# Rotor noise power spectral density informed sound source enhancement and localisation for unmanned aerial vehicles

Benjamin Yen[*1] Yusuke Hioka[*2]

[*1] Tokyo Institute of Technology [*2] Acoustics Research Centre, University of Auckland

Recent years saw a significant increase in attention to using unmanned aerial vehicles (UAVs) to perform audio signal processing-related applications, such as sound source enhancement and localisation. However, the high levels of UAV rotor noise often result in extremely low input signal-to-noise ratios (SNR), thus rendering the problem highly challenging. This article reviews selected methods shown to perform well in these scenarios. The methods primarily rely on estimating the rotor noise power spectral densities (PSD) by utilising multi-sensory information and or machine learning to achieve the desired accuracy and robustness, thereby creating an effective postfilter or noise masking envelope to reduce the effects of rotor noise.

## 1. Introduction

Audio source enhancement and localisation have spanned decades of extensive research over many applications. Naturally, with the significant increase in the popularity of unmanned aerial vehicles (UAVs), research for UAV-specific applications has also been gaining increasing attention over the recent decade. For most studies, this involves attaching a microphone array [1] onto the UAV to perform audio recording. However, the proximity of the microphone array to the UAV and the high levels of UAV rotor noise make this problem a uniquely challenging task.

Effectively, the setting results in very low signal-to-noise ratios (SNR), for which many of the classic audio processing techniques are insufficient to be useful. Despite this, recent years saw several studies attempting to perform audio-related applications using UAVs, such as sound source localisation [2, 3, 4], sound source separation [5], and sound source enhancement [6, 7, 8, 9].

For sound source enhancement, a natural solution would involve using a noise mask or a noise filter to reduce the effects of rotor noise. For example, authors in [7] utilised a Kurtosis based noise estimator to design a noise mask, assisted by information regarding the sound source's location. A number of studies also utilised deep neural networks (DNN) to improve source enhancement performance, whether to design a noise mask to use for a postfilter [8] or to optimise beamformer steering [10].

As shown in the studies mentioned earlier, despite the multitude of different approaches, most require a denoising scheme for their methods to be effective. This is also well-demonstrated with the study in [6, 11], utilising the well-known *beamforming with Wiener postfilter* framework [1]. Here, the critical requirement for the method to perform effectively is to have accurate estimates of each sound source's power spectral densities (PSD). The study in [6] realised this using multiple beamformers to design a multi-channel Wiener postfilter (MWF), which was later extended by the studies [9, 12] with a multi-sensory, machine learning based rotor noise PSD estimator. In particular, using non-acoustical features yielded from the rotor's state improves the accuracy and robustness in estimating the rotor noise PSDs. This is demonstrated in [9] with its strong source enhancement performance with real-life experiments using a flying drone, which will be reviewed in this article along with its baseline from [6].

For sound source localisation, a common approach includes the use of the multiple signal classification algorithm (MUSIC), combined with denoising techniques, to achieve the desired performance [2, 3, 4]. Recent studies also showed the use of multiple UAVs to triangulate and improve localisation performance [13]. On the other hand, several studies have also shown the use of the generalised cross-correlation - phase transform (GCC-PHAT) method along with a denoising scheme [14, 15]. This was particularly apparent in the 2019 IEEE Signal Processing Cup, where many of the top participating teams made use of such a framework [16]. The study in [14] in particular demonstrated the effectiveness of using a DNN-driven noise envelope to improve source localisation performance by reducing the effects of rotor noise directly, which has shown to be particularly effective over many input scenarios. This method will also be reviewed in this article.

The rest of this article is organised as follows. First, studies in sound source enhancement for UAVs mentioned earlier will be reviewed in Section 2.. This includes the problem setup, the *beamforming with Wiener postfilter* framework proposed by [6], and extensions introduced in [9]. A sound source localisation algorithm that carries means to reduce rotor noise from [14] will be reviewed in Section 3.. Finally, the article is concluded with some comments and discussions in Section 4..

Contact: Benjamin Yen, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552, E-mail: benjamin@ra.sc.e.titech.ac.jp

Benjamin Yen is an International Research Fellow of the Japan Society for the Promotion of Science.

## 2. Sound source enhancement

This section reviews methods from studies [6, 9].

## 2.1 Problem setup

The problem considers a UAV, mounted with an $M$-sensor microphone array, receiving a target source $S(\omega, t)$, $K$ *spatially coherent* interfering noise sources $N_\theta(\omega, t)$ (including noise generated by $U$ ($\leq K$) UAV rotors) arriving from different angles $\theta$, and ambient *spatially incoherent* noise. The system aims to extract a clear target source signal from the $M$-channel noisy recordings [6].

The short-time Fourier transform (STFT) of the $M$-microphone input signals are expressed in vector form as

$$\begin{aligned}\mathbf{x}(\omega, t) &:= \Big[X_1(\omega, t), \quad \cdots, \quad X_M(\omega, t)\Big]^T \\ &= \mathbf{a}_{\theta_0}(\omega) S(\omega, t) + \sum_{u=1}^{U} \mathbf{a}_{\theta_u}(\omega) N_{\theta_u}(\omega, t) \\ &+ \sum_{n=U+1}^{K} \mathbf{a}_{\theta_n}(\omega) N_{\theta_n}(\omega, t) + \mathbf{v}(\omega, t), \end{aligned} \tag{1}$$

where $^T$ denotes the transpose, and $X_m(\omega, t)$ is the STFT of the $m$-th microphone's input signal. $\theta_0$, $\theta_u$ and $\theta_n$ indicate the angles to the target, the $u$-th rotor, and the $n$-th spatially coherent interfering noise source, respectively. $\omega = 1, \ldots, F$ and $t$ denote the angular frequency (of $F$ frequency bins) and frame index, respectively. $\mathbf{a}_\theta(\omega) = [A_{1,\theta}(\omega), \cdots, A_{M,\theta}(\omega)]^T$ and $\mathbf{v}(\omega, t) = [V_1(\omega, t), \cdots, V_M(\omega, t)]^T$ are the steering vector between the source located at angle $\theta$ and each microphone $m$, and the incoherent noise vector observed by the microphone array, respectively.

In the study in [9], given that both UAV rotor noise and spatially coherent interfering noise can be modelled as spatially coherent sources, $\mathbf{v}(\omega, t)$ is considered negligible for simplicity. In addition, sound sources are assumed to be mutually uncorrelated. For a UAV which typically operates in open outdoor environments, sound propagation is assumed to closely resemble a free field. Regardless, $A_{m,\theta}(\omega)$ is modelled as the transfer function between each sound source and microphone or impulse response (IR) measurements in practice. Furthermore, the problem is assumed to be limited to overdetermined cases, where $M \geq K + 1$. Finally, the problem assumes that the sound arrival angles of the target source and all noise sources are given *a priori*.

## 2.2 Beamforming with rotor noise informed Wiener postfilter

Figure 1 shows an overview of the source enhancement algorithm from [9]. Beamforming is a commonly used technique to perform source enhancement. In the studies [9, 12], $K + 1$ fixed beamformers are used, with the main lobe of each beamformer directed towards the angle of each sound source $\theta$ outlined in Section 2.1 (i.e. $\theta_0$ for the target, $\theta_u$ for the $u$-th rotor noise and $\theta_n$ for the $n$-th interfering noise source). The beamformer outputs $Y_\theta(\omega, t)$ are then calculated as

$$Y_\theta(\omega, t) = \mathbf{w}_\theta^H(\omega) \mathbf{x}(\omega, t), \tag{2}$$

where $\mathbf{w}_\theta(\omega) = [W_{1,\theta}(\omega), \cdots, W_{M,\theta}(\omega)]$ denotes the vector of the beamformer's filter weights and $^H$ denotes the Hermitian conjugate. The framework from [6] used the minimum variance distortionless response (MVDR) beamforming technique, where its weights $\mathbf{w}_\theta$ are obtained as [17]

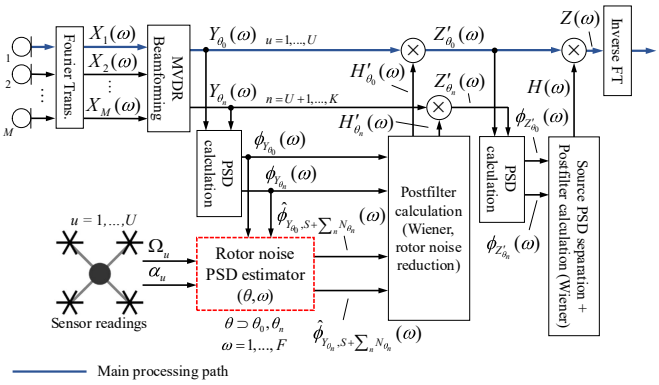


Figure 1: Overall framework of the beamforming with rotor noise informed postfilter method [9].

$$\mathbf{w}_\theta(\omega) = \frac{R^{-1}(\omega)\mathbf{a}_{\theta_{0'}}(\omega)}{\mathbf{a}_{\theta_{0'}}^H(\omega) R^{-1}(\omega) \mathbf{a}_{\theta_{0'}}(\omega)}, \tag{3}$$

where $\mathbf{a}_{\theta_{0'}}(\omega)$ denotes the steering vector of the chosen target source direction for the beamformer (i.e. the angle to which the directivity of the beamformer points). Assuming the measured IRs of each angle of interest is known, $R(\omega)$ is the normalised noise covariance matrix modelled using the steering vector of the $N'$ chosen noise source directions for the beamformer $\mathbf{a}_{\theta_{n'}}(\omega)$ (i.e. the "nulls" of the beamformer).

Often, beamforming is coupled with a postfilter to provide additional reduction of unwanted noise, especially when the number of microphones available is limited [18]. However, as mentioned in Section 1., this framework requires highly accurate estimates of the individual source PSDs to work well. This is particularly challenging due to the very low SNR levels in the UAV problem setting, leading to the target sources and other interfering noise sources being masked heavily by rotor noise. However, rotor noise is strongly correlated to the UAV's state characteristics, which opens a gateway to accurately estimate rotor noise PSDs (i.e. $\phi_{Y_\theta, N_{\theta_u}}(\omega, t)$), which can later be utilised to infer source PSDs of other types. The studies in [19, 20] leveraged this idea and utilised machine learning-based algorithms to estimate $\phi_{Y_\theta, N_{\theta_u}}(\omega, t)$, which was later utilised for source enhancement in [9, 12]. As a result, two Wiener filters (WF) are used. We note that PSDs from microphone signals from studies reviewed in this article are calculated using the Welch method [21].

The first postfilter is dedicated to suppressing rotor noise (hereafter referred to as $\text{WF}_{\text{rot}}$). Here, $\text{WF}_{\text{rot}}$ carries out rotor noise suppression directly on beamformer outputs pointing the target source $Y_{\theta_0}(\omega, t)$ and interfering noise source $Y_{\theta_n}(\omega, t)$, respectively. Here, $\text{WF}_{\text{rot}}$ is designed using rotor noise PSDs estimated by a rotor noise PSD estimation module. The module estimates the rotor noise PSDs for beamformer outputs that point towards the target source $\sum_{u=1}^{U} \hat{\phi}_{Y_{\theta_0}, N_{\theta_u}}(\omega, t)$ and interfering noise source $\sum_{u=1}^{U} \hat{\phi}_{Y_{\theta_n}, N_{\theta_u}}(\omega, t)$ via a machine learning-based

mapping function, taking the UAV's non-acoustical parameters as its input features (see Section 2.3). Using estimates $\sum_{u=1}^{U} \hat{\phi}_{Y_{\theta_0},N_{\theta_u}}(\omega,t)$ and $\sum_{u=1}^{U} \hat{\phi}_{Y_{\theta_n},N_{\theta_u}}(\omega,t)$, the beamformer output PSDs after removing rotor noise $\hat{\phi}_{Y_{\theta_0},S+\sum_n N_{\theta_n}}(\omega,t)$ and $\hat{\phi}_{Y_{\theta_n},S+\sum_n N_{\theta_n}}(\omega,t)$ are then obtained as

$$\hat{\phi}_{Y_\theta,S+\sum_{n=U+1}^{K} N_{\theta_n}}(\omega,t) = \phi_{Y_\theta} - \sum_{u=1}^{U} \hat{\phi}_{Y_\theta,N_{\theta_u}}. \tag{4}$$

Using these PSDs, $\mathrm{WF}_{\mathrm{rot}}$ is then calculated to reduce rotor noise in $\phi_{Y_{\theta_0}}(\omega,t)$ and $\phi_{Y_{\theta_n}}(\omega,t)$ as

$$H'_\theta(\omega,t) = \frac{\hat{\phi}_{Y_\theta,S+\sum_{n=U+1}^{K} N_{\theta_n}}}{\phi_{Y_\theta}}. \tag{5}$$

Using these Wiener filters, the rotor noise reduced output signals $Z'_{\theta_0}(\omega,t)$ and $Z'_{\theta_n}(\omega,t)$ are then obtained as

$$Z'_\theta(\omega,t) = H'_\theta(\omega,t) Y_\theta(\omega,t). \tag{6}$$

Following (6), a second stage postfiltering process $\mathrm{WF}_{\mathrm{int}}$ is used to further reduce interfering noise, using *PSD estimation in beamspace* [6]. Here, the corresponding PSDs of the multiple beamformers and sound sources are then represented in a set of equations as

$$\underbrace{\begin{bmatrix} \phi_{Z'_{\theta_0}} \\ \phi_{Z'_{\theta_{U+1}}} \\ \vdots \\ \phi_{Z'_{\theta_K}} \end{bmatrix}}_{\mathbf{\Phi}_{Z'_\theta}(\omega,t)} = \underbrace{\begin{bmatrix} |D_{0,\theta_0}|^2 & |D_{0,\theta_{U+1}}|^2 & \cdots & |D_{0,\theta_K}|^2 \\ |D_{U+1,\theta_0}|^2 & |D_{1,\theta_{U+1}}|^2 & \cdots & |D_{U+1,\theta_K}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |D_{K,\theta_0}|^2 & |D_{K,\theta_{U+1}}|^2 & \cdots & |D_{K,\theta_K}|^2 \end{bmatrix}}_{\mathbf{G}(\omega)} \underbrace{\begin{bmatrix} \phi_S \\ \phi_{N_{\theta_{U+1}}} \\ \vdots \\ \phi_{N_{\theta_K}} \end{bmatrix}}_{\mathbf{\Phi}_{S+N}(\omega,t)}, \tag{7}$$

where $\phi_{Z'_\theta}(\omega,t)$ are the PSDs calculated from $Z'_\theta(\omega,t)$ (i.e. outputs of (6)). Note that since rotor noise is removed beforehand, the rotor noise source $N_{\theta_u}(\omega,t)$ and consequently the beamformer output PSD pointing towards it $\phi_{Y_{\theta_u}}(\omega,t)$, need no longer to be considered.

Again, assuming that the measured IRs of each angle of interest are known, the matrix $\mathbf{G}(\omega)$ can be calculated beforehand. The source PSDs can then be estimated as

$$\mathbf{\Phi}_{S+N}(\omega,t) = \mathbf{G}^{-1}(\omega)\mathbf{\Phi}_{Z'_\theta}(\omega,t). \tag{8}$$

Following (8), $\mathrm{WF}_{\mathrm{int}}$ is designed using $\mathbf{\Phi}_{S+N}(\omega,t)$ is utilised to separate the target and coherent interfering noise sources, and thereby extracting the final output signal $Z(\omega,t)$. The weights of $\mathrm{WF}_{\mathrm{int}}$ is given as

$$H(\omega,t) = \frac{\phi_S(\omega,t)}{\phi_S(\omega,t) + \sum_{n=U+1}^{K} \phi_{N_{\theta_n}}(\omega,t)}. \tag{9}$$

Finally, $Z(\omega,t)$ is obtained as

$$Z(\omega,t) = H(\omega,t) Z'_{\theta_0}(\omega,t). \tag{10}$$

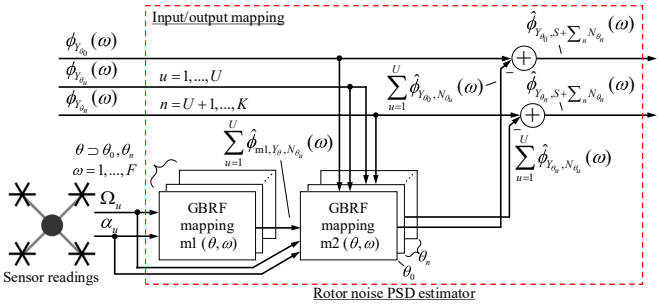


Figure 2: Rotor noise PSD estimation module [9], which contains the GBRF based input/output mapping functions with multi-sensory input information.

### 2.3 Rotor noise PSD estimator input/output mapping

In the study [9], the multi-sensory, machine learning-based rotor noise PSD estimator uses the gradient boosted random forest mapping function (GBRF). This is a combination of regression tree (RT), gradient boosting (GB), and random forest (RF) [22]. Details of its architecture are well described in common textbooks and in [9]. The GBRFs are prepared for $F$ independent frequency bins and each beamformer output (except those pointing towards the rotors), giving a total of $(1 + K - U) \times F$ GBRF models.

For the input feature used for the GBRFs in [9], two mapping configurations (m1 and m2, see Section 2.3) are used. Common features utilised by both configurations are **rotor speed** $(\Omega_u(t))$ and **rotor acceleration** $(\alpha_u(t))$. It is shown in [9] that rotor noise PSD consists of many tonal harmonics, of which its frequencies follow $\Omega_u(t)$, which means such information can be utilised to infer the rotor noise PSD without concerns from acoustical disturbances.

In addition to $\Omega_u(t)$ and $\alpha_u(t)$, m2 also utilises acoustic signals, specifically **beamformer output PSD** $(\phi_{Y_\theta}(\omega,t))$ and **output of m1** $(\sum_{u=1}^{U} \hat{\phi}_{\mathrm{m1},Y_\theta,N_{\theta_u}}(\omega,t))$. While $\phi_{Y_\theta}(\omega,t)$ contains disturbance from sources other than rotor noise, it is the closest representation of the true rotor noise PSD while the UAV is in operation. Therefore, it could serve as a useful reference to capture the details of rotor noise PSD $\sum_{u=1}^{U} \phi_{Y_\theta,N_{\theta_u}}(\omega,t)$. On the other hand, $\sum_{u=1}^{U} \hat{\phi}_{\mathrm{m1},Y_\theta,N_\theta}(\omega,t)$ provides an undisturbed estimate of the rotor noise PSD was well demonstrated to be an important input feature that supplements $\phi_{Y_\theta}(\omega,t)$ in [9].

More details concerning the sound source localisation method and its performance evaluation through in-flight experiments can be found in [9].

## 3. Sound source localisation

This section reviews the method from [14].

### 3.1 Problem setup

For the source localisation problem for UAVs, we slightly modify the problem setup from source enhancement. The UAV utilised in this section is also different from Section 2.. While still assuming a UAV system with a $M$-sensors, receiving a single target sound source, $K$ interfering *spatially*

*coherent* noise sources (including $U$ UAV rotors), and ambient *spatially incoherent* noise, the objective of the system here is to accurately locate the target sound source using the $M$-channel noisy recordings. Therefore, we re-express (1) to explicitly account for a source signal's direction of arrival (DOA) as

$$\begin{aligned}\mathbf{x}(\omega,t) &:= \Big[X_1(\omega,t), \quad \cdots, \quad X_M(\omega,t)\Big]^T \\ &= \mathbf{a}(\omega,\vec{\theta}_S)S(\omega,\vec{\theta}_S,t) \\ &+ \sum_{u=1}^{U}\mathbf{a}(\omega,\vec{\theta}_u)N(\omega,\vec{\theta}_u,t) \\ &+ \sum_{n=U+1}^{K}\mathbf{a}(\omega,\vec{\theta}_n)N(\omega,\vec{\theta}_n,t)+\mathbf{v}(\omega,t), \end{aligned} \tag{11}$$

where $\mathbf{a}(\omega,\vec{\theta}) = [A_1(\omega,\vec{\theta}),\cdots,A_M(\omega,\vec{\theta})]^T$ and $\mathbf{v}(\omega,t)$ are the vector of transfer functions between the source at DOA $\vec{\theta} = [\theta_{\text{el}},\theta_{\text{az}}]^T$ (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. $S(\omega,\vec{\theta}_S,t)$, $N(\omega,\vec{\theta}_u,t)$ and $N(\omega,\vec{\theta}_n,t)$ are the STFT of the target sound source at angle $\vec{\theta}_S$, the noise source coming from the $u$-th rotor at angle $\vec{\theta}_u$, and the $n$-th *spatially coherent interfering* noise source at angle $\vec{\theta}_n$, respectively. Like in Section 2.1, we assume $\mathbf{v}(\omega,t)$ is negligible. For the 3D problem, $\vec{\theta}_S$, $\vec{\theta}_u$ and $\vec{\theta}_n$ are expressed in spherical coordinates as

$$\vec{\theta}_S = [\theta_{S,\text{el}},\theta_{S,\text{az}}]^T, \quad \vec{\theta}_u = [\theta_{u,\text{el}},\theta_{u,\text{az}}]^T, \quad \vec{\theta}_n = [\theta_{n,\text{el}},\theta_{n,\text{az}}]^T. \tag{12}$$

The assumptions imposed in Section 2.1 are also applicable to the source localisation problem. In order to identify the directions of the target sound source, knowledge regarding the transfer function $\mathbf{a}(\omega,\vec{\theta})$ is required, which, unfortunately, is generally unavailable. Thus, we assume the UAV operates at some height above ground and is mostly open air. Therefore, the environment is approximately a free field, and that $\mathbf{a}(\omega,\vec{\theta})$ is assumed as the steering vector of a plane wave [6], described as $\mathbf{a}(\omega,\vec{\theta}) = \Big[e^{-j\omega\tau_{\vec{\theta},1}},\cdots,e^{-j\omega\tau_{\vec{\theta},M}}\Big]^T$, where $\tau_{\vec{\theta},m}$ is the TDOA at the $m$-th 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.

The problem in this section is based on the DRone EGonoise and localizatiON (DREGON) database [23], which considers three distinct tasks for the UAV and the target sound source:

Task 1. Hovering UAV - where both the target sound source and UAV are fixed in position throughout the audio recording.

Task 2. Flying (i.e. moving) UAV, broadband sound source - Here, the UAV is assumed to be moving while the target sound source (continuous broadband signal) remains fixed.

Task 3. Flying (i.e. moving) UAV, speech sound source - Here, the UAV is assumed to be moving while the target sound source (speech signal) remains fixed.

We assume that for tasks 2 and 3, the UAV moves relatively smoothly, such that there are no erratic variations in the rotor noise signature. In addition, the DREGON database only contains the target sound source and UAV rotor noise. Thus no additional coherent interfering noise sources exist (i.e. $K = U$).

### 3.2 Proposed method

Figure 3 shows a block diagram of the localisation method from [14]. The method is based on the method from [24], however with extensions proposed to address the very low SNR unique to the UAV problem setup. We first introduce the baseline method from [24] in Section 3.2.1, followed by the extensions and modifications made to the baseline method.

#### 3.2.1 Multi-source TDOA estimation in reverberant audio using angular spectra

This section outlines the baseline method from [24]. First, the SNR is calculated in the angular TDOA and time-frequency spectrum using pairs of microphones within the array, giving $K_{\text{p}} = {}_M C_2$ unique spectrum, which we refer it as the *SNR response*. Prior to calculating the SNR response, a mapping between a grid of TDOAs $\tau$ and a relevant range of $\vec{\theta}$ in the elevation and azimuth plane (i.e. the angular spectra) for each $k$-th microphone pair is established as follows

$$\tau_k(\theta_{\text{el}},\theta_{\text{az}}) = \frac{p_k \sin\left(\alpha_k\left(\theta_{\text{el}},\theta_{\text{az}}\right)\right)}{c_0}, \tag{13}$$

$$\alpha_k(\theta_{\text{el}},\theta_{\text{az}}) = \cos^{-1}\left(\frac{\mathbf{d}_k\left(\theta_{\text{el}},\theta_{\text{az}}\right)\cdot\Delta\mathbf{p}_k}{p_k}\right), \tag{14}$$

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 $k$-th pair of microphones in Cartesian coordinates and $p_k$ is the magnitude of the separating distance. Here, the target sound source is assumed to be located within this angular range of interest.

The baseline method from [24] provides several localisation techniques to calculate the SNR response (hereby denoted as $\psi_k(t,\omega,\tau_k)$ for each $k$-th microphone pair) for localisation. As such, the study in [14] explored methods such as the delay-and-sum [25] and MVDR beamforming techniques, as well as the generalised cross-correlation - phase transform (GCC-PHAT) [26]. In addition, a non-linear extension of GCC-PHAT proposed in [27], and a modified MVDR approach developed in [24] to improve robustness against diffuse ambient noise, named diffuse noise model, is also utilised in [14].

Following the calculation of $\psi_k(t,\omega,\tau_k)$, the SNR responses are aggregated together across the frequency bins, time frames, and the $K_{\text{p}}$ microphone pairs, to give an overall SNR response in terms of $\vec{\theta}$ (i.e. an angular spectrum). Aggregation across the frequency bins and the microphone pairs is carried out via summing, while time frames can

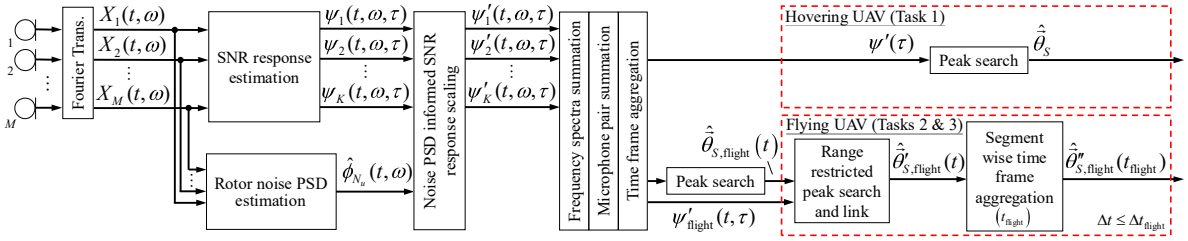


Figure 3: Block diagram of the proposed sound source localisation method in [14].

be summed or taken to the maximum. In [14], time frame aggregation depends on the task. In task 1 (i.e. hovering UAV), all $T_{\text{hover}}$ time frames are aggregated to give a single location estimate. This is done as the relative location between the microphone array, and the target sound source remains fixed. Aggregation for task 1 is calculated as

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

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

Subsequently, the overall SNR response for tasks 2 and 3 are calculated as

$$\psi'^{\text{sum}}_{\text{flight}}(t, \tau) = \sum_{t=1}^{T_{\text{flight}}} \sum_{k=1}^{K_{\text{p}}} \sum_{\omega=1}^{F} \psi_k(t, \omega, \tau_k), \tag{17}$$

$$\psi'^{\max}_{\text{flight}}(t, \tau) = \max_{t=1}^{T_{\text{flight}}} \sum_{k=1}^{K_{\text{p}}} \sum_{\omega=1}^{F} \psi_k(t, \omega, \tau_k). \tag{18}$$

The TDOA $\tau$ that gives the maximum overall SNR response from $\psi'(\tau)$, denoted $\hat{\tau}_S$ is then calculated as

$$\hat{\tau}_S = \underset{\tau}{\text{argmax}} \left( \psi'(\tau) \right), \tag{19}$$

$$\hat{\tau}_{S,\text{flight}}(t) = \underset{\tau}{\text{argmax}} \left( \psi'_{\text{flight}}(t, \tau) \right). \tag{20}$$

Finally, following (13) and (14) using $\hat{\tau}_S$ and $\hat{\tau}_{S,\text{flight}}(t)$, we obtain the source location in terms of angle for tasks 1 ($\hat{\vec{\theta}}_S$), 2 and 3 ($\hat{\vec{\theta}}_{S,\text{flight}}(t_{\text{flight}})$).

#### 3.2.2 Noise PSD informed SNR response scaling

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

While it is shown in [9] that adopting a machine learning-based rotor noise PSD estimator enables highly accurate PSD estimation, one of the challenges with the DREGON database is the limited amount of rotor noise data available for training. Therefore, conventional neural networks would not be suitable for the task. On the other hand, denoising autoencoders (DAE) learn a compressed representation of the uncorrupted input rather than a full mapping of the training data. Therefore, it can be used for feature extraction, and denoising [28]. Here, the input of the DAE is the PSDs from the microphone recordings $\phi_m(t, \omega)$ to output an accurate estimate of the rotor noise PSD $\phi_u(\omega)$. This is then used to create an envelope to scale and denoise the SNR response $\psi(t, \omega, \tau)$.

To form the encoder component of the DAE, the input audio PSDs $\phi_m(t, \omega)$ from each microphone are used to map towards the hidden representations $z$. Subsequently, the rotor noise PSD $\hat{\phi}_u(\omega)$ is reconstructed from $z$, which forms the decoder component of the DAE.

The size of the input audio PSD data is $T_{\text{DAE}} \times F$, where $T_{\text{DAE}} = 1$ corresponds to the number of PSD frames taken per observation. Details regarding the DAE architecture can be found in [14]. The DAE is optimised with respect to the mean squared error (MSE) between the output PSD $\hat{\phi}_u(t, \omega)$ and the true rotor noise PSD $\phi_u(t, \omega)$. To optimise MSE loss, the Adam optimiser is used [29]. The DAE is trained for each $m$ microphone channel, giving a total of $M$ DAEs for producing the SNR response scaling weighting envelope. To maximise noise removal effectiveness, the estimated PSD with the most prominent amplitude response out of the $M$ microphones for each frequency bin $\omega$ is selected and applied to scale the SNR responses for all $K_{\text{p}}$ microphone pairs. In addition, the estimated PSD frames are grouped and averaged to match the time frames for the localisation process.

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

$$\psi'_k(t, \omega, \tau) = \frac{\psi_k(t, \omega, \tau)}{\hat{\phi}_u(t, \omega)}. \tag{21}$$

This response then follows the aggregation process (15)-(18) to obtain the overall SNR response, which will be used to calculate $\hat{\vec{\theta}}_S$ using (13) and (14).

### 3.3 Angular spectral range restricted peak search and link

As discussed in Section 3.2.1, in tasks 2 and 3, time frame aggregation is carried out with smaller groups of frames $T_{\text{flight}}$, which potentially causes a loss in angular spectral resolution. To combat this issue, an angular spectral range restricted peak search and link post-processing algorithm (RPSL) is proposed in [9]. The algorithm is applied towards the localisation output $\hat{\vec{\theta}}_{S,\text{flight}}(t)$ before time frame aggregation is carried out (see Figure 3).

A flowchart describing the algorithm is shown in Figure 4. The algorithm carries out SNR response peak searching in the angular spectrum for several iterations to obtain the correct sound source travel path, which generally follows these main steps:

Step 1. Using localisation output $\hat{\vec{\theta}}_{S,\text{flight}}(t)$ as the reference

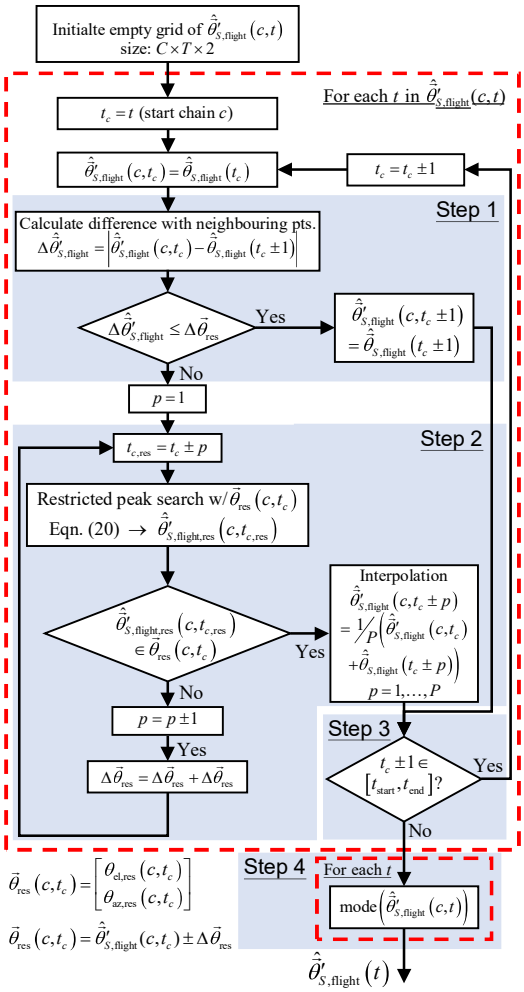


Figure 4: Flowchart of the RPSL algorithm. The **Steps** highlighted follows the descriptions in Section 3.3.

path of locations, check the difference $\Delta\hat{\vec{\theta}}'_{S,\text{flight}}$ between $\hat{\vec{\theta}}_{S,\text{flight}}(t)$ with respect to $\hat{\vec{\theta}}_{S,\text{flight}}(t-1)$ and $\hat{\vec{\theta}}_{S,\text{flight}}(t+1)$ for each time frame $t$.

Step 2. Perform restricted peak search using (20) with $\psi'_{\text{flight}}(t,\tau)$ (see Section 3.2.2) and $\vec{\theta}_{\text{res}}(c,t_c)$ (see 4) around time frames where $\Delta\hat{\vec{\theta}}'_{S,\text{flight}}$ exceeds threshold $\Delta\vec{\theta}_{\text{res}}$).

Step 3. Repeat Step 1 and Step 2 until no valid locations can be found, or if the start/end of the localisation path is reached (i.e. $t_c \pm 1 \notin [t_{\text{start}}, t_{\text{end}}]$). This forms the $c$-th "chain" of locations/local path (see Figure 4).

Step 4. Upon obtaining all $C$ chains of local paths, we find the final path $\hat{\vec{\theta}}'_{S,\text{flight}}(t)$ by selecting locations that appear most frequently amongst the $C$ chains at each $t$-th time frame.

Finally, the $T_{\text{flight}}$ time frames in $\hat{\vec{\theta}}_{S,\text{flight}}(t)$ are then aggregated together to obtain $\hat{\vec{\theta}}'_{S,\text{flight}}(t_{\text{flight}})$ (see Figure 3). Note that the threshold $\Delta\hat{\vec{\theta}}_{\text{res}}$ is heuristically tuned based on whether the estimated path of locations appears to be sensible overall (i.e. no aggressive jumps or unnatural changes in direction).

The RPSL post-processing algorithm is carried out in 2 second frame batches of $\hat{\vec{\theta}}_{S,\text{flight}}(t)$, except for the last batch, which would depend on the number of frames remaining. If the restricted peak search fails to obtain a valid location in a particular local path/time frame, the algorithm will skip the time frame and proceed to the next. Following this, the skipped locations are later obtained via interpolation between two valid time frames.

More details concerning the sound source localisation method, and its performance evaluation through experiments using the DREGON database can be found in [14].

## 4. Conclusions

This article has overviewed sound source enhancement and localisation techniques designed specifically for audio recording systems for UAVs using microphone arrays. Both application results in highly distinct algorithms, where source enhancement is based on beamforming with rotor noise informed postfiltering, and sound source localisation uses an extended multi-source TDOA estimation method. However, both applications share a common trait of having a dedicated means to reduce the influence of rotor noise on their respective received audio.

Regardless, UAV-specific signal processing remains a challenging task and has many aspects open to future studies. Methods to further improve rotor noise reduction and, more importantly, maintain or improve the output audio quality require much research in source enhancement. For source localisation, tracking moving sources more effectively also requires much work. For example, collecting more rotor noise data to better train the rotor noise PSD estimation DAE would drastically improve its performance. Utilising multiple UAVs to triangulate the localisation results (such as that proposed in [13]) would be another viable approach.

## References


[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing. Springer, 2001.

[2] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2013, pp. 3943–3948.

[3] K. Nakadai, M. Kumon, H. G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, and O. Sugiyama, "Development of microphone-array-embedded UAV for search and rescue task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2017, pp. 5985–5990.

[4] K. Yamada, M. Kumon, and T. Furukawa, "Belief-driven control policy of a drone with microphones for multiple

sound source search," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 5326–5332.

[5] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2016, pp. 1299–1304.

[6] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2016, pp. 1–5.

[7] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570–4582, June 2018.

[8] Z-W. Tan, A. H-T. Nguyen, and A. W-H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *2019 Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Nov. 2019, pp. 1885–1892.

[9] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Applied Acoustics*, vol. 189, pp. 108590, 2022.

[10] Y. Song, S. Kindt, and N. Madhu, "Drone ego-noise cancellation for improved speech capture using deep convolutional autoencoder assisted multistage beamforming," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–8.

[11] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," *Applied Acoustics*, vol. 155, pp. 423 – 427, 2019.

[12] B. Yen, Y. Hioka, and B. Mace, "Source enhancement for unmanned aerial vehicle recording using multi-sensory information," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 850–857.

[13] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Assessment of sound source tracking using multiple drones equipped with multiple microphone arrays," *International journal of environmental research and public health*, vol. 18, no. 17, pp. 9039, 2021.

[14] B. Yen and Y. Hioka, "Noise power spectral density scaled snr response estimation with restricted range search for sound source localisation using unmanned aerial vehicles," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–26, 2020.

[15] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 508–519, 2022.

[16] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: highlights from the ieee signal processing cup 2019 student competition [sp competitions]," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 138–144, 2019.

[17] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug 1969.

[18] Y. Hioka and K. Niwa, "Estimating power spectral density for spatial audio signal separation: An effective approach for practical applications," *Acoustical Science and Technology*, vol. 38, no. 4, pp. 175–184, 2017.

[19] B. Yen, Y. Hioka, and B. Mace, "Estimating power spectral density of unmanned aerial vehicle rotor noise using multisensory information," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2434–2438.

[20] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 545–549.

[21] 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 Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, June 1967.

[22] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2001.

[23] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[24] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950 – 1960, Aug. 2012, Latent Variable Analysis and Signal Separation.

[25] I. McCowan, "Microphone arrays: A tutorial," *Queensland University, Australia*, pp. 1–38, 2001.

[26] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[27] B. Loesch and B Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 1–8, 2010.

[28] P. Vincent, H. Larochelle, Y. Bengio, and P-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, Dec. 2014.