Source enhancement for unmanned aerial vehicle recording using multi-sensory information

Benjamin Yen* and Yusuke Hioka* and Brian Mace*
* Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, Auckland, 1010 New Zealand
E-mail: benjamin.yen@ieee.org, yusuke.hioka@ieee.org, b.mace@auckland.ac.nz

***Abstract*—A method to effectively capture desired sound signals from an unmanned aerial vehicle (UAV)-mounted audio recording system by utilising accurate rotor noise power spectral density (PSD) estimations of a UAV is proposed. The method seeks to improve the estimation accuracy and robustness of rotor noise PSD by incorporating UAV rotor characteristics in conjunction with microphone signals. Experiment results show rotor noise PSD estimation accuracy to within 5.5 dB log spectral distortion regardless of the presence of surrounding sound sources, with consistent ∼28 dB improvement in signal-to-noise ratio, in particular, reduction of rotor noise from the noisy microphone recordings.**

***Index Terms*—Microphone array, unmanned aerial vehicle, source enhancement, power spectral density, rotor noise**

I. Introduction

Unmanned aerial vehicles (UAVs) have shown a significant increase in popularity over the past few years across a range of applications, such as filming [1], search and rescue [2], and more recently, security and surveillance [3]. Such applications take advantage of capturing visual information (i.e. video and imagery) that are otherwise impossible without making use of UAVs. Audio capturing, on the other hand, remains a challenging task, due to the high levels of noise radiated by the UAVs rotors, as well as environmental noise such as wind or traffic.

Numerous studies attempt to perform clear audio extraction or related applications using UAVs. These include sound source localisation, [4], [5], [6], [7], [8], [9], or sound source separation [10], for which most studies make use of a single or an array of microphones [11] mounted on the UAV. Recently, there has been an increase in attention on studies focusing on directly improving audio recording quality [12], [13], [14]. Among these, one such series of studies was presented by the authors in [15], [16], utilising the well-known *beamforming with Wiener postfilter* framework [11]. By designing a Wiener filter via accurate estimation of each source's power spectral density (PSD), the desired signals can be extracted from its noisy mixture. However, these estimated PSDs have to be as accurate as possible, which is a challenging task for approaches purely dependent on acoustical information, as each microphone would receive a mixture of sound arriving from both the desired and undesired sources. Therefore, a means of predicting the PSDs that is uninfluenced by these practical constraints is necessary, namely, by using non-acoustical information.

Fortunately, rotor noise possesses a relatively structured and predictable behaviour, with aerodynamic and aeroacoustic studies showing that there exists a strong correlation between rotor noise and the characteristics of the rotor's behaviour such as rotor speed [17], [18]. In addition, some studies have attempted to suppress rotor noise by utilising various non-acoustical parameters [4], [14], [19]. Therefore, with appropriate utilisation of sensors, these parameters can be measured and, most importantly, they are immune to acoustical interferences. However, when it is used for source enhancement, such parameters require mapping to their desired output (i.e. rotor noise PSD), and typically such a mapping function would be a non-linear process.

A recent trend to model complex, non-linear, input-output mapping that has gained much attention is machine learning-based methods. Many works have been carried out utilising such techniques, such as deep neural networks (DNN) for source separation [20], [21], multichannel speech enhancement [22], [23] and speech recognition [24], [25], with considerable performance improvement over conventional methods. Estimating the PSD of sound sources from beamformer outputs using DNN [26], [27] has also shown promising results.

In the case of rotor noise PSD estimation, studies carried out in [28] and [29] combat the practical challenges seen in [15] by taking both the UAV rotor's state and microphone signals (i.e. acoustical and non-acoustical information) into account, and utilising machine learning approaches such as regression trees (RT) [30] and DNN as its mapping function. In this study, we extend these two studies by 1) further developing the rotor noise PSD estimation algorithm towards a source enhancement problem with a multi-sensory framework and, 2) evaluate the algorithm with a one-rotor practical experiment setup against rotor noise PSD accuracy and source enhancement performance.

II. UAV System and Problem Setup

Fig. 1 shows an overview of the audio recording UAV, including the microphone array setup used in this study.

A. Problem Setup

The problem assumes a UAV-mounted 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

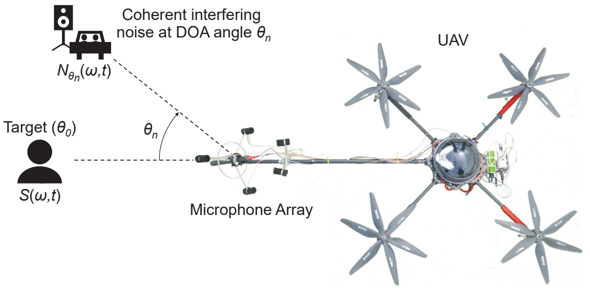

Fig. 1: Audio recording UAV overview (top view).

angles, and ambient *spatially incoherent* noise. The system aims to extract a clear target source signal from the M-channel noisy recordings [15]. The short-time Fourier transform (STFT) of the array's input signals are expressed in vector form as

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

where ^{T} denotes the transpose, and $X_m(\omega,t)$ is the STFT of the m-th microphone's input signal. θ_0, θ_u and θ_n indicate the angles to the target, the u-th rotor, and the n-th spatially coherent interfering noise source, respectively. ω and t denote the angular frequency (of F frequency bins) and frame index. $\mathbf{a}_\theta(\omega) = [A_{1,\theta}(\omega), \cdots, A_{M,\theta}(\omega)]^T$ and $\mathbf{v}(\omega,t) = [V_1(\omega,t), \cdots, V_M(\omega,t)]^T$ are the steering vectors between the source located at angle θ and each microphone m, and the incoherent noise vector observed by the microphone array, respectively.

Given each of the target source and all noise sources are essentially unique, the problem setup assumes the sound sources to be mutually uncorrelated. Since the most common flight scenarios for a UAV are outdoor environments, sound propagation is assumed to be a free field. Regardless, $A_{m,\theta}(\omega)$ is modelled as the transfer function between each sound source and microphone. Furthermore, it is assumed that the problem is 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*.

B. Source enhancement using beamforming with postfiltering

This section briefly explains the beamforming with postfiltering framework from [15] that is adapted in this study. First, the input signals are filtered via fixed beamformers, with the mainlobe of each directed towards the angle of a particular sound source θ (i.e. θ_0, θ_u and θ_n). 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 $\mathbf{w}_\theta(\omega)$ denotes the vector of the beamformer's filter weights and ^{H} denotes the Hermitian conjugate.

Following the beamformer, the noise source's signals are further reduced via a Wiener postfilter by using the PSD estimates of the target source and the other noise sources. From (1) and (2), the PSD of the beamformer outputs can be approximated as

$$\begin{aligned}\phi_{Y_\theta}(\omega,t) &\cong \phi_{Y_\theta,S}(\omega,t) + \sum_{u=1}^{U}\phi_{Y_\theta,N_{\theta_u}}(\omega,t) \\ &+ \sum_{n=U+1}^{K}\phi_{Y_\theta,N_{\theta_n}}(\omega,t) + \phi_{Y_\theta,\bar{V}}(\omega,t),\end{aligned} \tag{3}$$

where $\phi_{Y_\theta,S}(\omega,t)$, $\phi_{Y_\theta,N_{\theta_u}}(\omega,t)$, $\phi_{Y_\theta,N_{\theta_n}}(\omega,t)$ and $\phi_{Y_\theta,\bar{V}}(\omega,t)$ are the PSDs of the target source, the u-th UAV rotor noise, the n-th spatially coherent interfering noise and the incoherent noise, respectively. These PSDs are calculated by using the Welch method [31] given by $\phi_\mathcal{X}(\omega,t) = \lambda\phi_\mathcal{X}(\omega,t-1) + (1-\lambda)|\mathcal{X}(\omega,t)|^2$, where λ is the forgetting factor, and $\mathcal{X}(\omega,t)$ denotes the STFT of an arbitrary signal. As this study focuses on reducing UAV rotor noise in recordings which can be modelled as spatially coherent sources, $\phi_{Y_\theta,\bar{V}}(\omega,t)$ is considered negligible for simplicity. The Wiener filter coefficients are then estimated as

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

Note that ω and t are omitted for brevity in (4) and also for the rest of this paper unless otherwise specified. The operator $\hat{\cdot}$, in this case, denotes an estimate. Finally, the postfilter output signal $Z(\omega,t)$ is obtained as

$$Z(\omega,t) = H(\omega,t)Y_{\theta_0}(\omega,t). \tag{5}$$

III. Proposed Method

As mentioned in Section I, this study seeks a source enhancement algorithm using the rotor noise PSD $\phi_{Y_\theta,N_{\theta_u}}(\omega,t)$ predicted from non-acoustical information (i.e. UAV rotor's state) for target signal extraction. To model the non-trivial relationship between the UAV rotor's state to the rotor noise PSD, a machine learning-based approach is proposed.

This section introduces a source enhancement framework that incorporates the non-acoustically estimated PSDs, followed by the mapping function used to estimate the rotor noise PSDs and its input feature preparation.

A. General framework

Fig. 2 shows a block diagram of the proposed framework. The framework is an extension to that of the study from [15], using the *beamforming with postfiltering* framework. In this study, the minimum variance distortionless response (MVDR) beamformer technique [32] was used, with $\mathbf{w}_\theta(\omega)$ given by

$$\mathbf{w}_\theta(\omega) = \frac{R^{-1}(\omega)\mathbf{a}_\theta(\omega)}{{\mathbf{a}_\theta}^H(\omega)R^{-1}(\omega)\mathbf{a}_\theta(\omega)}, \tag{6}$$

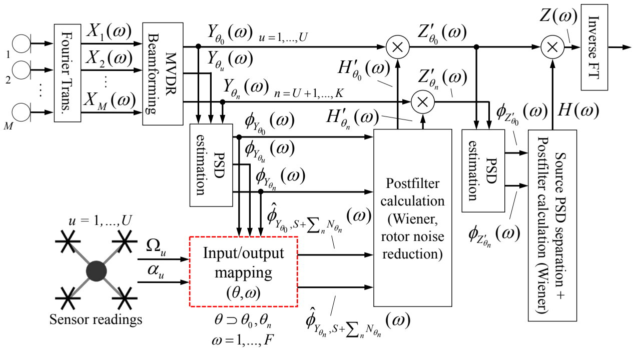
1
2
M
Fourier Trans.
$X_1(\omega)$
$X_2(\omega)$
$X_M(\omega)$
MVDR Beamforming
$Y_{\theta_0}(\omega)$ $u=1,\dots,U$
$Y_{\theta_u}(\omega)$
$Y_{\theta_n}(\omega)$ $n=U+1,\dots,K$
$H'_{\theta_0}(\omega)$
$H'_{\theta_n}(\omega)$
$Z'_{\theta_0}(\omega)$
$Z'_{\theta_n}(\omega)$
$Z(\omega)$
Inverse FT
PSD estimation
$\phi_{Y_{\theta_0}}(\omega)$
$\phi_{Y_{\theta_u}}(\omega)$
$\phi_{Y_{\theta_n}}(\omega)$
$\hat{\phi}_{Y_{\theta_0},S+\sum_n N_{\theta_n}}(\omega)$
$\hat{\phi}_{Y_{\theta_n},S+\sum_n N_{\theta_n}}(\omega)$
Postfilter calculation (Wiener, rotor noise reduction)
PSD estimation
$\phi_{Z'_{\theta_0}}(\omega)$
$\phi_{Z'_{\theta_n}}(\omega)$
$H(\omega)$
Source PSD separation + Postfilter calculation (Wiener)
$u=1,\dots,U$
Ω_u
α_u
Input/output mapping (θ,ω)
Sensor readings
$\theta\supset\theta_0,\theta_n$
$\omega=1,\dots,F$

Fig. 2: Overall framework of the proposed method. Rotor noise removal follows (9) to (12), and the final output signal extraction follows (13) to (16).

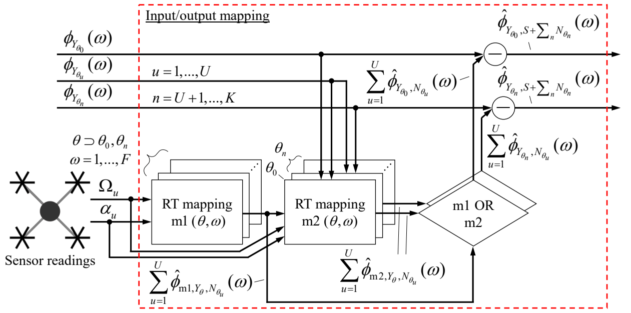
Input/output mapping
$\phi_{Y_{\theta_0}}(\omega)$
$\phi_{Y_{\theta_u}}(\omega)$
$\phi_{Y_{\theta_n}}(\omega)$
$u=1,\dots,U$
$n=U+1,\dots,K$
$\sum_{u=1}^{U}\hat{\phi}_{Y_{\theta_0},N_{\theta_u}}(\omega)$
$\hat{\phi}_{Y_{\theta_0},S+\sum_n N_{\theta_n}}(\omega)$
$\hat{\phi}_{Y_{\theta_n},S+\sum_n N_{\theta_n}}(\omega)$
$\sum_{u=1}^{U}\hat{\phi}_{Y_{\theta_n},N_{\theta_u}}(\omega)$
$\theta\supset\theta_0,\theta_n$
$\omega=1,\dots,F$
θ_n
θ_0
Ω_u
α_u
RT mapping m1 (θ,ω)
RT mapping m2 (θ,ω)
m1 OR m2
Sensor readings
$\sum_{u=1}^{U}\hat{\phi}_{\mathrm{m1},Y_\theta,N_{\theta_u}}(\omega)$
$\sum_{u=1}^{U}\hat{\phi}_{\mathrm{m2},Y_\theta,N_{\theta_u}}(\omega)$

Fig. 3: Input/output mapping of the proposed method, taking rotor speed, acceleration, and beamformer output PSDs as input information. The rotor noise removed PSD outputs are obtained using (7) and (8).

assuming the free field assumption holds and R is the normalised noise covariance matrix modelled using $A_{m,\theta}(\omega)$ mentioned in Section II-A. However, different to that from [15], a two-stage postfiltering system is used, with rotor noise and interfering noise source suppression carried out separately. The method incorporates a rotor noise PSD estimation module, highlighted by the red dashed box in Fig. 2, with details given in Fig. 3 and Section III-C. The module estimates the rotor noise PSDs in the output of the beamformer pointing 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 mapping function with the UAV's non-acoustical parameters (see Section III-C). Using these estimates, 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

$$\hat{\phi}_{Y_{\theta_0},S+\sum_{n=U+1}^{K}N_{\theta_n}}(\omega,t)=\phi_{Y_{\theta_0}}-\sum_{u=1}^{U}\hat{\phi}_{Y_{\theta_0},N_{\theta_u}}, \tag{7}$$

$$\hat{\phi}_{Y_{\theta_n},S+\sum_{n=U+1}^{K}N_{\theta_n}}(\omega,t)=\phi_{Y_{\theta_n}}-\sum_{u=1}^{U}\hat{\phi}_{Y_{\theta_n},N_{\theta_u}}, \tag{8}$$

respectively. These PSDs are used to design the Wiener filter for reducing rotor noise in $\phi_{Y_{\theta_0}}(\omega,t)$ and $\phi_{Y_{\theta_n}}(\omega,t)$, which are given by

$$H'_{\theta_0}(\omega,t)=\frac{\hat{\phi}_{Y_{\theta_0},S+\sum_{n=U+1}^{K}N_{\theta_n}}}{\phi_{Y_{\theta_0}}}, \tag{9}$$

$$H'_{\theta_n}(\omega,t)=\frac{\hat{\phi}_{Y_{\theta_n},S+\sum_{n=U+1}^{K}N_{\theta_n}}}{\phi_{Y_{\theta_n}}}, \tag{10}$$

respectively. The postfilter output signals after rotor noise reduction $Z'_{\theta_0}(\omega,t)$ and $Z'_{\theta_n}(\omega,t)$ are then obtained as

$$Z'_{\theta_0}(\omega,t)=H'_{\theta_0}(\omega,t)Y_{\theta_0}(\omega,t), \tag{11}$$

$$Z'_{\theta_n}(\omega,t)=H'_{\theta_n}(\omega,t)Y_{\theta_n}(\omega,t). \tag{12}$$

The PSDs of $Z'_{\theta_0}(\omega,t)$ and $Z'_{\theta_n}(\omega,t)$ are then utilised to perform a second stage postfiltering process. The PSD estimation in this stage makes use of the *PSD estimation in beamspace*, identical to that used in [15], to obtain $\Phi_{S+N}(\omega,t)=[\phi_S,\phi_{N_{\theta_{U+1}}},\dots,\phi_{N_{\theta_K}}]^T$, where ϕ_S and $\phi_{N_{\theta_n}}$ represents the estimated target source and interfering noise source PSD, respectively. Given the rotor noise is removed beforehand, the method assumes that the instantaneous PSD of the beamformer outputs can be approximated as

$$\begin{aligned}\phi_{Z'_\theta}(\omega,t)\cong{}&G_{0,\theta_0}(\omega)\phi_S(\omega,t)\\&+\sum_{n=U+1}^{K}G_{n,\theta_n}(\omega)\phi_{N_{\theta_n}}(\omega,t),\end{aligned} \tag{13}$$

where $\phi_{Z'_\theta}(\omega,t)$ are the PSDs from $Z_\theta(\omega,t)$. $G_{0,\theta_0}(\omega)$ and $G_{n,\theta_n}(\omega)$ correspond to $|D_{0,\theta_0}(\omega)|^2$ and $|D_{0,\theta_0}(\omega)|^2$, where $D_{0,\theta_0}(\omega)$ and $D_{n,\theta_n}(\omega)$ are the directivity of the beamformer to the angle of the target θ_0 and interfering noise sources θ_n, defined as

$$D_{0,\theta_0}(\omega)=\mathbf{w}_{0,\theta_0}^H(\omega)\mathbf{a}_{0,\theta_0}(\omega), \tag{14}$$

$$D_{n,\theta_n}(\omega)=\mathbf{w}_{n,\theta_n}^H(\omega)\mathbf{a}_{n,\theta_n}(\omega), \tag{15}$$

respectively. These estimated PSDs are then used to design another Wiener filter for separating the target and coherent interfering noise sources that extracts the final output signal $Z(\omega,t)$, with the weights given as

$$H(\omega,t)=\frac{\hat{\phi}_S(\omega,t)}{\hat{\phi}_S(\omega,t)+\sum_{n=U+1}^{K}\hat{\phi}_{N_{\theta_n}}(\omega,t)}. \tag{16}$$

Overall, the source enhancement framework is *multi-sensory*, utilising non-acoustical information during rotor noise reduction, to combat source leakage and allow increased effectiveness in reducing the remaining interfering noise sources.

B. Input/output mapping

The study in [29] showed that RT and DNN were both equally effective techniques for mapping input features to the rotor noise PSD $\sum_{u=1}^{U} \phi_{Y_\theta, N_{\theta_u}}(\omega, t)$. However, given the less computationally intensive nature of RTs and the shorter training time, it is the only mapping function used in this study. RT is a non-parametric regression technique where, given a set of observations, splits the input space into an expanding sequence of partitions via recursive binary subdivision. By optimising against a given objective function, an optimal sequence of partitioning is found [30], resembling a tree-like structure. In this study, the RTs are optimised with respect to the mean square error (MSE) between the true and estimated rotor noise PSD. The RTs are first grown, followed by pruning via a separate validation dataset to avoid overfitting. The validation dataset contains PSD similar in specifications to the training data. However, they are recorded separately with non-overlapping signals. The RTs are prepared for each independent frequency bin and beamformer output, giving a total of $(1 + K - U) \times F$ RT models. Note that since the primary objective of this study is to extract clear target signals from the noisy recordings, rotor noise PSDs are only estimated for the output PSDs of the beamformer that points its mainlobe towards the target ($\phi_{Y_{\theta_0}}(\omega, t)$) and interfering noise ($\phi_{Y_{\theta_n}}(\omega, t)$) sources.

C. Input feature configurations

The two mapping configurations ($\mathrm{m1}$ and $\mathrm{m2}$, see Fig. 3) follow that from the study in [28] with minor simplifications made. The following describes the motivation behind each configuration and their input features.

Configuration $\mathbf{m1}$: $\mathrm{m1}$ utilises rotor speed and acceleration ($\Omega_u(t)$, $\alpha_u(t)$), that are captured via speed sensors, as input features.

Configuration $\mathbf{m2}$: in addition to using $\Omega_u(t)$ and $\alpha_u(t)$, $\mathrm{m2}$ also utilises acoustic information. Namely, the output PSD of the beamformer that points its mainlobe towards the UAV rotors $\phi_{Y_{\theta_u}}(\omega, t)$ is used, as well its rate of change $\Delta\phi_{Y_{\theta_u}}(\omega, t)$, which was shown to be useful input features in simulations from [28] and [29]. Previous frames of $\phi_{Y_{\theta_u}}(\omega, t)$ and $\Delta\phi_{Y_{\theta_u}}(\omega, t)$ are also used in hoping to capture temporal changes of the PSD spectra. However, as $\phi_{Y_{\theta_u}}(\omega, t)$ is derived from microphone recordings (i.e. acoustical signals), it would also contain target and interfering noise. To remedy this, the output of $\mathrm{m1}$ ($\sum_{u=1}^{U} \hat{\phi}_{\mathrm{m1}, Y_\theta, N_{\theta_u}}(\omega, t)$) serves as a supplementary feature for $\phi_{Y_{\theta_u}}(\omega, t)$.

IV. Experiments

A. Experiment setup

The performance of the proposed method was evaluated via experiments based on the UAV used in the previous study [15], with the setup shown in Fig. 4. The UAV system utilised an array of six unidirectional microphones mounted in the plane of the UAV rotors. It is divided into two sub-arrays. The front sub-array is a circular array of three unidirectional microphones with a centre shotgun microphone, and the back sub-array consists of two unidirectional microphones. As opposed to the front-array, where all microphones point to the front of the UAV, the microphones in the back sub-array are oriented to point towards the UAV rotors.

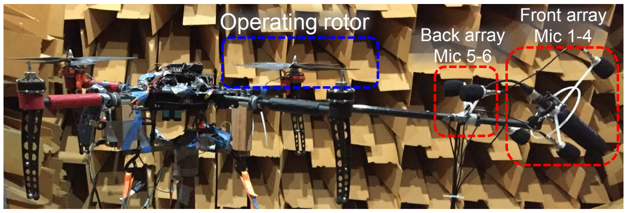

Fig. 4: UAV system and microphone array setup for experimental measurements.

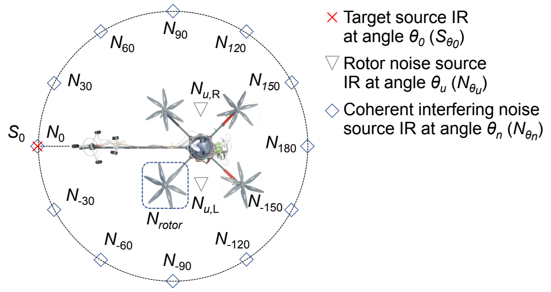
N_{90}
N_{60}
N_{120}
N_{30}
$N_{u,\mathrm{R}}$
N_{150}
S_0
N_0
N_{180}
N_{-30}
$N_{u,\mathrm{L}}$
N_{-150}
N_{rotor}
N_{-60}
N_{-120}
N_{-90}
× Target source IR at angle θ_0 (S_{θ_0})
▽ Rotor noise source IR at angle θ_u (N_{θ_u})
◇ Coherent interfering noise source IR at angle θ_n (N_{θ_n})

Fig. 5: Impulse response measurement positions and sound source locations.

In order to establish a controlled environment for audio data measurements that mimic the free field assumption as close as possible, all recordings were made in an anechoic chamber. However, due to the size of the UAV relative to the anechoic chamber, the UAV was mounted in a fixed position roughly at the centre of the anechoic chamber, as shown in Fig. 4. The measured noise was then mixed with a corpus of target source and interfering noise patterns that were produced by convolution of the sources with the impulse response (IR) measurements in the configuration shown in Fig. 5, except with rotor noise which was directly recorded in the chamber separately. The beamformers were configured by the specifications shown in Table I.

To capture the speed data, a custom built speed sensing module was developed to achieve the required sampling rate and resolution (see Fig. 6).

B. Experiment parameters

In this study, we focus on enhancing the target source signal while removing an interfering noise source. As a result, rotor noise PSD estimation is carried out for beamformer outputs $Y_{\theta_0}(\omega, t)$ and $Y_{\theta_n}(\omega, t)$, with microphone usage specification

TABLE I: Beamformer specifications used in experiments.

Beamformer	Microphone	Pole	Nulls
0	$1, 2, 3, 4$	P_0	$\mathrm{P}_{u,\mathrm{L}}$, P_{θ_1}
1	$1, 5, 6$	$\mathrm{P}_{u,\mathrm{L}}$	P_0, P_{θ_1}
2	$1, 2, 3, 4$	P_{θ_1}	$\mathrm{P}_{u,\mathrm{L}}$, P_0

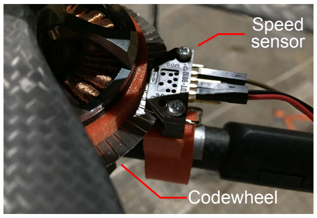

Fig. 6: Speed sensing module.

TABLE II: Experiment parameter specifications.

Sampling rate	48 kHz
STFT length (overlap shift)	2048 (1024)
Forgetting factor λ for ϕ_{Y_θ}, $\phi_{Z'_\theta}$ calc. [31]	0.3
# of beamformers	3

for each beamformer outlined in Fig. 4. As this paper presents a conceptual study, some assumptions are made to simplify the experiment setup. First, the target source is limited to speech, and secondly, all sound sources are assumed to exist in an environment with near free-field conditions. Lastly, as an initial proof of concept investigation, only one rotor (rotor 1, see Fig. 4) is used.

Table II summarises the audio processing parameters used. For observed microphone array signals, the target and noise sources are mixed and prepared under two metrics: i) signal-to-*rotor*-noise-ratio (SRNR) and ii) signal-to-*interfering*-noise-ratio (SINR) [15], which quantifies the power ratio of the target source to the rotor noise, and the coherent interfering noise sources, respectively. These are measured based on microphone 1 from the front sub-array (see Fig. 4). Tests are also conducted for different interfering noise source angles θ_n (see Fig. 5). Table III summarises the training and testing data preparation for the RTs of the proposed method.

The performance of the proposed method was evaluated under several measures. Firstly, the accuracy of the rotor noise PSD produced by the proposed method was evaluated against the previous study [15]. To achieve this, the log spectral distortion (LSD) [33] between the estimated rotor noise PSD and the relative true rotor noise PSD (i.e. without target or interfering noise) is measured and compared. To evaluate source enhancement improvement, similar to the input noise conditions prepared for testing i) SRNR and ii) SINR improvement [15], were measured and compared.

TABLE III: Experiment data specifications (* denotes m2 only, ** denotes for PSD accuracy evaluation only, *** i.e. rotor noise only).

	Training	Testing
UAV speed range (rpm)	3000-4000	3000-3300
# of target source patterns	12 (6 male, 6 female)*	4 (2 male, 2 female)
# of interfering noise patterns	8 (4 traffic, 4 music)*	4 (2 traffic, 2 music)
# of interfering noise angles	12	12
Input SRNR (dB)	$-\infty^{*,***}$, -10*, 0*, 10*	-30, -20, -10, 0, 10**
Input SINR (dB)	$-\infty^{*,***}$,-10*, 5*, 20*	-10, 0, 10, 20
# of datasets	130	672 + 128**
Total # of observations (per beamformer)	1271013	185472 + 35328**

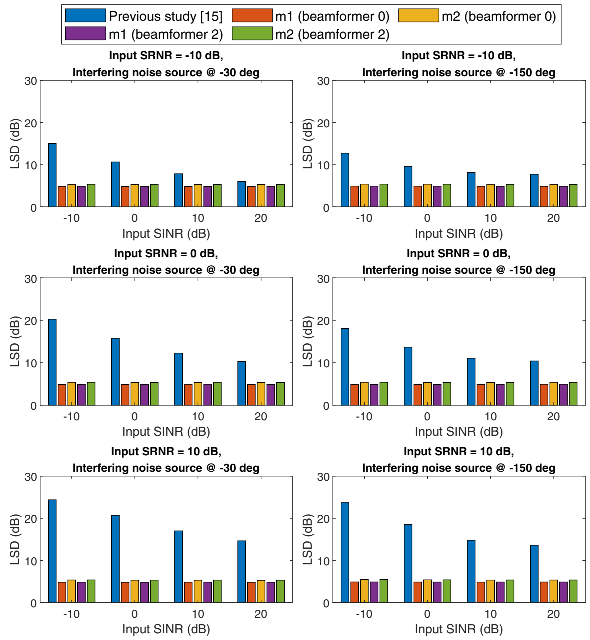

Fig. 7: Rotor noise PSD estimation accuracy: LSD for different input SRNR (each row of the graphs) and SINR (horizontal axis of each graph). See Table I for details of the specified beamformers.

V. Results and Discussion

A. Evaluation of PSD estimation accuracy

Fig. 7 summarises the results for different input SRNRs. Note that different from the studies [28] and [29], results were evaluated against real-world measurements as opposed to numerical simulations. In the following discussion, m1 and m2 will be referred to the two proposed input configurations as described in Section III-C unless otherwise specified.

For results shown in Fig. 7, different to that seen from numerical simulations under ideal condition presented in [28] and [29], the previous study [15] gave a much higher LSD, indicating larger PSD estimation error than that from the proposed method. This is perhaps due to the deviations of the practical environment from an ideal free-field. Fig. 7 shows that higher LSD figures are seen as the SRNR increases or the SINR decreases. This is expected as the interfering noise becomes more dominant in comparison to the rotor noise, causing estimation of rotor noise PSD via acoustical signals much more challenging with the increase in interfering noise levels relative the rotor noise in the microphone recordings. In contrast, the proposed method delivered a consistent performance regardless of the input noise level because the method takes non-acoustical information as its input, making the algorithm immune to interference noise.

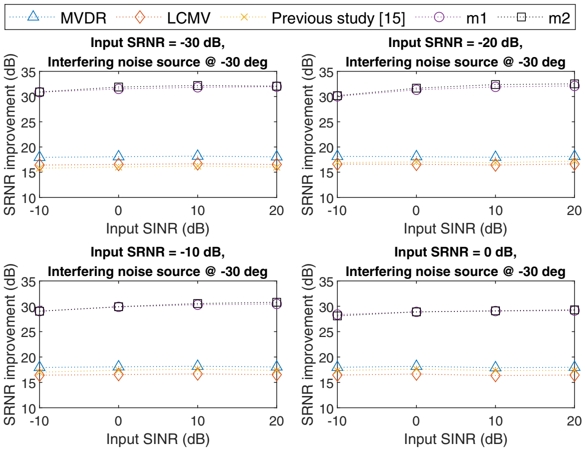

Fig. 8: SRNR improvement for different input SRNR (each subplot) and SINR (horizontal axis of each subplot).

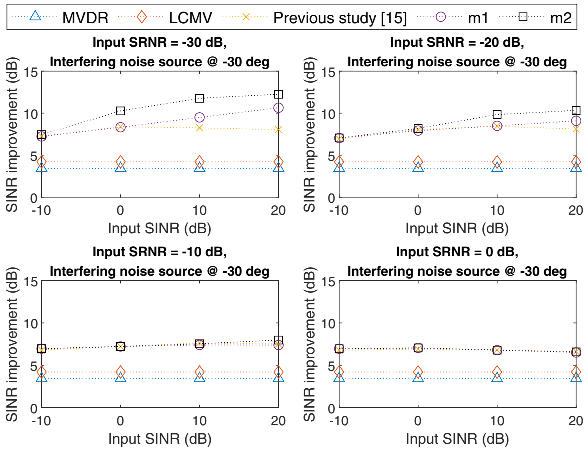

Fig. 10: SINR improvement for different input SRNR (each subplot) and SINR (horizontal axis of each subplot).

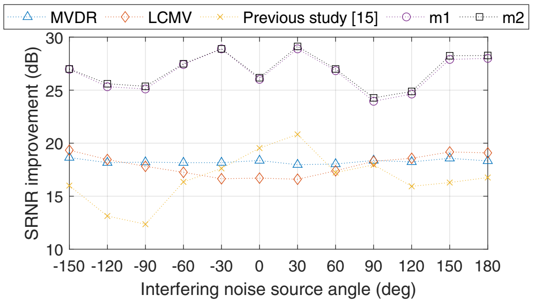

Fig. 9: SRNR improvement with different interfering noise source angles (at input SRNR/SINR of 0 dB).

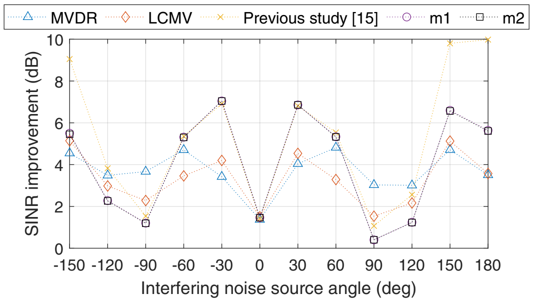

Fig. 11: SINR improvement with different interfering noise source angles (at input SRNR/SINR of 0 dB).

Comparing $\mathrm{m1}$ and $\mathrm{m2}$ shows very similar performance. In most cases, $\mathrm{m1}$ delivers LSDs of on average 0.5 dB lower. This is significantly different to that seen from [28] and [29], where a mixed performance was seen. This is potentially driven by the acoustical input feature $\phi_{Y_{\theta_u}}(\omega, t)$, which could have brought forward influences of cross-leakage and other practical IR effects due to deviations from an ideal free-field, as experienced similarly by the precious study [15]. However, utilising the input feature $\phi_{Y_{\theta_u}}(\omega, t)$ still brings performance advantages which will later be evident in Section V-B.

B. Evaluation of source enhancement performance

This section compares the performances of the previous study and the proposed method for source enhancement. However, as conventional source enhancement techniques, results from MVDR and the linearly constrained minimum variance (LCMV) [34] beamformers are also compared.

Fig. 8 summarises the SRNR improvement with respect to the microphone inputs for different input SRNRs and SINRs. Here, the proposed method shows significant SRNR improvement over the previous study and those from MVDR and LCMV beamformers, where the lower the input SRNR, the higher the SRNR improvement the proposed method was able to deliver. The proposed method is also able to present a relatively consistent performance for different input SINRs, indicating increased robustness under harsher input rotor noise conditions (i.e. lower SRNR). Performance against the variation of the angle of the interfering noise source (as outlined in Fig. 5) is presented in Fig. 9. Here, there are variations in SRNR improvements depending on the interfering noise source placement, which vary similarly to that from the previous study [15]. This is expected as the source enhancement process from (13) to (16) follows that from [15], which makes use of the directivity matrices $G_{0,\theta_0}(\omega)$ and $G_{n,\theta_n}(\omega)$ to reduce the remaining interfering noise sources. Thus some of the inherent spatial aliasing effects would also be carried over.

The proposed method outperformed MVDR and LCMV while delivering comparable performance compared to the previous study in SINR improvement performance over different input noise conditions, as shown in Fig. 10. While the overall reduction in performance of the proposed method is not significant, it indicates that some loss in SINR performance is introduced from the rotor noise reduction filter. This is further highlighted in Fig. 11, which summarises the SINR performance over different placement angle of the interfering noise source (as outlined in Fig. 5). Here, the proposed method and the previous study have similar performance for angles between 0 deg to ± 90 deg, after which the previous study [15] performs somewhat better. As mentioned previously, since the input signal would require both the rotor noise reduction process ((7) to (12)), as well as the interfering noise reduction postfiltering process ((13) to (16)), potential distortion and spatial separation effects between the target and interfering noise sources would have reflected in the SINR performance. However, this could potentially be overcome by modifying the general framework to minimise the impact on the spatial properties coming from the rotor noise reduction process. However, this remains part of future investigation.

Like the rotor noise PSD estimation performance seen in Fig. 7, $\mathrm{m1}$ and $\mathrm{m2}$ show very similar source enhancement performance in both SRNR and SINR improvement. Contrary to the rotor noise PSD estimation accuracy performance, overall $\mathrm{m2}$ delivers slightly higher SRNR improvement over $\mathrm{m1}$, as shown in Fig. 8 and Fig. 9. This indicates that the input feature $\phi_{Y_{\theta_u}}(\omega, t)$ gives $\mathrm{m2}$ the benefit to track minute details (such as broadband spectra) in the rotor noise PSD due to it having an acoustic reference. This achieves the objective of the utilisation of *multi-sensory information*, where the hybrid input information complements another to refine the algorithm's source enhancement performance.

It should be noted that while the proposed method yields significant improvement in the overall source enhancement performance over the previous study [15], the experiment setup is simplified (i.e. one rotor). With a full UAV setup of multiple rotors, it is expected that variations in the combinations of speeds would increase the overall complexity of the rotor noise PSD spectrum. In addition, a four-rotor setup would require an extra beamformer pointing towards $P_{u,\mathrm{R}}$, while the other existing beamformers (i.e. beamformer 0-2) require an extra null placed at $P_{u,\mathrm{R}}$ (see Table I). This results in changes in the beamformer properties, and as a result affecting the PSDs spectral characteristics. However, given the capabilities of machine learning techniques, it is expected that the proposed method would adapt without much difficulty.

Overall, the results show that appropriate utilisation of rotor motion characteristics was able to consistently deliver significantly improved source enhancement performance compared to [15] in reducing rotor noise despite the existence of other coherent interfering noise sources.

VI. Conclusion

We proposed a source enhancement method for a UAV-mounted audio recording system, using multi-sensory information which includes characteristics of the UAV rotor's state, to accurately estimate rotor noise PSD. The mapping between the multi-sensory information and the rotor noise PSDs are carried out using RTs. The rotor noise PSDs were then subsequently utilised in a source enhancement algorithm for rotor noise reduction and interfering noise removal using Wiener postfilters.

This study evaluated the algorithm by a practical system compared to simulation in previous studies, with results showing consistent PSD estimation accuracy improvements with robustness against interferences from non-rotor noise-related sources such as target source and coherent interfering noise. Source enhancement results showed the method's ability to perform well under most input noise conditions. However, as an initial conceptual study of the method, only a one-rotor problem setup was considered. Future work sees the need to expand the experimental evaluation to a full, four-rotor UAV setup, as well as developing the general framework further to account for the drop in SINR improvement compared to the previous study.

References

[1] R. Verrier, "Drones are providing film and TV viewers a new perspective on the action," *Los Angeles Times*, Jun. 2017.

[2] M. Margaritoff, "An English lifeboat crew is testing drones for search and rescue," *The Drive*, Sept. 2017.

[3] A. Charlton, "Police drone to fly over New Years Eve celebrations in Times Square," *salon*, Dec. 2018.

[4] P. Marmaroli, X. Falourd, and H. Lissek, "A UAV motor denoising technique to improve localization of surrounding noisy aircrafts: Proof of concept for anti-collision systems," in *Acoustics*, Société Française d'Acoustique, Ed., Apr. 2012.

[5] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2015, pp. 6143–6148.

[6] K. Washizaki, M. Wakabayashi, and M. Kumon, "Position estimation of sound source on ground by multirotor helicopter with microphone array," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2016, pp. 1980–1985.

[7] 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.

[8] 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.

[9] T. Spadini, G. S. Imai Aldeia, G. Barreto, K. Alves, H. Ferreira, R. Suyama, and K. Nose-Filho, "On the application of segan for the attenuation of the ego-noise in the speech sound source localization problem," in *2019 Workshop on Communication Networks and Power Systems (WCNPS)*, Oct 2019, pp. 1–4.

[10] 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.

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

[12] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2019, pp. 5320–5325.

[13] 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.

[14] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2016, pp. 219–222.

[15] 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.

[16] 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.

[17] J. E. Ffowcs Williams and D. L. Hawkings, "Theory relating to the noise of rotating machinery," *Journal of Sound and Vibration*, vol. 10, no. 1, pp. 10–21, 1969.

[18] G. Sinibaldi and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Applied Acoustics*, vol. 74, no. 1, pp. 79–88, Jan. 2013.

[19] R. P. Fernandes, E. C. Santos, A. L. L Ramos, and J. A. Apolinário, "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proceedings of the 20th International Conference on Transformative Science and Engineering, Business and Social Innovation*, 2015, pp. 536–541.

[20] J. Hou, S. Wang, Y. Lai, Y. Tsao, H. Chang, and H. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, April 2018.

[21] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 26–30.

[22] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 116–120.

[23] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 1577–1581.

[24] Y. Tu, J. Du, Y. Xu, L. Dai, and C-H. Lee, "Deep neural network based speech separation for robust speech recognition," in *2014 12th IEEE International Conference on Signal Processing (ICSP)*, Oct. 2014, pp. 532–536.

[25] G. Hinton, L. Deng, D. Yu, G. Dahl, A-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[26] T. Kawase, K. Niwa, K. Kobayashi, and Y. Hioka, "Application of neural network to source PSD estimation for Wiener filter based array sound source enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2016, pp. 1–5.

[27] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi, and Y. Hioka, "Supervised source enhancement composed of nonnegative auto-encoders and complementarity subtraction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 266–270.

[28] 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.

[29] 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.

[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, CRC Press, 1984.

[31] 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.

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

[33] L. R. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, PTR Prentice Hall, 1993.

[34] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.