Beamforming informed independent low-rank matrix analysis for sound source enhancement in unmanned aerial vehicles

Jin Xuan Teh*, Norihiro Takamune†, Hiroshi Saruwatari†, Benjamin Yen‡, Michael Kingan* and Yusuke Hioka*
* Acoustics and Vibration Research Centre, University of Auckland, Auckland, New Zealand
† The University of Tokyo, Tokyo, Japan
‡ Tokyo Institute of Technology, Tokyo, Japan

***Abstract*—This study proposes the integration of informed, supervised, and blind sound source enhancement approaches for unmanned aerial vehicle (UAV) applications. The proposed method incorporates a beamformer, representing the informed approach, a pre-recorded noise database for the supervised approach, and independent low-rank matrix analysis (ILRMA) for the blind approach. This method aims to improve sound source enhancement performance while addressing the permutation ambiguity problem of the output channels inherent to ILRMA. The method leverages the fixed spatial relationship between the UAV's propellers and microphones to capture spatial information of the noise generated the propellers. This is achieved by deriving a noise covariance matrix from pre-recorded propeller noise signals and incorporating it into a general eigenvalue beamformer to effectively suppress these noises. The filter weights of the beamformer are then used to inform the spatially regularised ILRMA, guiding the algorithm with additional spatial information of the sound sources. Experimental results demonstrate significant performance improvements, including a 13 dB increase in the source-to-distortion ratio and a 0.21 point increase in the short-time objective intelligibility score. The proposed method outperforms both the original ILRMA and the beamformer in most scenarios and effectively addresses the global permutation ambiguity problem.**

I. Introduction

Unmanned aerial vehicles (UAVs) are increasingly used in applications such as surveillance, filming, rescue operations, and wildlife monitoring. However, the disruptive noise generated by their rotors and propellers, known as 'ego-noise' significantly hinders their effectiveness in auditory applications [1]. This noise degrades the quality of audio recordings captured by onboard microphones, reducing their practical utility. High-quality audio recordings are essential in various domains. For example, environmental monitoring relies on clear audio to track natural events, and wildlife research depends on high-quality recordings to study animal behaviour and habitats [2]. In emergency response scenarios, precise audio data are crucial for effective decision-making, highlighting the potential of UAVs in outdoor search and rescue missions [3].

Addressing the challenge of ego-noise in UAV applications has led to the development of several approaches to enhance sound sources. These approaches can be categorised into three main types: informed, supervised, and blind. The informed approach utilises spatial cues to enhance the target signal, with beamforming being a widely used method [4]. The supervised approach leverages additional information, such as pre-recorded noise data, rotor speed, or additional microphones to capture noise-only signals. These techniques are often used in conjunction with machine learning techniques to estimate noise characteristics for effective noise suppression [5], [6]. The blind approach employs blind source separation (BSS) algorithms, which do not rely on prior information and separate sound sources based on their statistical properties, with independent low-rank matrix analysis (ILRMA) being a notable example [7].

Given the dominant ego-noise signal in UAV applications, achieving good performance in sound source enhancement is challenging. Researchers have increasingly combined different approaches to develop frameworks that maximise performance. Yen et al. [5] combined informed and supervised approaches, using rotor state information to estimate the noise covariance matrix with machine learning methods, which then informed a minimum variance distortionless respons (MVDR) beamformer to further suppress UAV ego-noise. Similarly, our previous work [8] used pre-recorded noise signals to inform a beamformer and estimate the power spectral density (PSD) of the target source. For integration with the blind approach, Lin et al. [9] combined BSS with spatial information of the target source using time-spatial filtering. These studies demonstrate the potential for superior performance through the integration of different approaches.

This study explores integrating informed, supervised, and blind approaches to improve sound source enhancement performance for UAV applications. The proposed framework combines a beamformer (informed approach), a pre-recorded noise database (supervised approach), and ILRMA (blind approach), leveraging each of their strengths while addressing their limitations. This integration aims to guide ILRMA to achieve enhanced performance under real-world conditions. The framework leverages the stationary nature of spatial characteristics between the microphone and UAV noise to estimate a noise spatial covariance matrix. It also uses a target spatial covariance matrix estimated from a range of potential target positions to design a beamformer weight to spatially regularise ILRMA.

II. Background

The proposed method integrates a beamformer framework from a previous study [8] with ILRMA [7]. This section presents the mathematical formulation of the UAV sound source enhancement problem (Section II-A), followed by an overview of the fundamental components, which include the beamforming framework (Section II-B) and ILRMA (Section II-C). The formulations provided here form the basis for the proposed method discussed in Section III.

A. Problem setup

Consider a problem where U independent source signals comprise one target source and $(U-1)$ spatially coherent noise sources are considered. Each target and noise source is assumed to be independent; thus, the sources are considered mutually uncorrelated. These sound sources are captured by an M-channel microphone array under the assumption that the system is determined $(M = U)$. The sound propagation from each source to the microphone array is modelled using a linear and convolutive mixture model. To analyse these sound signals in the time-frequency domain, the short-time Fourier transform (STFT) is applied to the observed signals. This effectively converts the convolutive mixture model into an instantaneous mixture model within each frequency bin. The resulting STFT of the target source and the noise sources are denoted as s_{ij} and $n_{ij,u}$, respectively, where $i = 1, ..., I$ indexes the frequency bins, $j = 1, ..., J$ indexes the time frames, and $u = 1, ..., U$ indexes the sources. Additionally, the STFT of ambient noise, represented as $\mathcal{V}_{ij} \in \mathbb{C}^M$, encompasses both background noise and inherent microphone self-noise. These signals make up the signals observed by the microphone array $x_{ij,m}$, which are expressed in vector form as follows:

$$\begin{aligned}\mathbf{x}_{ij,m} &= [x_{ij,1}, \ldots, x_{ij,M}]^{\mathsf{T}} \\ &= \mathbf{a}_{i,1} s_{ij} + \sum_{u=2}^{U} \mathbf{a}_{i,u} n_{ij,u} + \mathcal{V}_{ij}.\end{aligned} \tag{1}$$

Here, $m = 1, ..., M$ indicates the microphone index, and $^{\mathsf{T}}$ denotes a transpose. The steering vector [10] for each source, $\mathbf{a}_{i,u}$, is described as an array of transfer functions $a_{i,u,m}$ from the source u to the microphone m and is defined as follows:

$$\mathbf{a}_{i,u} = [a_{i,u,1}, \ldots, a_{i,u,M}]^{\mathsf{T}}. \tag{2}$$

B. Beamforming framework

Beamforming is a spatial filtering technique that isolates target signals while attenuating noise signals. The output of a beamformer can be expressed as follows:

$$\mathbf{y}_{ij} = \mathbf{w}_i^{\mathsf{H}} \mathbf{x}_{ij}. \tag{3}$$

Here, $^{\mathsf{H}}$ denotes a Hermitian transpose. The vector $\mathbf{w}_i = [w_{i,1}, \ldots, w_{i,M}]^{\mathsf{T}}$ represents the beamformer's filter weights, $w_{i,m}$, applied to the signals observed by each microphone m. In the absence of ambient noise sources $\mathcal{V}_{ij}$, the optimal filter weight can be found by maximising the signal-to-noise ratio (SNR) [11], which is expressed as follows:

$$\xi_i = \frac{{\mathbf{w}_i}^{\mathsf{H}} \mathbf{R}_{i,\mathrm{s}} \mathbf{w}_i}{{\mathbf{w}_i}^{\mathsf{H}} \mathbf{R}_{i,\mathrm{n}} \mathbf{w}_i}, \tag{4}$$

where $\mathbf{R}_{i,\mathrm{s}}$ and $\mathbf{R}_{i,\mathrm{n}}$ represent the spatial covariance matrices of frequency i for the target and noise signal, respectively. The solution to this optimisation problem, known as generalised eigenvalue (GEV) beamforming [12], [13], is given by:

$$\mathbf{w}^{\mathsf{GEV}} = \mathcal{P}\{\mathbf{R}_{\mathrm{n}}^{-1} \mathbf{R}_{\mathrm{s}}\}. \tag{5}$$

For brevity, i is omitted unless otherwise specified hereafter. Here, $\mathcal{P}\{\cdot\}$ denotes the operator that extracts the eigenvector corresponding to the largest eigenvalue through generalised eigenvalue decomposition [14]. To achieve a distortionless magnitude response, blind analytical normalisation (BAN) compensation factor was proposed [13]:

$$\psi = \frac{\sqrt{{\mathbf{w}^{\mathsf{GEV}}}^{\mathsf{H}} \mathbf{R}_{\mathrm{n}} \mathbf{R}_{\mathrm{n}} \mathbf{w}^{\mathsf{GEV}}}}{{\mathbf{w}^{\mathsf{GEV}}}^{\mathsf{H}} \mathbf{R}_{\mathrm{n}} \mathbf{w}^{\mathsf{GEV}}}, \tag{6}$$

where, ψ corrects the distortions in magnitude inherent to the GEV beamformer. Consequently, applying ψ allows the adjusted $\mathbf{w}^{\mathsf{GEV}}$ to reassemble the MVDR's distortionless magnitude response, albeit with phase distortions. The resulting beamformer is denoted as GEV beamformer with BAN compensation factor (GEVB) and is expressed as: $\mathbf{w}^{\mathsf{GEVB}} = \psi \mathbf{w}^{\mathsf{GEV}}$.

C. ILRMA

In BSS, the goal is to recover source signals from observed signals. Let the STFT of the source signals in each time-frequency bin be described as, $\mathbf{h}_{ij} = [h_{ij,1}, \ldots, h_{ij,U}]^{\mathsf{T}}$, where the order of sources in $\mathbf{h}_{ij}$ is unknown, reflecting the "blind" nature of BSS. This stands in contrast to the distinguishable sources described in (1). The mixing system is modelled as $\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{h}_{ij}$, where $\mathbf{A}_i$ is the mixing matrix, and each component $\mathbf{a}_{i,u}$ within $\mathbf{A}_i$ is a steering vector for each source. The separation process can be written as:

$$\mathbf{y}_{ij} = {\mathbf{W}^{\mathsf{BSS}}}_i \mathbf{x}_{ij}, \tag{7}$$

where $\mathbf{y}_{ij} = [y_{ij,1}, \ldots, y_{ij,U}]^{\mathsf{T}}$ are the separated signals, and ${\mathbf{W}^{\mathsf{BSS}}}_i = \left[{\mathbf{w}^{\mathsf{BSS}}}_{i,1}, \ldots, {\mathbf{w}^{\mathsf{BSS}}}_{i,U}\right]^{\mathsf{H}}$ is the demixing matrix. In a determined problem where $\mathbf{A}_i$ is invertible, ideal ${\mathbf{W}^{\mathsf{BSS}}}_i$ is defined as $\mathbf{A}_i^{-1}$.

In ILRMA, the separated signal is assumed to follow a time-varying complex Gaussian distribution with a zero mean and variance $l_{ij,u}$, corresponding to $y_{ij,u}$ [7]. $l_{ij,u}$, is arranged in a two-dimensional array $\mathbf{L}_u \in \mathbb{R}_+^{I \times J}$ representing the spectral model for the u-th source [15]. Similar to non-negative matrix factorization (NMF), $\mathbf{L}_u$ can be represented as $\mathbf{L}_u = \mathbf{T}_u \mathbf{V}_u$, where $\mathbf{T}_u \in \mathbb{R}_+^{I \times K}$ and $\mathbf{V}_u \in \mathbb{R}_+^{K \times J}$ are non-negative matrices with K NMF bases. The cost function for ILRMA, expressing the negative log-likelihood function, is given as:

$$\mathcal{J} = \sum_{ij,u} \left[\frac{\left| \mathbf{w}^{\mathsf{BSS}\mathsf{H}}_{i,u} \mathbf{x}_{ij} \right|^2}{l_{ij,u}} + \log l_{ij,u} \right] - 2J \sum_i \log \left| \det \mathbf{W}^{\mathsf{BSS}}{}_i \right|. \tag{8}$$

Here, constant terms have been omitted for simplicity.

To further improve the separation accuracy and optimisation stability, a method called spatially regularised ILRMA (SR-ILRMA) have been proposed [16]. It incorporates a regularisation term into the ILRMA framework, thereby resolving the global permutation ambiguity problem while improving the convergence of the cost function during iterative optimisation [16]. The global permutation ambiguity problem refers to the uncertainty regarding which output channel corresponds to the target speech, a challenge that arises from the random channel allocation in the original ILRMA method. By resolving this ambiguity, SR-ILRMA enhances the reliability of source separation. The SR-ILRMA cost function, denoted by $\mathcal{J}_R$, extends the original ILRMA cost function $\mathcal{J}$ by incorporating a regularisation term. This term is scaled by a weight parameter γ_u and quantifies the difference between the current estimate $\mathbf{w}^{\mathsf{BSS}}{}_{i,u}$ and a supervisor matrix $\widehat{\mathbf{w}}^{\mathsf{BSS}}_{i,u}$, as follows:

$$\mathcal{J}_{\mathrm{R}} = \mathcal{J} + \sum_{i,u} \gamma_u \| \mathbf{w}^{\mathsf{BSS}}{}_{i,u} - \widehat{\mathbf{w}}^{\mathsf{BSS}}_{i,u} \|^2. \tag{9}$$

The update rule and detailed description for minimising the SR-ILRMA cost function, $\mathcal{J}_R$, can be found in [16].

III. Proposed Method

In UAV audition, extracting a clear target signal is challenging due to the low SNR [5]. To address this, a novel framework is proposed that combines several established approaches to maximise enhancement performance. First, informed and supervised approaches are integrated, using a pre-recorded noise database and a GEVB beamformer to create beamformer with practical utility. The filter weights of this beamformer are then employed to inform ILRMA through spatial regularisation, improving source enhancement and addressing the global permutation ambiguity problem inherent to ILRMA.

The proposed framework is illustrated in Figure 1, with the input and pre-processing elements depicted in grey. The spatial covariance matrices, marked in green, are discussed in Section III-A. The beamformer-informed ILRMA component, highlighted in blue, is detailed in Section III-B.

A. Spatial covariance matrix

The noise spatial covariance matrix, $\mathbf{R}_{\mathrm{n}}$, is crucial for both the beamformer and the spatially regularised ILRMA. Assuming that the spatial characteristics of the UAV ego-noise are stationary due to the fixed position of the microphone array [5], $\mathbf{R}_{\mathrm{n}}$ can be estimated using a sampled noise covariance matrix, here i is omitted for simplicity:

$$\mathbf{R}_{\mathrm{n}} = \frac{1}{JP} \sum_{j=1}^{J} \sum_{p=1}^{P} \mathbf{x}_{j,p} \mathbf{x}^{\mathsf{H}}_{j,p}, \tag{10}$$

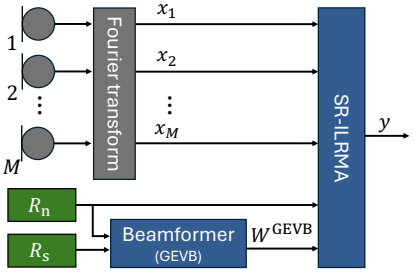

Fig. 1: Block diagram of the framework. For clarity, the frequency and frame indices have been omitted.

where $\mathbf{x}_p$ is the p-th noise sample signal, with $p = 1, ..., P$ indicating the noise sample index. By leveraging the fixed spatial relationship between the microphone array and the UAV noise sources, $\mathbf{R}_{\mathrm{n}}$ can be pre-calculated using pre-recorded rotor noise data. This method eliminates the need for complex methods, such as voice activity detection (VAD) or machine learning techniques, which are commonly employed to estimate the spatial covariance matrix in sound source enhancement for UAVs. This simplifies implementation and enhances practicality.

The target spatial covariance matrix, $\mathbf{R}_{\mathrm{s}}$, is used in the beamformer. It is derived by considering the target source's varying positions relative to the microphone array during UAV operations. Figure 2 illustrates the relative position of the target source and the UAV from both top and side views. The top view assumes the azimuth angle relative to the UAV direction is zero, indicating the target source is directly ahead. The side view shows the elevation angle varying within a predefined range. This assumption enhances the framework's robustness in real-world scenarios where the precise elevation angle is often unknown. $\mathbf{R}_{\mathrm{s}}$ is calculated as:

$$\mathbf{R}_{\mathrm{s}} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{a}_q \mathbf{a}^{\mathsf{H}}_q. \tag{11}$$

Here, $\mathbf{a}_q$ denotes the steering vector for various potential positions of the target source relative to the microphone array, with $q = 1, ..., Q$ indicating the target position index. These steering vectors are pre-measured.

B. Beamformer informed ILRMA

Beamformer informed ILRMA aims to enhance separation performance and address the global permutation ambiguity in ILRMA. The proposed method uses beamformer filter weights to inform ILRMA by embedding spatial information of sound sources into the demixing process and dedicating one output channel specifically for the target source. This strategy ensures consistent allocation of the target source to a designated channel, reducing uncertainty in standard ILRMA outputs.

The beamformer in this framework employs the GEVB beamformer described in Section II-B and uses the spatial covariance matrix detailed in Section III-A to calculate the filter

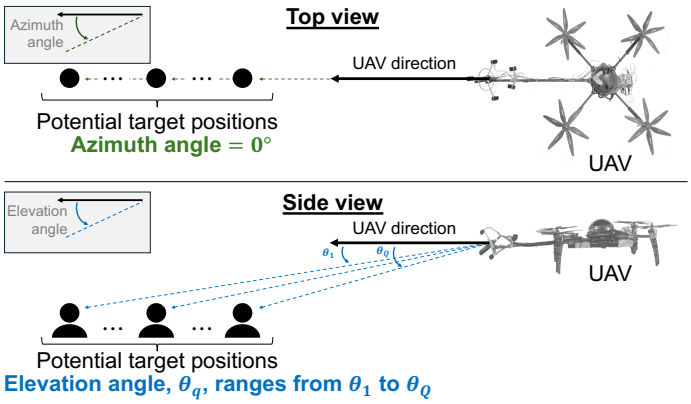

Fig. 2: Target source positions: azimuth (top view) and elevation (side view).

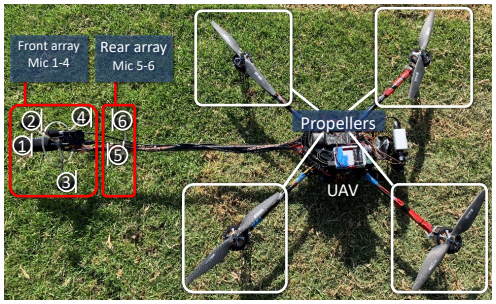

Fig. 3: UAV setup (from [5]). Red boxes highlight microphone arrays, white circles show microphone numbers, and white boxes highlight propellers.

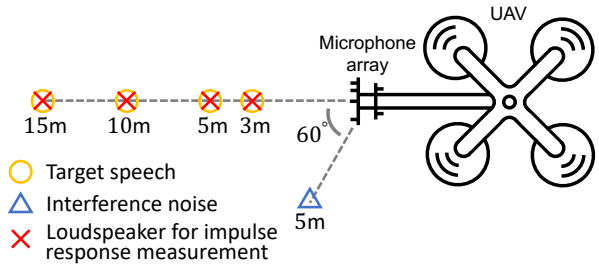

Fig. 4: Illustration of the relative positions.

weight $\mathbf{w}^{\text{GEVB}}$. This approach has shown robust performance across various target positions in a previous study [8].

Subsequently, ILRMA is informed by the beamformer through supervisor matrix, $\widehat{\mathbf{W}}_i^{\text{BSS}}$, derived via a multi-step process using $\mathbf{R}_{\text{n}}$ and $\mathbf{w}^{\text{GEVB}}$. This process begins with the eigenvalue decomposition of $\mathbf{R}_{\text{n}}$ into its eigenvectors and eigenvalues sorted in descending order. The first $U-1$ eigenvectors, corresponding to the largest eigenvalues, are selected to capture the most significant noise components. These eigenvectors and the filter weights $\mathbf{w}^{\text{GEVB}}$ are then used to construct $\widehat{\mathbf{W}}_i^{\text{BSS}}$. Specifically, $\widehat{\mathbf{W}}_i^{\text{BSS}}$ incorporates the first $U-1$ eigenvectors for the initial $U-1$ channels and $\mathbf{w}^{\text{GEVB}}$ for the U-th channel. This configuration provides spatial information on the dominant noise components and the target source while dedicating the final output channel specifically for the target source. Finally, $\widehat{\mathbf{W}}_i^{\text{BSS}}$ is used in the cost function for SR-ILRMA.

IV. Experimental Results

The proposed method was evaluated in real-world scenarios with varying target speech positions and sound levels. Performance was measured using objective metrics such as the source-to-distortion ratio (SDR) [17] and short-time objective intelligibility (STOI) [18]. Additionally, SDR improvement (SDRi) and STOI improvement (STOIi) were assessed based on the differences between output and input performance metrics.

A. Experiment

The proposed framework was evaluated using UAV recordings detailed in a previous study [5]. The system's microphone array includes six microphones: four in the front sub-array (Microphone 1 is a shotgun, Microphones 2, 3, and 4 are cardioids) and two cardioid microphones in the rear sub-array (Microphones 5 and 6) as shown in Figure 3. The database included the followings:

- **Target speech:** ten sentences from the Centre for Speech Technology Research VKTS Corpus [19], played at 60 dBA and 80 dBA (measured at one metre from the loudspeaker), and recorded at distances of 3 m, 5 m, 10 m, and 15 m, yielding 80 sets of target speech.
- **UAV noise:** 64 recordings of hovering UAV ego-noise without the target signal. Five sets were used for performance evaluation, and 59 sets for estimating the noise covariance matrix.
- **Interference noise:** background music and traffic noise recorded at 60 dBA from a distance of 5 m at a 60-degree angle to the front of the microphone array (Figure 4).
- **Impulse response:** measured at 3 m, 5 m, 10 m, and 15 m to generate the estimated spatial covariance matrix for the target source.

Recordings were conducted at Harry James Taupaki Reserve in Auckland, New Zealand, capturing ambient noises including wind, bird and insect calls [5]. Figure 4 shows the positions of the target signal, interference noise, and the loudspeaker for measuring impulse responses. These recordings enabled a comprehensive evaluation of the proposed method.

B. Results and discussion

1) Addressing global permutation ambiguity problem: The proposed method's effectiveness in correctly allocating the target output channel across various noise levels was evaluated. Accuracy was measured as the percentage of instances where the target output channel was correctly identified. Figure 5 compares the accuracy of the proposed method, beamformer-informed ILRMA (iILRMA), with the original ILRMA at different input SNRs. The original ILRMA assumes that speech signals exhibit higher kurtosis than noise signals [20] and uses kurtosis of the output signals to estimate the speech channel

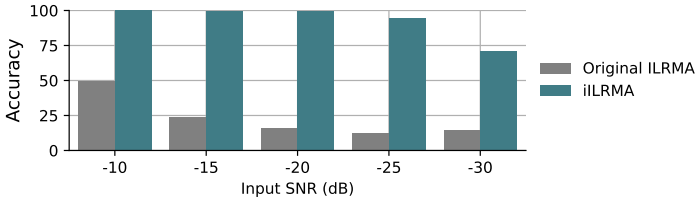

Fig. 5: Accuracy of channel allocation at varying input SNR.

[21]. The target channel is correctly identified if the channel with the highest kurtosis matches the channel with the highest SDR. In contrast, iILRMA deems the target channel correctly identified if the allocated target channel has the highest SDR among all channels. Results show that the accuracy of the original ILRMA decreases as SNR decreases, from 50% at -10 dB to 24% at -15 dB. Conversely, iILRMA achieves near-perfect accuracy at input SNRs from 0 to -20 dB, with accuracy declining to 95% at -25 dB and 71% at -30 dB. This demonstrates significant improvement over the original ILRMA and identified a critical threshold around -25 dB for the proposed method's effectiveness.

2) Performance comparison: The source enhancement performance of the proposed method is compared against the original ILRMA and GEVB beamformer. Figure 6 presents the SDRi and STOIi of various enhancement methods. The results indicate that iILRMA outperforms the original ILRMA in both SDRi and STOIi metrics. When compared to the GEVB beamformer, iILRMA shows better SDRi from -10 to -20 dB input SNR but performs worse at -25 and -30 dB input SNR. This decline in performance at lower SNRs is attributed to the reduced accuracy observed at -25 dB, as shown in Figure 5. Additionally, the beamformer's design, characterised by a distortionless magnitude response, may contribute to this outcome. For STOIi, iILRMA exhibits superior performance from -10 dB to -25 dB and comparable performance at -30 dB compared to the beamformer. The diminishing improvement in STOIi at lower SNRs can be attributed to iILRMA's reduced channel accuracy at -25 dB and below, as discussed in Section IV-B1. These results demonstrate that iILRMA provides better source enhancement performance in most scenarios.

3) Real world performance: The proposed method was evaluated under real-world conditions, unlike the scenarios with fixed input SNRs discussed in Sections IV-B1 and IV-B2. The evaluation involved varying sound levels and target speech positions to assess its robustness and performance. The results are shown in Figure 7. At a target sound level of 80 dBA, iILRMA demonstrated significant performance improvements, with SDR and STOI improvements of 13 dB and 0.21, respectively, at a target distance of 3 metres. However, at a target sound level of 60 dBA, the proposed method showed suboptimal performance, indicating that enhancing sound sources under such challenging conditions remains an area for future research. Audio samples are accessible at: https://github.com/JinXuanTeh/audio-sample-BF-ILRMA

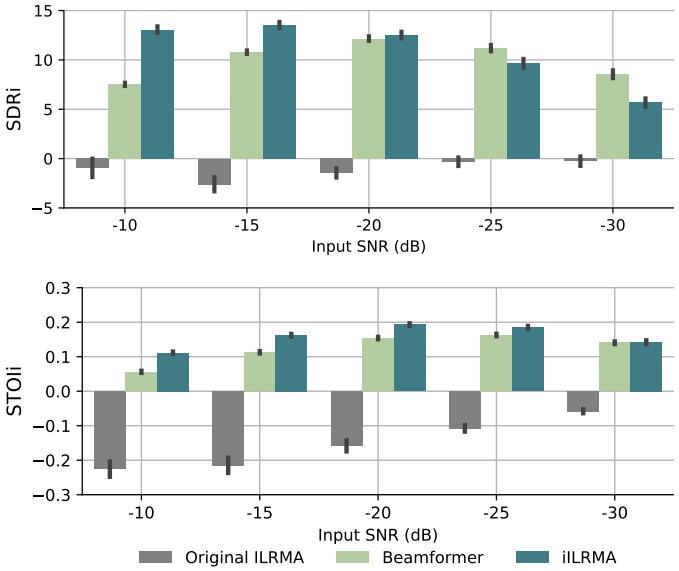

Fig. 6: Source enhancement performance of different methods with 95% confidence intervals.

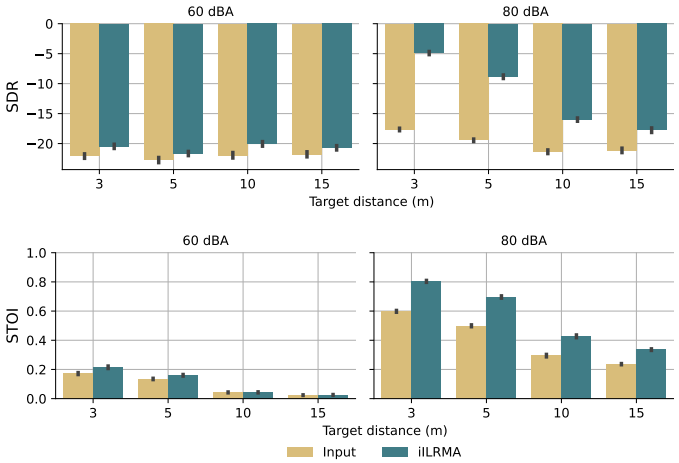

Fig. 7: Source enhancement performance in real-world conditions with 95% confidence intervals.

V. Conclusion

This study explored the integration of informed, supervised, and blind approaches to enhance sound source performance in UAV applications. The proposed framework combines a beamformer (informed approach), a pre-recorded noise database (supervised approach), and ILRMA (blind approach). Experimental results demonstrated that the proposed method outperforms both the original ILRMA and the beamformer in most scenarios. iILRMA showed substantial gains in SDR and STOI metrics, achieving an SDR improvement of 13 dB and a STOI improvement of 0.21. Additionally, the proposed method effectively addresses the global permutation ambiguity problem, ensuring more accurate channel allocation across various SNR levels. Future research should focus on enhancing the framework's performance under extremely low SNR conditions and exploring the integration of additional sound

source enhancement algorithms to further improve robustness and effectiveness in diverse environments. This comprehensive approach demonstrates the potential for superior performance through the integration of different methods.

Acknowledgment

This research was partially funded by the Royal Society of New Zealand Catalyst Seeding programme: New Zealand – Japan Joint Research Projects (JSP-UOA1901-JR), the Kajima Foundation's Support Program for International Joint Research Activities (2024-kyodoshin-05) and the Acoustics and Vibration Research Centre at the University of Auckland.

References

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

[2] D. P. Nowacek, F. Christiansen, L. Bejder, J. A. Goldbogen, and A. S. Friedlaender, "Studying cetacean behaviour: New technological approaches and conservation applications," en, *Animal Behaviour*, vol. 120, pp. 235–244, 2016.

[3] K. Nakadai, M. Kumon, H. G. Okuno, *et al.*, "Development of microphone-array-embedded uav for search and rescue task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5985–5990.

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

[6] 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," en, *Applied Acoustics*, vol. 216, p. 109 818, 2024.

[7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] J. X. Teh, B. Yen, M. Kingan, and Y. Hioka, "Source enhancement with different mvdr beamformer designs for unmanned aerial vehicle audition," 2024.

[9] 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.

[10] M. Brandstein and D. Ward, *Microphone Arrays*. Springer Berlin Heidelberg, 2001.

[11] J. Li and P. Stoica, *Robust adaptive beamforming* (Wiley Series in Telecommunications and Signal Processing ; v.88), eng. Hoboken, NJ: John Wiley, 2006.

[12] H. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory* (Detection, Estimation, and Modulation Theory). Wiley, 2004.

[13] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[14] G. Golub and C. Van Loan, *Matrix Computations* (Johns Hopkins Studies in the Mathematical Sciences). Johns Hopkins University Press, 2013.

[15] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," en, *APSIPA Transactions on Signal and Information Processing*, vol. 8, no. 1, 2019.

[16] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 746–750, 2018.

[17] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, *First stereo audio source separation evaluation campaign: data, algorithms and results*. Springer Berlin Heidelberg, 2007, pp. 552–559.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[19] J. Yamagishi, V. Christophe, and M. Kirsten, *CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, Sound, Data retrieved from Edinburgh DataShare, Nov. 2019.

[20] G. Li, M. E. Lutman, S. Wang, and S. Bleeck, "Relationship between speech recognition in noise and sparseness," en, *International Journal of Audiology*, vol. 51, no. 2, pp. 75–82, 2011.

[21] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Blind speech extraction based on rank-constrained spatial covariance matrix estimation with multivariate generalized Gaussian distribution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1948–1963, 2020.