Sound source enhancement using power spectral density estimation in beamspace for a dual unmanned aerial vehicle system

Mingxue Song*, Jin Xuan Teh*, Yusuke Hioka*, Benjamin Yen† and Hiroshi Saruwatari‡
* Acoustics and Vibration Research Centre, University of Auckland, Auckland, New Zealand
† Department of Systems and Control Engineering, Institute of Science Tokyo, Japan
‡ Graduate School of Information Science and Technology, The University of Tokyo, Japan
E-mail: mingxue.song@auckland.ac.nz

***Abstract*—A sound source enhancement framework is proposed for a dual unmanned aerial vehicle (UAV) system, where each UAV is equipped with a microphone array. While introducing more UAVs inevitably increases interfering noises, it simultaneously provides additional information that could enable improved robustness against various noise conditions and ultimately lead to superior overall performance compared to a single-UAV setup. This paper extends the power spectral density (PSD) estimation in beamspace by jointly using the combination of beamformers' directivity gains from both arrays to improve estimation accuracy. Moreover, the cross-power spectral density between the beamformers' outputs is incorporated to further refine the PSD estimation of the target source. The resulting PSD estimates are utilised to compute a non-linear post-filter, which is then applied to the averaged beamforming output to suppress residual noise. Experimental results show that the proposed framework outperforms the single-UAV baseline, achieving significant improvement in terms of signal-to-interference-plus-rotor-noise ratio and extended short-time objective intelligibility.**

I. Introduction

Unmanned aerial vehicles (UAVs), or drones, have been widely deployed with various sensors, such as cameras, radar, and microphones. Among these, drone audition [1], the technique to capture audio signals using microphones mounted on UAVs, is attracting growing interest due to its potential in a variety of applications, such as filming and search and rescue [2]—particularly when the visual information is obstructed. However, the noise generated by their rotors and propellers [3], also known as ego noise, severely contaminates the target source signals, limiting its practical applications.

To enhance the quality of the target sound source in drone audition, previous research has investigated various approaches to mitigate the effect of ego noise from audio recordings, such as [4]–[10]. Among these approaches, studies using the framework consisting of a beamformer and a Wiener post-filter has shown to be particularly effective [4], [5]. Based on that, Yen et al. [6] further improved the accuracy of noise covariance matrix estimation by exploiting rotor state information, resulting in improved ego noise reduction.

In recent years, distributed processing systems containing multiple sensor nodes, each equipped with computing and communication capabilities, have been actively investigated [11]. By deploying sensors over a wider area, a wireless sensor network can be formed. Compared to single-sensor systems, a wireless sensor network is preferred due to factors such as higher input signal-to-noise ratio at specific nodes [12], information sharing among nodes [13], and improved scalability and robustness in diverse environments [14]. This concept has also been applied to multiple UAVs, also referred to as drone swarms [15]. Such distributed UAV systems have been employed in object detection [16], tracking [17], and surveillance [18], etc., demonstrating significant practical utility. While some studies utilising multiple UAVs have shown potential in sound source localisation and tracking [17], noises from adjacent UAVs have been reported to significantly affect the tracking accuracy [19].

In the context of sound source enhancement for drone audition, the use of multiple UAVs has not been studied yet. Similar to the localisation and tracking problems [19], the additional UAVs act as interfering noise sources, which, on top of the already very loud ego noise, make the conditions particularly challenging. As a result, directly applying sound source enhancement methods designed for a single UAV may lead to degraded performance, thereby undermining the benefit of deploying multiple UAVs. To address this challenge, this study investigates a framework for sound source enhancement in a dual-UAV system as an early-stage exploration of utilising multi-UAV configurations for drone audition. We base our approach on the previous studies for a single-UAV system [4], [5]. Power spectral density (PSD) estimation in beamspace [20] is utilised for computing the Wiener post-filter weights, which are then applied to the output of a minimum variance distortionless response (MVDR) beamformer. Building upon this foundational framework, we extend the PSD in beamspace to dual-UAV systems, aiming to use the spatial information from both arrays to cooperatively estimate the PSDs of the sources. Although the proposed framework is somewhat similar to the study in [21], the current study focuses on the unique problem in drone audition. Moreover, we further enhance the PSD estimates by considering the cross-power spectral density (CSD) between the beamforming outputs of two arrays, each coming from one of the two UAVs. The refined PSDs

979-8-3315-7206-8/25/$31.00

are then used to compute the Wiener post-filter weights to remove the residual noise. Similarly, one could apply methods on source enhancement using distributed microphone arrays e.g. [22], [23] estimating the "common components" from multiple beamforming outputs by using non-negative matrix factorisation or non-negative tensor factorisation. However, their effectiveness in drone audition also remains unclear, as the dominant drone noise may be incorrectly clustered into the target source.

II. PSD ESTIMATION IN BEAMSPACE

As preparation, this section briefly reviews the PSD estimation in beamspace [20], which was used in the previous study [5] for sound source enhancement on a single-UAV system. Assume an M-channel microphone array receives signals from a target sound source and N interfering noise sources. In the short-time Fourier transform (STFT) domain, the observed signal $\mathbf{x}_{ij}$ is given by:

$$\mathbf{x}_{ij} = \mathbf{a}_{i,\theta_0} S_{ij} + \sum_{n=1}^{N} \mathbf{a}_{i,\theta_n} N_{ij,\theta_n}, \tag{1}$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ denote the frequency and time indices, respectively. S_{ij} is the target sound source arriving from direction θ_0 and N_{ij,θ_n} is the n-th interfering noise arriving from direction θ_n. $\mathbf{a}_{i,\theta_0}$ denotes the transfer function from the target source to the microphone array, and $\mathbf{a}_{i,\theta_n}$ denotes that from the n-th noise source to the microphone array.

A fixed beamformer pointing its mainlobe towards the target source is applied to $\mathbf{x}_{ij}$, with its weight vector denoted as $\mathbf{w}_{i,l_0}$. The output of the beamformer is then given by $y_{ij,l_0} = \mathbf{w}_{i,l_0}^H \mathbf{x}_{ij}$, where $(\cdot)^H$ denotes the Hermitian transpose. By assuming all sources are mutually uncorrelated, the instantaneous PSD of the beamformer's output $\phi_{Y_{ij,l_0}}$ can be approximated as

$$\phi_{Y_{ij,l_0}} \approx |G_{i,l_0,\theta_0}|^2 \, \phi_{S_{ij}} + \sum_{n=1}^{N} |G_{i,l_0,\theta_n}|^2 \, \phi_{N_{ij,\theta_n}}, \tag{2}$$

where $\phi_{S_{ij}}$ and $\phi_{N_{ij,\theta_n}}$ denote the instantaneous PSD of S_{ij} and N_{ij,θ_n}, respectively. The instantaneous PSD of an arbitrary signal $\mathcal{X}_{ij}$ may be calculated by using Welch's method [24] given by $\phi_{\mathcal{X}_{ij}} = \alpha \phi_{\mathcal{X}_{ij-1}} + (1-\alpha)|\mathcal{X}_{ij}|^2$, where α is the forgetting factor. $|G_{i,l_0,\theta}|^2$ is the directivity gain of the beamformer with $G_{i,l_0,\theta} := (\mathbf{w}_{i,l_0})^H \mathbf{a}_{i,\theta}$.

Now assume $N-1$ more beamformers with different directivity patterns are applied to the microphone array in addition to the one pointing towards the target source. These beamformers are typically designed to point their mainlobes to the angles of the noise sources θ_n [20]. The PSDs of all beamformers' outputs can then be represented in a form of simultaneous equation as

$$\underbrace{\begin{bmatrix} \phi_{Y_{ij,l_0}} \\ \phi_{Y_{ij,l_1}} \\ \vdots \\ \phi_{Y_{ij,l_N}} \end{bmatrix}}_{\Phi_{Y_{ij}}} \approx \underbrace{\begin{bmatrix} |G_{i,l_0,\theta_0}|^2 & |G_{i,l_0,\theta_1}|^2 & \cdots & |G_{i,l_0,\theta_N}|^2 \\ |G_{i,l_1,\theta_0}|^2 & |G_{i,l_1,\theta_1}|^2 & \cdots & |G_{i,l_1,\theta_N}|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |G_{i,l_N,\theta_0}|^2 & |G_{i,l_N,\theta_1}|^2 & \cdots & |G_{i,l_N,\theta_N}|^2 \end{bmatrix}}_{\mathbf{G}_i} \underbrace{\begin{bmatrix} \phi_{S_{ij}} \\ \phi_{N_{ij,\theta_1}} \\ \vdots \\ \phi_{N_{ij,\theta_N}} \end{bmatrix}}_{\Phi_{P_{ij}}}, \tag{3}$$

where $\Phi_{P_{ij}}$ represents the PSD vector of the target and noise sources, which can be estimated by solving this equation using the least squares method as follows:

$$\widehat{\mathbf{\Phi}}_{P_{ij}} = \left(\mathbf{G}_i^H \mathbf{G}_i\right)^{-1} \mathbf{G}_i^H \mathbf{\Phi}_{Y_{ij}}. \tag{4}$$

These resulting estimates can then be utilised to design a non-linear filter for the reduction of residual noise.

III. PROPOSED METHOD

This section elaborates the proposed method. We first define the problem of sound source enhancement with the dual-UAV setup (Section III-A), which is followed by the details of the proposed extension of PSD estimation in beamspace (Section III-B) and its implementation for sound source enhancement (Section III-C).

A. Problem definition

Unlike many applications of sound source enhancement, the positions of the UAV-mounted microphone array with respect to the target source and other UAVs can be flexibly adjusted due to the mobility and multisensory nature of UAVs. Therefore, in the current study, we specifically assume that two UAVs are positioned equidistant from the target source, as shown in Figure 1. The microphone array is mounted on a boom sticking out from the UAV's main body so that the target source and the UAV's rotors are located in opposite directions with respect to the microphone array. The k-th UAV's observation $\mathbf{x}_{ij}^k$ with $k \in \{1, 2\}$, includes the target source S_{ij} and its own ego noises N_{ij,θ_n^k}, which are assumed to be uncorrelated. Furthermore, $\mathbf{x}_{ij}^k$ also contains the interfering rotor noises $N_{ij}^{k'}$ from another

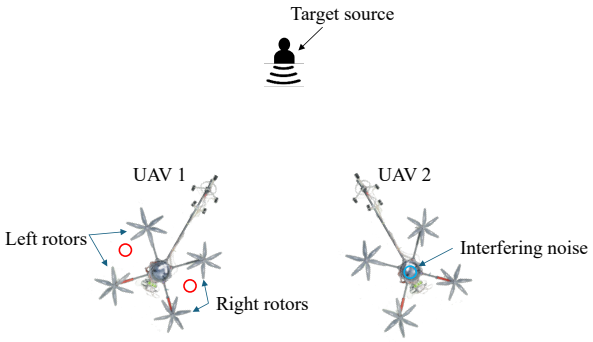

Fig. 1. A dual-UAV system for target source enhancement. With respect to UAV 1, ego noises are represented by point sources from left and right rotors, denoted by red circles. Interfering noise is also represented by a point source from UAV 2, denoted by a blue circle.

UAV k', such that $(k, k') = (1, 2)$ or $(2, 1)$. Following [5], we simplify the ego noise model by assuming that the noises from the left and right rotors are respectively represented by point sources as shown in Figure 1, i.e. the number of ego noises is $N = 2$. Together with the transfer functions from the target source to each microphone array $\mathbf{a}_{i,\theta_0^k}$, the transfer functions of the ego noises $\mathbf{a}_{i,\theta_n^k}$ are assumed to be known *a priori* and will be used in the subsequent beamformer design. The interfering rotor noise from another UAV is also modelled as a single point source, as shown in Figure 1, where its transfer function is denoted by $\mathbf{a}_{i,\theta_{N_{k'}}}$. However, this transfer function is assumed to be unknown, as such information is generally unavailable in practice. Overall, $\mathbf{x}_{ij}^k$ can be expressed as

$$\mathbf{x}_{ij}^k = \mathbf{a}_{i,\theta_0^k} S_{ij} + \sum_{n=1}^{N} \mathbf{a}_{i,\theta_n^k} N_{ij,\theta_n^k} + \mathbf{a}_{i,\theta_{N_{k'}}} N_{ij}^{k'}. \tag{5}$$

The problem of this study is to enhance only the target source S_{ij} using the observed signals from the dual UAVs $\mathbf{x}_{ij}^1$ and $\mathbf{x}_{ij}^2$.

B. PSD estimation in beamspace for dual-array

The proposed framework for sound source enhancement in a dual-UAV system is illustrated in Figure 2. It comprises multiple processing blocks, including beamforming, PSD estimation, and post-filtering. Fixed beamformers are first applied to enhance the target signal for each array independently, where MVDR beamformers are adopted in this study. For the k-th UAV, the target beamforming output is given by $y_{ij,l_0^k} = (\mathbf{w}_{i,l_0^k})^H \mathbf{x}_{ij}^k$, where $\mathbf{w}_{i,l_0^k}$ denotes the weight of the MVDR beamformer. Additionally, two MVDR beamformers are steered towards the directions of left and right rotors for each UAV. The PSDs and CSDs of the outputs of these beamformers are calculated, which are then used to estimate the PSDs of the sources. These PSD estimates are subsequently used to derive the Wiener filter weights, which are applied to the averaged target beamforming outputs to obtain the final enhanced signal. This framework effectively exploits the spatial information from both arrays, which offers increased robustness against diverse noise environments. Thus, it is expected to improve the overall performance for sound source enhancement.

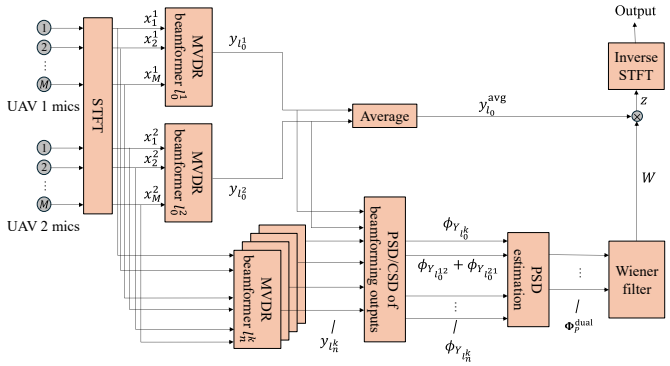

Fig. 2. Block diagram of the proposed sound source enhancement framework for a dual-UAV system (time and frequency indices omitted).

For the PSD estimation, we expand the approach introduced in Section II by first stacking the PSDs of all beamformers' outputs from both UAVs together and formulating a simultaneous equation as follows:

$$\mathbf{\Phi}_{Y_{ij}}^{\text{dual}} = \begin{bmatrix} \phi_{Y_{ij,l_0^1}} \\ \phi_{Y_{ij,l_1^1}} \\ \phi_{Y_{ij,l_2^1}} \\ \phi_{Y_{ij,l_0^2}} \\ \phi_{Y_{ij,l_1^2}} \\ \phi_{Y_{ij,l_2^2}} \end{bmatrix} \approx \mathbf{G}_i^{\text{dual}} \underbrace{\begin{bmatrix} \phi_{S_{ij}} \\ \phi_{N_{ij,\theta_1^1}} \\ \phi_{N_{ij,\theta_2^1}} \\ \phi_{N_{ij,\theta_1^2}} \\ \phi_{N_{ij,\theta_2^2}} \end{bmatrix}}_{\mathbf{\Phi}_{P_{ij}}^{\text{dual}}}, \tag{6}$$

where $\mathbf{G}_i^{\text{dual}}$ is given by

$$\mathbf{G}_i^{\text{dual}} = \begin{bmatrix} \left|G_{i,l_0^1,\theta_0^1}\right|^2 & \left|G_{i,l_0^1,\theta_1^1}\right|^2 & \left|G_{i,l_0^1,\theta_2^1}\right|^2 & 0 & 0 \\ \left|G_{i,l_1^1,\theta_0^1}\right|^2 & \left|G_{i,l_1^1,\theta_1^1}\right|^2 & \left|G_{i,l_1^1,\theta_2^1}\right|^2 & 0 & 0 \\ \left|G_{i,l_2^1,\theta_0^1}\right|^2 & \left|G_{i,l_2^1,\theta_1^1}\right|^2 & \left|G_{i,l_2^1,\theta_2^1}\right|^2 & 0 & 0 \\ \left|G_{i,l_0^2,\theta_0^2}\right|^2 & 0 & 0 & \left|G_{i,l_0^2,\theta_1^2}\right|^2 & \left|G_{i,l_0^2,\theta_2^2}\right|^2 \\ \left|G_{i,l_1^2,\theta_0^2}\right|^2 & 0 & 0 & \left|G_{i,l_1^2,\theta_1^2}\right|^2 & \left|G_{i,l_1^2,\theta_2^2}\right|^2 \\ \left|G_{i,l_2^2,\theta_0^2}\right|^2 & 0 & 0 & \left|G_{i,l_2^2,\theta_1^2}\right|^2 & \left|G_{i,l_2^2,\theta_2^2}\right|^2 \end{bmatrix}. \tag{7}$$

The PSD vector of the source components $\mathbf{\Phi}_{P_{ij}}^{\text{dual}}$ can be estimated by solving (6) using the least squares method as well.

While (6) expands the original PSD estimation in beamspace [20] to leverage both arrays' information for enhancing the PSD estimates, it only considers the auto-PSDs of the beamformers' outputs, i.e. the phase information is ignored. Such information related to phase can be taken into account by employing the CSDs of the target beamforming outputs from both arrays to further enhance the estimation accuracy. These CSDs are denoted as $\phi_{Y_{ij,l_0^{12}}}$ and $\phi_{Y_{ij,l_0^{21}}}$, which are given by $y_{ij,l_0^1}(y_{ij,l_0^2})^*$ and $y_{ij,l_0^2}(y_{ij,l_0^1})^*$, respectively, where $(\cdot)^*$ denotes conjugate. By assuming that the ego noises from each array are also uncorrelated, the sum of the CSDs can be approximated as

$$\begin{aligned} &\phi_{Y_{ij,l_0^{12}}} + \phi_{Y_{ij,l_0^{21}}} \\ &\approx \left(G_{i,l_0^1,\theta_0^1}(G_{i,l_0^2,\theta_0^2})^* + G_{i,l_0^2,\theta_0^2}(G_{l_0^1,\theta_0^1})^*\right)\mathbb{E}[S_{ij}S_{ij}^*] \\ &\quad + \sum_{n,n'=1}^{2} \Big(G_{i,l_0^1,\theta_n^1}(G_{i,l_0^2,\theta_{n'}^2})^*\mathbb{E}[N_{ij,\theta_n^1}(N_{ij,\theta_{n'}^2})^*] \\ &\quad + G_{i,l_0^2,\theta_n^2}(G_{i,l_0^1,\theta_{n'}^1})^*\mathbb{E}[N_{ij,\theta_n^2}(N_{ij,\theta_{n'}^1})^*]\Big) \\ &\approx 2\,\mathrm{Re}\left[G_{i,l_0^1,\theta_0^1}\left(G_{i,l_0^2,\theta_0^2}\right)^*\right]\phi_{S_{ij}}, \end{aligned} \tag{8}$$

where $\mathrm{Re}[\cdot]$ denotes the real part of a complex value. The resulting equation can be incorporated into (6) to refine the

target PSD as

$$\begin{bmatrix} \mathbf{\Phi}^{\text{dual}}_{Y_{ij}} \\ \phi_{Y_{ij,l_0^{12}}} + \phi_{Y_{ij,l_0^{21}}} \end{bmatrix} \approx \begin{bmatrix} \mathbf{G}^{\text{dual}}_{i} \\ 2\,\text{Re}\left[G_{i,l_0^1,\theta_0^1}(G_{i,l_0^2,\theta_0^2})^*\right] \quad \mathbf{0}_{1\times 4} \end{bmatrix} \mathbf{\Phi}^{\text{dual}}_{P_{ij}}, \tag{9}$$

where $\mathbf{0}_{1\times 4}$ denotes a 1×4 zero vector. $\mathbf{\Phi}^{\text{dual}}_{P_{ij}}$ can be estimated by the same manner as in (4).

C. Implementation for sound source enhancement

A Wiener filter can be computed using these estimated PSDs and applied to the averaged beamforming output of both arrays: $y^{\text{avg}}_{ij,l_0} := \frac{1}{2}(y_{ij,l_0^1} + y_{ij,l_0^2})$, to suppress the noises from both UAVs as

$$\begin{aligned} z_{ij} &= W_{ij}\, y^{\text{avg}}_{ij,l_0} \\ &= \frac{\phi_{S_{ij}}}{\phi_{S_{ij}} + \frac{1}{4}\sum_{k=1}^{2}\sum_{n=1}^{2}\phi_{N_{ij,\theta_n^k}}}\, y^{\text{avg}}_{ij,l_0}. \end{aligned} \tag{10}$$

Finally, z_{ij} is transformed back to the time domain using the inverse STFT to obtain the enhanced target signal.

IV. Experiments

This section presents the experimental settings (Section IV-A), followed by a comparison between the single-UAV system and the proposed dual-UAV system (Section IV-B). Finally, a discussion on the synchronisation problem is provided (Section IV-C).

A. Experimental settings

The proposed framework was evaluated using recordings collected in a previous study [25] on a prototype UAV equipped with a microphone array. As shown in Figure 3, the array consisted of two sub-arrays: the Front array and the Rear array. The Front array comprised a shotgun microphone (mic 1) surrounded by three unidirectional microphones (mic 2-4) all pointing approximately in the target direction θ_0. The Rear array consisted of two unidirectional microphones (mic 5-6) oriented towards the left and right UAV rotors, respectively. An additional omnidirectional microphone (mic 7) was also installed on the boom to record the UAV noise, which was later used to simulate the interfering noise for another UAV.

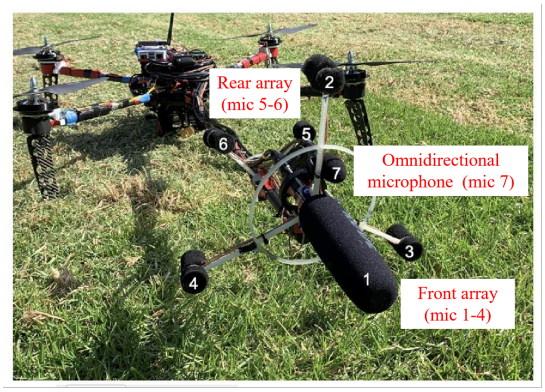

Fig. 3. UAV prototype from [25].

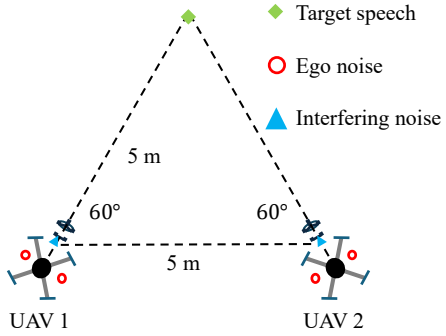

Fig. 4. UAV setup with marked assumed point source positions for impulse response measurement.

The impulse responses were also measured, which were utilised to calculate the transfer functions and synthesise the interfering UAV noise. Figure 4 indicates the assumed source positions for measuring the impulse responses. The target source was assumed to be located directly in front of the UAV at a distance of $5\,\text{m}$ for both UAVs, while the interfering UAV noises were assumed to arrive from approximately $60°$ to their left or right of the target source, respectively, also $5\,\text{m}$ away. To simulate the dual-UAV environment, the interfering noise for UAV 1 was synthesised by convolving the impulse response from the UAV 2's position to the UAV 1's microphone array with the ego noise recorded by mic 7 on UAV 2, and vice versa for UAV 2's interfering noise. Target speech and ego noise were recorded separately in an open outdoor field with a fixed hovering UAV [25]. Since we assumed that the ego noises from each array are uncorrelated, we used different sets of ego noise recordings for each UAV. The target speech, ego noise and interfering noise were then summed up at each UAV's microphone array to generate the simulated observed signals.

All input signals and impulse responses were measured at $96\,\text{kHz}$ and were resampled to $16\,\text{kHz}$ in implementation. The STFT was performed using a Hanning window with a length of $32\,\text{ms}$ and an overlap of $16\,\text{ms}$. To assess the signal quality, two metrics were used: signal-to-interference-plus-rotor-noise-ratio (SINR) and extended short-time objective intelligibility (ESTOI) [26]. Here, SINR was defined as the ratio between the power of the target speech and the combined power of interfering UAV noise and ego noise. The shotgun microphone was used as the reference microphone, where the input signal-to-interference-ratio (SIR) was fixed at $0\,\text{dB}$, which defines the power ratio between the target source and the interfering UAV noise. Two cases with input signal-to-*rotor*-noise ratios (SRNRs) [6] at $-20\,\text{dB}$ and $-30\,\text{dB}$ were tested. The performance of the proposed method was compared against that of the foundational framework [5] under two conditions: with and without an interfering UAV noise. The performance at the beamforming stage was also compared. In addition, the contribution of incorporating CSD to PSD estimation was quantified by evaluating the performance gain of the dual-UAV system using PSDs estimated from (9) over those from (6).

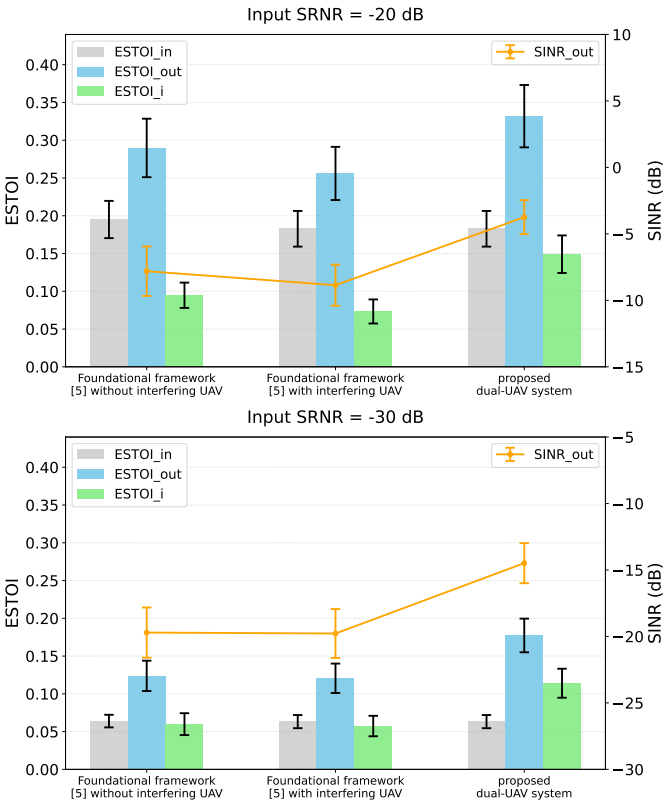

Fig. 5. ESTOI (bars) and SINR (lines) at $-20\,\mathrm{dB}$ and $-30\,\mathrm{dB}$ input SRNRs with 95% confidence intervals. The three conditions (left to right) are: foundational framework [5] without the interfering UAV, foundational framework [5] with the interfering UAV, and proposed framework for dual UAVs.

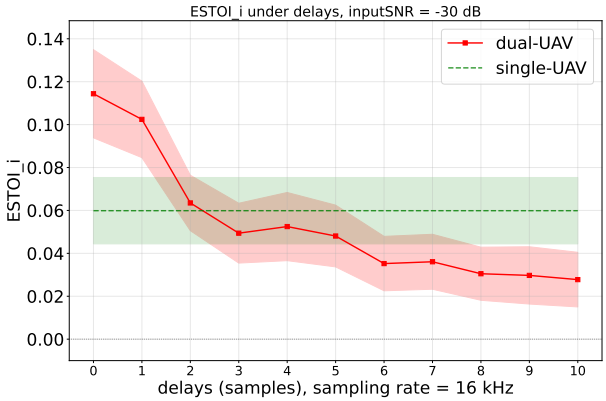

Fig. 6. ESTOI under various delays applied to UAV 2 with 95% confidence intervals shown by error bars. The dotted line denotes the single-UAV case.

TABLE I
SINR AND ESTOI OF TARGET BEAMFORMING OUTPUTS (WITH 95% CONFIDENCE INTERVALS). PERFORMANCE IS EVALUATED BETWEEN SINGLE AND AVERAGED BEAMFORMING OUTPUTS.

Metric	Beamformer	Input SRNR = -20 dB	Input SRNR = -30 dB
SINR (dB)	Single	-12.0692 ± 1.0068	-21.5244 ± 1.1313
	Averaged	-9.7937 ± 0.7417	-18.6358 ± 0.7954
ESTOI	Single	0.2518 ± 0.0307	0.1162 ± 0.0159
	Averaged	0.3102 ± 0.0373	0.1598 ± 0.0195

TABLE II
PERFORMANCE GAIN (WITH 95% CONFIDENCE INTERVALS) OF THE DUAL-UAV SYSTEM USING PSDS ESTIMATED FROM (9) OVER THOSE FROM (6).

Performance gain	Input SRNR = -20 dB	Input SRNR = -30 dB
SINR_gain (dB)	1.8052 ± 0.3446	1.0317 ± 0.4503
ESTOI_gain	0.0163 ± 0.0040	0.0134 ± 0.0042

B. Results and discussions

Figure 5 presents the SINR and ESTOI metrics, where ESTOI_in and ESTOI_out denote the ESTOI of the input and output signals, respectively. ESTOI_i is defined as ESTOI_out − ESTOI_in. SINR_out represents the SINR of the output signal. Under the assumption of input SIR being $0\,\mathrm{dB}$, the plots for the two foundational framework cases illustrate that directly applying the foundational framework shows little to no performance difference. In contrast, the proposed framework demonstrates superior performance under the same condition. In addition, an improvement over the single-UAV system [5] in both metrics can be observed across different input SRNRs. This advantage is particularly evident in lower input SRNR scenarios, where SINR demonstrates approximately $5.3\,\mathrm{dB}$ improvement and ESTOI shows 0.06 point improvement compared to using a single UAV, highlighting the benefit of deploying dual UAVs. The improvement would be attributed to two main factors. First, the noise components could be averaged out during the beamforming stage, while the target signal is preserved. This is supported by the performance of the beamforming presented in Table I, where the performance of the averaged target beamforming output outperforms that of the single target beamforming output in both metrics. Second, incorporating the CSD between the beamforming outputs would yield better PSD estimates for more accurate post-filter design. Table II reports the SINR_gain and ESTOI_gain, defined as the differences in SINR_out and ESTOI_out resulting from using (9) rather than (6) for the proposed framework. The positive values across all cases indicate that using (9) leads to better performance, thus suggesting it provides more accurate PSD estimates.

C. Effect of synchronisation error

When distributed microphone arrays are employed, maintaining synchronisation between arrays is challenging in practice. In addition, we assume that the target source is equidistant from both arrays. While this assumption may be perfectly realised in simulation, it is often difficult to guarantee in real-world scenarios, leading to similar problems as asynchronous conditions. To investigate the effect of such synchronisation error, we manually introduced different delays to the signals observed by UAV 2. The result shown in Figure 6 suggests that while the proposed dual-UAV system can achieve better performance when the UAVs are perfectly synchronised, it is highly sensitive to synchronisation errors. This is because the estimation of source PSDs relies on the CSDs of the beamforming outputs, which utilise phase information. Such information will be inaccurate when asynchronisation occurs, subsequently degrading the accuracy of the PSD estimates. In

addition, averaging the outputs of two beamformers inherently requires strict time alignment. Addressing such sensitivity issue remains as open question to future research.

V. Conclusions

This study has investigated the use of a dual-UAV system for sound source enhancement. The proposed framework extends the PSD estimation in beamspace for dual UAVs to leverage the spatial information from both arrays, which effectively refines the PSD of the target source subsequently used for designing a non-linear filter. Experimental results demonstrated that the proposed method outperformed the previous method designed for single-UAV system in both SINR and ESTOI, showing its robustness under dual-UAV scenarios. Future research can further improve the practicality of current approaches by removing assumptions currently imposed, as well as addressing the synchronisation problems.

Acknowledgment

This study was supported by China Scholarship Council, the Acoustics and Vibration Research Centre at the University of Auckland and Kajima Foundation's Support Program for International Joint Research Activities (2024-kyodoshin-05).

References

[1] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura-Algarve, Portugal, 2012, pp. 3288–3293.

[2] Y. Sun, F. Zhang, Y. Liu, *et al.*, "Acoustic event detection for drone search and rescue system based on bi-directional long and short-term memory beamforming method to remove rotor noise," *Digital Signal Processing*, vol. 157, p. 104 881, 2025.

[3] R. McKay and M. J. Kingan, "Multirotor unmanned aerial system propeller noise caused by unsteady blade motion," in *25th AIAA/CEAS Aeroacoustics Conference*. 2019.

[4] 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)*, IEEE, 2016.

[5] 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.

[6] B. Yen, Y. Li, and Y. Hioka, "Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2491–2506, 2023.

[7] W. N. Manamperi, T. D. Abhayapala, P. N. Samarasinghe, and J. Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones using multichannel wiener filtering and gaussian-mixture based post-filtering," *Applied Acoustics*, vol. 216, p. 109 818, 2024.

[8] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[9] J. X. Teh, N. Takamune, H. Saruwatari, B. Yen, M. Kingan, and Y. Hioka, "Beamforming informed independent low-rank matrix analysis for sound source enhancement in unmanned aerial vehicles," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Macau, Macao, 2024, pp. 1–6.

[10] Z.-W. Tan and A. W. H. Khong, "SMoLnet-T: An efficient complex-spectral mapping speech enhancement approach with frame-wise CNN and spectral combination transformer for drone audition," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, pp. 1–6.

[11] S. Pasha, J. Lundgren, C. Ritz, and Y. Zou, "Distributed microphone arrays, emerging speech and audio signal processing platforms: A review," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 4, pp. 331–343, 2020.

[12] R. Chang, Z. Chen, and F. Yin, "Robust distributed noise suppression in acoustic sensor networks," *IEEE Sensors Journal*, vol. 22, no. 18, pp. 18 151–18 161, Sep. 2022.

[13] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks—part i: Sequential node updating," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5277–5291, 2010.

[14] Y. Hioka and W. B. Kleijn, "Distributed blind source separation with an application to audio signals," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 233–236.

[15] G. Chmaj and H. Selvaraj, "Distributed processing applications for UAV/drones: A survey," in *Advances in Intelligent Systems and Computing*, vol. 1089, Springer, 2014.

[16] V. Sadhu, C. Sun, A. Karimian, R. Tron, and D. Pompili, "AerialDeepSearch: Distributed Multi-Agent Deep Reinforcement Learning for Search Missions," in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Delhi, India, 2020, pp. 165–173.

[17] 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, p. 9039, 2021.

[18] D. Kingston, R. W. Beard, and R. S. Holt, "Decentralized perimeter surveillance using a team of UAVs," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1394–1404, Dec. 2008.

[19] B. Yen, T. Yamada, K. Itoyama, and K. Nakadai, "Evaluation of multi-drone sound source tracking algorithms," 2024.

[20] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1240–1250, Jun. 2013.

[21] Y. Hioka, K. Kobayashi, K. Furuya, and A. Kataoka, "Enhancement of sound sources located within a particular area using a pair of small microphone arrays," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 91, no. 2, pp. 561–574, 2008.

[22] S. Ayano, L. Li, S. Seki, and D. Kitamura, "Audio spotforming using nonnegative tensor factorization with attractor-based regularization," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Lyon, France, 2024, pp. 121–125.

[23] Y. Kagimoto, K. Itoyama, K. Nishida, and K. Nakadai, "Spotforming by NMF using multiple microphone arrays," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 9253–9258.

[24] P. D. 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, Jun. 1967.

[25] B. Yen, Y. Hioka, G. Schmid, and B. Mace, "Multi-sensory sound source enhancement for unmanned aerial vehicle recordings," *Applied Acoustics*, vol. 189, p. 108 590, 2022.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.