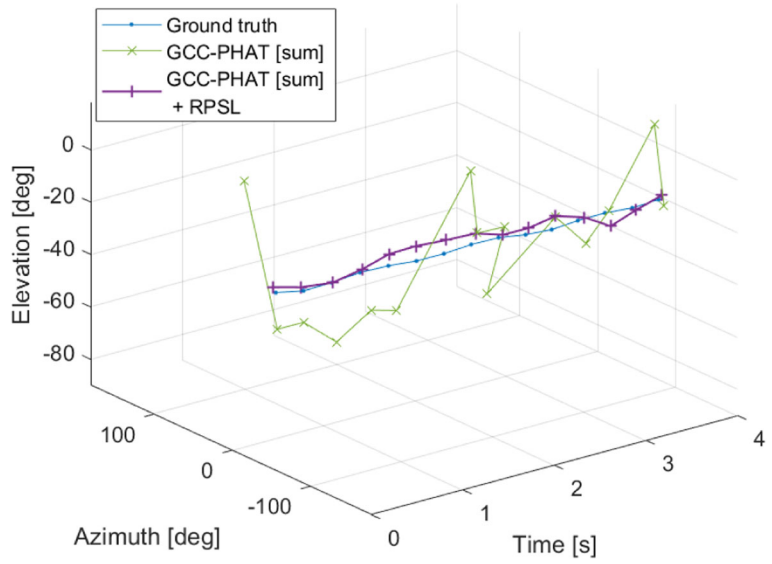

Fig. 13 Example of localisation path estimated for flying UAV broadband sound source (task 2) case 18 [1], using the baseline and proposed method (RPSL only)

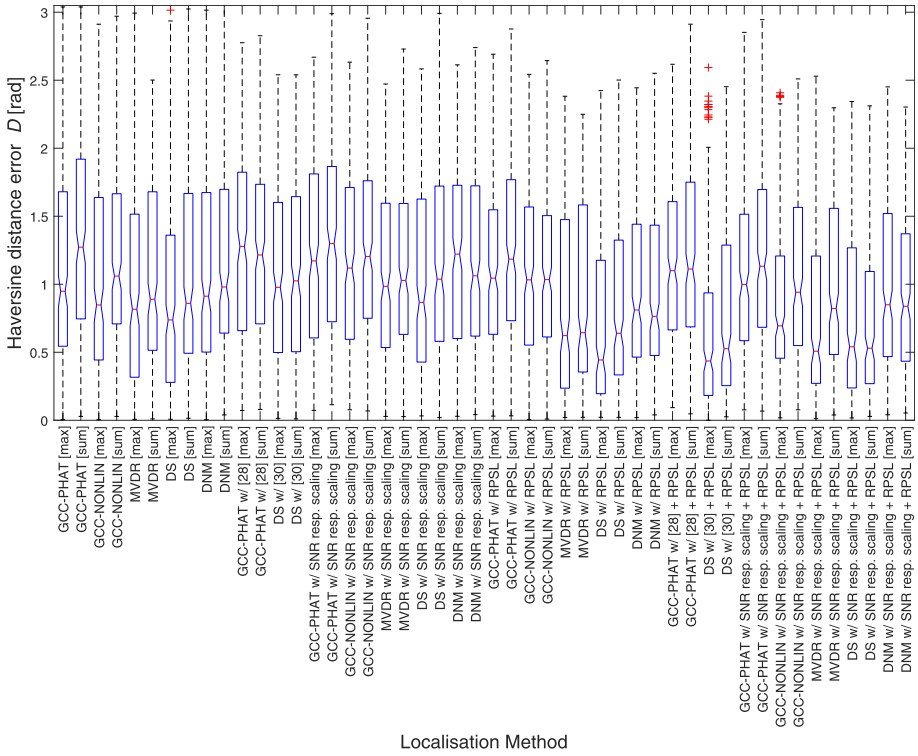

Fig. 14 Flying UAV—speech sound source (task 3) haversine distance error distribution. Red line indicates the median, upper and lower edges of the blue box indicate the 75% and 25% quartiles, upper and lower black bars indicate the maximum and minimum, and upper and lower corners of the trapezoidal indicate the 95% upper and lower confidence limits

Table 5 Haversine distance error performance comparison (task 3—flying UAV, speech sound source)

	Haversine distance error *D* (rad)							***p* value: paired sample *t* test (ref. best case)**
	Mean	**Median**	**Min**	**Max**	**0.25 quartile**	**0.75 quartile**	**RMSE**	
Baseline								
GCC-PHAT (max)	1.143	0.9486	0.00440	3.039	0.5444	1.680	1.362	1.81×10^{-17}
GCC-PHAT (sum)	1.375	1.2730	0.02862	3.039	0.7456	1.920	1.582	2.11×10^{-29}
GCC-NONLIN (max)	1.022	0.8480	0.00440	2.912	0.4434	1.638	1.233	2.13×10^{-11}
GCC-NONLIN (sum)	1.206	1.0604	0.02862	2.971	0.7088	1.665	1.390	8.98×10^{-21}
MVDR (max)	0.913	0.8164	0.00708	2.993	0.3166	1.515	1.138	5.83×10^{-7}
MVDR (sum)	1.045	0.8897	0.01019	2.501	0.5146	1.680	1.232	2.39×10^{-18}
DS (max)	0.881	0.7378	0.00800	3.015	0.2789	1.361	1.141	1.27×10^{-4}
DS (sum)	1.083	0.8598	0.01492	3.024	0.4927	1.668	1.328	2.47×10^{-16}
DNM (max)	1.088	0.9123	0.00437	3.015	0.5015	1.674	1.300	2.31×10^{-16}
DNM (sum)	1.160	0.9798	0.03962	3.056	0.6410	1.698	1.356	1.19×10^{-20}
w/ [28] T-F mask								
GCC-PHAT (max)	1.277	1.2783	0.07181	2.776	0.6593	1.824	1.441	1.18×10^{-31}
GCC-PHAT (sum)	1.247	1.2153	0.07979	2.827	0.7090	1.735	1.398	1.06×10^{-28}
w/ [30] T-F mask								
DS (max)	1.052	0.9775	0.01399	2.540	0.4983	1.602	1.243	3.05×10^{-17}
DS (sum)	1.097	1.0249	0.01038	2.540	0.5042	1.644	1.281	7.08×10^{-20}

Table 5 Haversine distance error performance comparison (task 3—flying UAV, speech sound source) (*Continued*)

	Haversine distance error *D* (rad)							***p* value: paired sample *t* test (ref. best case)**
	Mean	**Median**	**Min**	**Max**	**0.25 quartile**	**0.75 quartile**	**RMSE**	
w/ SNR response scaling								
GCC-PHAT (max)	1.207	1.1717	0.07321	2.669	0.6054	1.811	1.390	5.77×10^{-22}
GCC-PHAT (sum)	1.325	1.2988	0.11448	2.989	0.7246	1.866	1.501	3.25×10^{-30}
GCC-NONLIN (max)	1.170	1.1192	0.07740	2.633	0.5953	1.711	1.347	2.85×10^{-21}
GCC-NONLIN (sum)	1.250	1.2045	0.06879	2.956	0.7501	1.761	1.404	1.75×10^{-29}
MVDR (max)	1.045	0.9847	0.02859	2.472	0.5349	1.596	1.204	2.97×10^{-16}
MVDR (sum)	1.103	1.0282	0.02711	2.729	0.6316	1.594	1.253	1.22×10^{-22}
DS (max)	0.999	0.8668	0.03225	2.583	0.4278	1.626	1.198	1.15×10^{-11}
DS (sum)	1.115	1.0375	0.01969	2.991	0.5803	1.722	1.292	4.33×10^{-21}
DNM (max)	1.174	1.2215	0.02945	2.613	0.6012	1.728	1.345	2.37×10^{-23}
DNM (sum)	1.160	1.0620	0.04191	2.740	0.6189	1.724	1.323	5.48×10^{-24}
w/ RPSL post-processing								
GCC-PHAT (max)	1.129	1.0451	0.03077	2.691	0.6323	1.547	1.282	3.22×10^{-22}
GCC-PHAT (sum)	1.294	1.1847	0.03292	2.877	0.7323	1.769	1.458	1.44×10^{-31}
GCC-NONLIN (max)	1.083	1.0325	0.00474	2.543	0.5527	1.568	1.265	1.66×10^{-17}
GCC-NONLIN (sum)	1.093	1.0342	0.00901	2.644	0.6126	1.505	1.251	4.22×10^{-20}
MVDR (max)	0.826	0.6226	0.02069	2.382	0.2362	1.476	1.063	3.07×10^{-5}
MVDR (sum)	0.864	0.6437	0.02214	**2.250**	0.3550	1.583	1.078	4.67×10^{-8}
DS (max)	0.706	0.4435	0.02207	2.424	0.1956	1.176	0.962	n.s.
DS (sum)	0.850	0.6395	0.02145	2.501	0.3336	1.325	1.071	1.28×10^{-6}
DNM (max)	0.982	0.8109	0.01968	2.444	0.4646	1.441	1.167	1.33×10^{-14}
DNM (sum)	0.980	0.7641	0.03962	2.551	0.4765	1.434	1.169	1.81×10^{-15}
w/ [28] T-F mask + RPSL post-processing								
GCC-PHAT (max)	1.167	1.0996	0.09304	2.617	0.6643	1.607	1.316	8.90×10^{-26}
GCC-PHAT (sum)	1.285	1.1122	0.04744	2.912	0.6868	1.751	1.473	4.50×10^{-30}
w/ [30] T-F mask + RPSL post-processing								
DS (max) (best case)	**0.684**	**0.4362**	**0.00038**	2.593	**0.1827**	**0.937**	**0.951**	N/A
DS (sum)	0.786	0.5264	0.02560	2.452	0.2546	1.288	1.015	2.66×10^{-3}
w/ SNR response scaling + RPSL post-processing								
GCC-PHAT (max)	1.067	0.9980	0.07649	2.852	0.5852	1.515	1.241	5.90×10^{-18}
GCC-PHAT (sum)	1.202	1.1322	0.06775	2.945	0.6837	1.696	1.373	4.18×10^{-23}
GCC-NONLIN (max)	0.893	0.6941	0.01799	2.408	0.4562	1.208	1.082	1.40×10^{-7}
GCC-NONLIN (sum)	1.066	0.9414	0.07722	2.510	0.5493	1.565	1.236	2.76×10^{-19}
MVDR (max)	0.770	0.5080	0.01199	2.530	0.2716	1.207	0.996	n.s.
MVDR (sum)	0.984	0.8222	0.03901	2.297	0.4840	1.558	1.167	2.44×10^{-15}
DS (max)	0.759	0.5406	0.01738	2.344	0.2379	1.268	0.996	n.s.
DS (sum)	0.753	0.5300	0.02859	2.311	0.2693	1.094	0.978	n.s.
DNM (max)	0.996	0.8491	0.04045	2.451	0.4684	1.520	1.189	9.52×10^{-15}
DNM (sum)	0.957	0.8381	0.05366	2.303	0.4342	1.371	1.142	1.19×10^{-12}

Results from the baseline method are first presented, followed by results using the T-F mask from [28] and [30] and the proposed method (SNR response scaling and RPSL). Best-performing numericals for each category are highlighted in bold

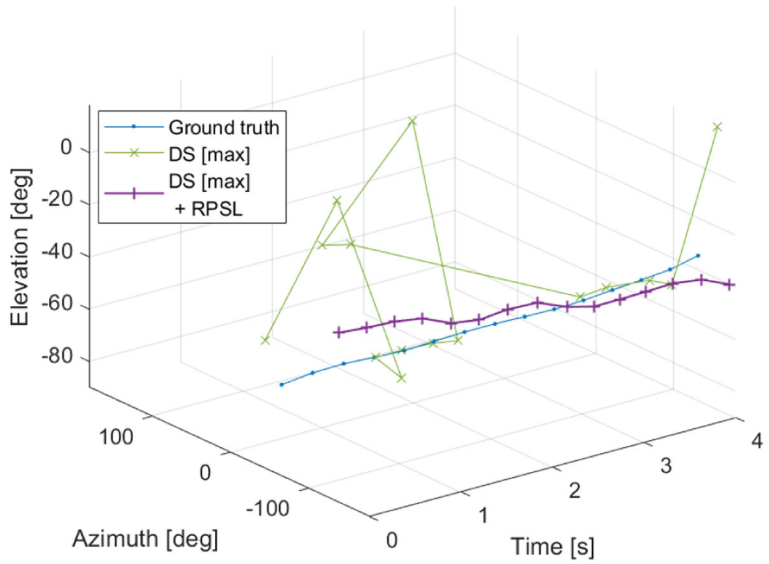

Fig. 15 Example of localisation path estimated for flying UAV speech sound source (task 3) case 2 [1], using the baseline and proposed method (RPSL only)

aggregation. As shown with the baseline method, the low input SNR coupled with speech source being temporally sparser than UAV rotor noise, there are major fluctuations in location estimations. On the other hand, the proposed method with only RPSL post-processing applied is able to estimate some of the locations along the path successfully. As some of the location estimates from the baseline method are correctly estimated, the RPSL post-processing algorithm is able to utilise these data points and perform restricted peak search (see Section 3.3), limiting influences coming from the UAV rotors and reverberation effects, and thereby give a much more accurate path estimate.

Figure 17 shows an unsuccessful example of localisation path estimation (case 10 [1]). Here, while there were a few correct estimates of $\hat{\hat{\theta}}_{S,\text{flight}}(t)$ using the baseline method, most are significantly different to that of the ground truth. Therefore, there was a limited basis for the RPSL post-processing algorithm to perform restricted peak search effectively. Therefore, while the variation in localisation

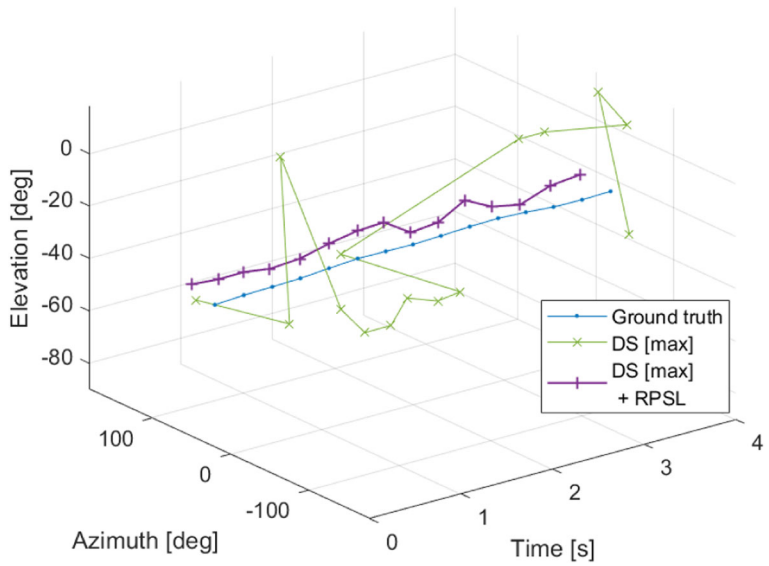

Fig. 16 Example of localisation path estimated for flying UAV speech sound source (task 3) case 3 [1], using the baseline and proposed method (RPSL only)

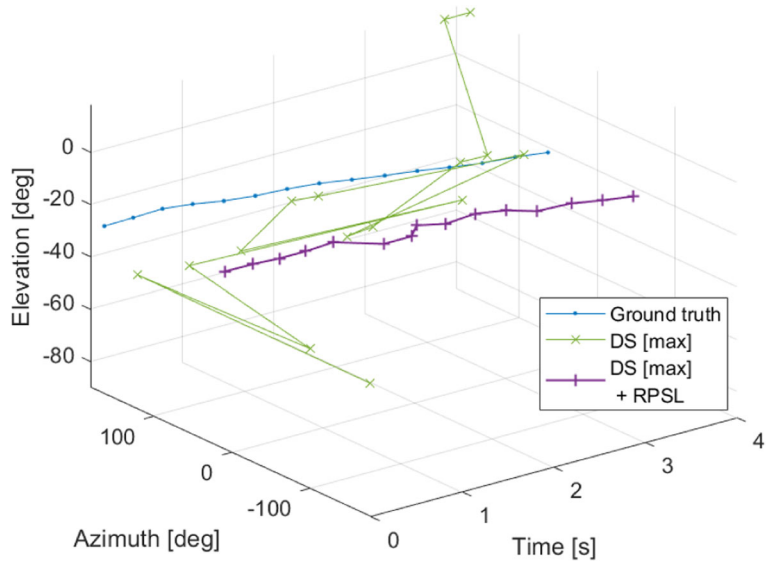

Fig. 17 Example of localisation path estimated for flying UAV speech sound source (task 3) case 10 [1], using the baseline and proposed method (RPSL only)

path estimate is much less chaotic compared to the baseline method, the overall path is incorrect. A potential method to resolve this deficiency would be to generate an initial $\hat{\tilde{\theta}}_{S,\text{flight}}(t)$ that takes in multiple peaks, such as the second and third largest peaks (instead of the single maximum peak), to create a wider grid of location points to perform local path search. Since it is unlikely that the UAV flight path would change in a severely rapid manner, having a larger number of potential local path estimates may grant a higher possibility in successful linking of the correct local paths together. However, exploring the problem further remains as future work.

Like with task 2, we compare the performance between GCC-PHAT with the proposed method against the T-F mask from [28], and the best-performing localisation technique using the T-F mask from [30] (DS). Again, as shown Fig. 14 and Table 5, both T-F masks and SNR response scaling could not outperform the baseline. Different to that from task 2, while SNR response scaling with max aggregation alone outperformed both [28], this was not the case with sum aggregation. However, when paired with the RPSL post-processing algorithm, SNR response scaling with max aggregation outperformed [28], with sum aggregation showing similar performance. On the other hand, comparing RPSL paired SNR response scaling against the pairing with [30] using DS showed similar performance. Although [30] was able to deliver slightly better performance figures with max aggregation (while SNR response scaling slightly outperformed [30] with sum aggregation), the p value from the paired sample t test indicates that this performance difference is not definitive. Therefore, the performance advantages delivered by SNR response scaling should not be overlooked. In addition, both T-F masks and SNR response scaling with RPSL post-processing outperformed RPSL post-processing alone. This could be driven by the target sound source being speech, where temporal sparsity can be expected, and some aspects of the UAV noise can be distinguished more successfully from the target source.

It should be noted that for both tasks 2 and 3, SNR response scaling suffered significantly from the lack of available training data. As mentioned in Section 4.1, given access to the UAV system for noise recordings, it is expected that noise removal performance would significantly increase with sufficient data available for proper DAE training, thereby improving localisation performance for both target source types. However, this remains a future investigation.

6 Conclusion

A method based on the multi-source TDOA estimation in reverberant audio using angular spectra, to perform sound localisation for a UAV-embedded audio recording system, is proposed. The study proposes extensions to improve localisation accuracy of the baseline method. The extensions include a means of reducing the UAV rotor noise effect via a weighting envelope based on the UAV rotor noise PSD. In addition, the proposed method also introduces an angular spectral range RPSL post-processing algorithm to improve localisation accuracies for the flying (moving) UAV scenario.

Experimental results using the dataset provided by the SPCup show that with proper DAE training, SNR response scaling improves SNR angular spectral response, resulting in a reduction in localisation error. The RPSL post-processing algorithm also displayed improvement in performance consistency even under low input SNR conditions when the source is a non-stationary signal, and the UAV is in motion. Future work includes accessing the UAV system for proper UAV rotor noise data collection, to properly investigate the SNR response scaling's ability to reduce UAV rotor noise effects under flying UAV scenarios. More challenging scenarios to investigate include rapid movement of the UAV or inclusion of spatially coherent interfering sound sources N_n.

Abbreviations

UAV: Unmanned aerial vehicle; SNR: Signal-to-noise ratio; TDOA: Time difference of arrival; PSD: Power spectral density; NN: Neural network; CNN: Convolutional neural network; DAE: Denoising autoencoder; MUSIC: Multiple signal classification; SRP-PHAT: Steered response power with phase transform; SPCup: 2019 IEEE Signal Processing Cup; STFT: Short-time Fourier transform; GCC-PHAT: Generalised cross-correlation-phase transform; GCC-NONLIN: Generalised cross-correlation- nonlinear; MVDR: Minimum variance distortionless response; DNM: Diffuse noise model; AE: Autoencoder; ReLU: Rectified linear unit; LeakyReLU: Leaky rectified linear unit; MSE: Mean square error; REHASP: REpeated HArvard Sentence Prompts; RMSE: Root mean square error; RPSL: Restricted peak search and link; ANOVA: Analysis of variance

Acknowledgements

We would like to give our thanks to the Acoustics Research Centre, University of Auckland, for supporting this project. We would also like to thank Alec Handyside, James Kennelly, Dylan Leslie, and Weichi Liu for their work and effort as a team in the development of the study's proposed method for the 2019 IEEE SPCup.

Authors' contributions

Benjamin Yen led the University of Auckland student team in the participation of the 2019 IEEE SPCup, including algorithm development, prototyping, and tuning. Benjamin Yen prepared the manuscript. Yusuke Hioka led the initiative of the study and supervised Benjamin Yen. Yusuke Hioka and Benjamin Yen reviewed edited the manuscript. All authors read and approved the manuscript.

Funding

This study was funded by the Acoustics Research Centre, University of Auckland.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the *DRone EGonoise and localizatiON dataset (DREGON)* repository, http://dregon.inria.fr/datasets/the-spcup19-egonoise-dataset/.

Competing interests

The authors declare that they have no competing interests.

Received: 29 May 2020 Accepted: 10 September 2020
Published online: 22 September 2020

References

1. 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. IEEE Signal Proc. Mag., 138–144 (2019)
2. R. Verrier, *Drones Are Providing Film and TV Viewers a New Perspective on the Action*. (Los Angeles Times, 2017). http://www.latimes.com/entertainment/envelope/cotown/la-et-ct-drones-hollywood-20151008-story.html
3. M. Margaritoff, *An English Lifeboat Crew Is Testing Drones for Search and Rescue*. (The Drive, 2017). http://www.thedrive.com/aerial/14528/an-english-lifeboat-crew-is-testing-drones-for-search-and-rescue
4. A. Charlton, *Police Drone to Fly Over New Year's Eve Celebrations in Times Square*. (Salon, 2018). https://www.salon.com/2018/12/31/police-drone-to-fly-over-new-years-eve-celebrations-in-times-square_partner
5. S. McCarthy, *Chinese Police Use Drone to Rescue Man Lost in XinJiang Desert*. (South China Morning Post, 2018). https://www.scmp.com/news/china/society/article/2168091/chinese-police-use-drone-rescue-man-lost-xinjiang-desert
6. A. Lusher, *Teenage Rape Victim Found by Police Drone with Thermal Imaging Camera*. (The Independent, 2018). https://www.independent.co.uk/news/uk/crime/police-drone-thermal-rape-victim-teenager-boston-lincolnshire-surveillance-technology-crime-fighting-a8572656.html
7. M. Ablon, *Crews: Both Climbers from Looking Glass Rock Rescued, One Taken to Hospital After Nearly 150-Foot Fall*. (FOX Carolina, 2019). https://www.foxcarolina.com/news/rescuers-searching-for-rock-climbers-at-looking-glass-rock/article_40faa27c-273e-11e9-8d7f-9f80965783a8.html
8. Y. Bando, H. Saruwatari, N. Ono, S. Makino, K. Itoyama, D. Kitamura, M. Ishimura, M. Takakusaki, N. Mae, K. Yamaoka, *et al*, Low latency and high quality two-stage human-voice-enhancement system for a hose-shaped rescue robot. J. Robot. Mechatron. **29**(1), 198–212 (2017)
9. Y. Hioka, M. Kingan, G. Schmid, R. McKay, K. A. Stol, Design of an unmanned aerial vehicle mounted system for quiet audio recording. Appl. Acoust. **155**, 423–427 (2019)
10. B. Yen, Y. Hioka, B. Mace, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information, (2018), pp. 545–549. https://ieeexplore.ieee.org/document/8521324. Accessed 19 Dec 2019
11. L. Wang, A. Cavallaro, in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Ear in the sky: ego-noise reduction for auditory micro aerial vehicles (IEEE, 2016), pp. 152–158. https://ieeexplore.ieee.org/document/7738063. Accessed 19 Dec 2019
12. L. Wang, A. Cavallaro, Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles. IEEE Sensors J. **17**(8), 2447–2455 (2017)
13. M. Brandstein, D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications, Digital Signal Processing*. (Springer, 2001). http://link.springer.com/10.1007/978-3-662-04619-7. Accessed 26 June 2017
14. P. Marmaroli, X. Falourd, H. Lissek, in *Acoustics 2012*, ed. by S. F. d'Acoustique. 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/document
15. K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, H. G. Okuno, in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Noise correlation matrix estimation for improving sound source localization by multirotor UAV, (2013), pp. 3943–3948. https://ieeexplore.ieee.org/document/6696920
16. K. Washizaki, M. Wakabayashi, M. Kumon, in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Position estimation of sound source on ground by multirotor helicopter with microphone array (IEEE, Daejeon, 2016), pp. 1980–1985. http://ieeexplore.ieee.org/abstract/document/7759312/. Accessed 29 June 2017
17. L. Wang, A. Cavallaro, Acoustic sensing from a multi-rotor drone. IEEE Sensors J. **18**(11), 4570–4582 (2018)
18. K. Okutani, T. Yoshida, K. Nakamura, K. Nakadai, in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter, (2012), pp. 3288–3293. https://ieeexplore.ieee.org/abstract/document/6385994. Accessed 19 Dec 2019
19. T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, K. Nakadai, in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array (IEEE, 2014), pp. 1902–1907. https://ieeexplore.ieee.org/document/6942813. Accessed 19 Dec 2019
20. K. Nakadai, M. Kumon, H. G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, O. Sugiyama, in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*

(IROS). Development of microphone-array-embedded UAV for search and rescue task, (2017), pp. 5985–5990. https://doi.org/10.1109/IROS.2017.8206494
21. M. Basiri, F. Schill, P. U. Lima, D. Floreano, in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Robust acoustic source localization of emergency signals from micro air vehicles, (2012), pp. 4737–4742. https://ieeexplore.ieee.org/document/6385608. Accessed 19 Dec 2019
22. M. Basiri, F. Schill, P. Lima, D. Floreano, On-board relative bearing estimation for teams of drones using sound. IEEE Robot. Autom. Lett. **1**(2), 820–827 (2016)
23. T. Ishiki, M. Kumon, in *2014 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. A microphone array configuration for an auditory quadrotor helicopter system, (2014), pp. 1–6. https://ieeexplore.ieee.org/document/7017653. Accessed 19 Dec 2019
24. T. Ishiki, M. Kumon, in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Design model of microphone arrays for multirotor helicopters, (2015), pp. 6143–6148. https://ieeexplore.ieee.org/document/7354252. Accessed 19 Dec 2019
25. T. Ishiki, K. Washizaki, M. Kumon, Evaluation of microphone array for multirotor helicopters. J. Robot. Mechatron. **29**(1), 168–176 (2017)
26. J. Choi, J. Chang, in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*. Convolutional neural network-based direction-of-arrival estimation using stereo microphones for drone, (Barcelona, 2020), pp. 1–5. https://ieeexplore.ieee.org/document/9051364
27. C. Blandin, A. Ozerov, E. Vincent, Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Process. **92**(8), 1950–1960 (2012). Latent Variable Analysis and Signal Separation
28. R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, H.-M. Park, Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments. IEEE Access. **8**, 7373–7382 (2020)
29. W. Zhang, Y. Zhou, Y. Qian, in *Proc. Interspeech 2019*. Robust DOA estimation based on convolutional neural network and time-frequency masking, (Graz-Austria, 2019), pp. 2703–2707. https://www.isca-speech.org/archive/Interspeech_2019/pdfs/3158.pdf
30. T. Gerkmann, R. C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Language Process. **20**(4), 1383–1393 (2012)
31. X. Li, S. Leglaive, L. Girin, R. Horaud, Audio-noise power spectral density estimation using long short-term memory. IEEE Signal Process. Lett. **26**(6), 918–922 (2019)
32. Z.-W. Tan, A. H.-T. Nguyen, A. W.-H. Khong, in *2019 Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*. An efficient dilated convolutional neural network for UAV noise reduction at low input SNR, (2019), pp. 1885–1892
33. Y. Hioka, M. Kingan, G. Schmid, K. A. Stol, in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. Speech enhancement using a microphone array mounted on an unmanned aerial vehicle, (2016), pp. 1–5. http://ieeexplore.ieee.org/abstract/document/7602937/. Accessed 24 June 2017
34. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976). https://doi.org/10.1109/TASSP.1976.1162830
35. I. McCowan, *Microphone Arrays: a Tutorial*. (Queensland University, Australia, 2001), pp. 1–38
36. H. Cox, R. M. Zeskind, M. M. Owen, Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1376 (1987)
37. C. Blandin, E. Vincent, A. Ozerov, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-source TDOA estimation using SNR-based angular spectra, (2011), pp. 2616–2619
38. B. Loesch, B. Yang, in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Blind source separation based on time-frequency sparseness in the presence of spatial aliasing, (2010), pp. 1–8
39. P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, in *Proceedings of the 25th International Conference on Machine Learning*. Extracting and composing robust features with denoising autoencoders (ICML, Helsinki, 2008), pp. 1096–1103. https://dl.acm.org/doi/abs/10.1145/1390156.1390294
40. P. Welch, The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoustics. **15**(2), 70–73 (1967)
41. A. L. Maas, A. Y. Hannun, A. Y. Ng, in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Rectifier nonlinearities improve neural network acoustic models (ICML, Atlanta, 2013). http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf
42. D. P. Kingma, J. Ba, in *International Conference on Learning Representations (ICLR)*. Adam: a method for stochastic optimization (ICLR, San Diego, 2015). https://arxiv.org/abs/1412.6980
43. Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, T. Kawahara, H. G. Okuno, Speech enhancement based on bayesian low-rank and sparse decomposition of multichannel magnitude spectrograms. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(2), 215–230 (2018)
44. M. Strauss, P. Mordel, V. Miguet, A. Deleforge, in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Dregon: dataset and methods for uav-embedded sound source localization (IROS, Madrid, 2018), pp. 1–8. https://ieeexplore.ieee.org/abstract/document/8593581/
45. F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, F. Michaud, The manyears open framework. Auton. Robot. **34**(3), 217–232 (2013)
46. G. E. Henter, T. Merritt, M. Shannon, C. Mayo, S. King, in *Fifteenth Annual Conference of the International Speech Communication Association*. Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech, (2014)
47. R. W. Sinnott, Virtues of the haversine. Sky Telesc. **68**, 159 (1984)
48. F. Curtin, P. Schulz, Multiple correlations and Bonferroni's correction. Biol. Psychiatry. **44**(8), 775–777 (1998)
49. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)

Publisher's Note