Acoustic Sensing From a Multi-Rotor Drone

Lin Wang and Andrea Cavallaro

***Abstract*—We propose a time-frequency processing method that localizes and enhances a target sound by exploiting spectral and spatial characteristics of the ego-noise captured by a microphone array mounted on a multi-rotor micro aerial vehicle. We first exploit the time-frequency sparsity of the acoustic signal to estimate at each individual time-frequency bin the *local* direction of arrival (DOA) of the sound and formulate spatial filters pointing at a set of candidate directions. Then we combine a kurtosis measure based on the spatial filtering outputs and a histogram measure based on the local DOA estimation to calculate a spatial likelihood function for source localization. Finally, we enhance the target sound by formulating a time-frequency spatial filter pointing at the estimated direction. As the ego-noise generally originates from specific directions, we propose a DOA-weighted spatial likelihood function that improves source localization performance by identifying noiseless sectors in the DOA circle. The DOA weighting scheme localizes the target sound even in extremely low signal-to-noise conditions when the target sound comes from a noiseless sector. We experimentally validate the performance of the proposed method with two array placements.**

***Index Terms*—Acoustic sensing, ego-noise reduction, micro aerial vehicle, microphone array, source localization**

I. Introduction

Multi-microphone acoustic sensing from a multi-rotor drone (or MAV: micro aerial vehicle) aims to record, localize and analyze sounds emitted by aerial or ground objects [1], [2]. Potential applications include recreational video capturing and broadcast, search and rescue, and surveillance [3]–[8]. The rotating motors and propellers generate strong ego-noise [9], which masks the target sound, degrades the sound quality and leads to extremely-low signal-to-noise ratios (e.g. SNR<-15 dB). The nonstationary spectrum of the ego-noise depends on the rotation speed of each motor, which changes over time [10]. Moreover, the microphones move with the MAV thus leading to a dynamic acoustic mixing network. Finally, the natural and motion-induced wind increases the noise captured by the microphones. All these issues make MAV-based acoustic sensing a very challenging task.

Most microphone-array noise reduction techniques are suitable for indoor sound processing when the input SNRs are relatively high [11]. Based on the observation that a target sound and the ego-noise usually have concentrated energy at sparsely isolated time-frequency bins, we proposed a time-frequency filtering approach [12], which formulates a spatial filter to enhance a target direction based on local direction of arrival (DOA) estimates at individual time-frequency bins. This approach works robustly under strong ego-noise but has two limitations. First, the performance of the time-frequency filter drops significantly when the target sound arrives from a direction close to that of the ego-noise. Second, to steer the spatial filter it needs knowledge of the DOA of the target sound, which is difficult to estimate due to the extremely low SNR and the nonstationarity of the ego-noise.

To address these problems, starting from a time-frequency processing framework [13], we propose a new source localization and enhancement method. The proposed method estimates the DOA of a sound by detecting the peak of a spatial likelihood function generated by combining the histogram of the local DOA estimates at individual time-frequency bins and a kurtosis measure computed by steering time-frequency spatial filters at a set of candidate directions. The combination of these two measures improves robustness to low SNRs and nonstationarity of the ego-noise. In addition, we divide the DOA circle into noisy and noiseless sectors and propose a DOA weighting scheme when calculating the spatial likelihood function. This scheme improves source localization performance when the target sound arrives from a noiseless sector. Finally, the proposed method steers the time-frequency spatial filter towards the estimated direction of the sound source.

The paper is organized as follows. Section II reviews the state of the art for MAV-based acoustic sensing. Section III formulates the problem and Section IV investigates the spectral and spatial characteristics of the ego-noise. Section V proposes the source localization and sound enhancement method. Experiments are conducted in Section VI and conclusions are drawn in Section VII.

II. Related work

MAV-based acoustic sensing approaches can be classified based on their strategy as supervised or unsupervised. Moreover, they can be grouped based on the task, e.g. source localization and sound enhancement (Table I).

Supervised approaches use additional sensors to monitor (i.e. to supervise) the status of the MAV in order to predict the ego-noise. Since the MAV ego-noise mainly consists of harmonic components whose fundamental frequency is proportional to the motor rotation speed, supervised approaches build a noise template database from which the spectrum [14] or the correlation matrix [15] of the ego-noise can be predicted depending on the MAV behaviour. The predicted ego-noise spectrum can be used to design a single-channel spectral filter for noise reduction [14]. The predicted noise correlation matrix can be used to design a GEVD-MUSIC (generalized eigenvalue decomposition - multiple signal classification) algorithm

Manuscript received: April 9, 2018

This work was supported by the ARTEMIS-JU and the U.K. Technology Strategy Board (Innovate U.K.) through the COPCAMS Project under grant 332913.

The authors are with Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, U.K. (e-mail: lin.wang@qmul.ac.uk; a.cavallaro@qmul.ac.uk).

TABLE I
ACOUSTIC SENSING METHODS FOR DRONES. KEY: S - SUPERVISED; U - UNSUPERVISED; d - DIAMETER.

Task	Strategy	Method	Ref.	MAV		Sensors	
				Multi-rotor	Fixed-wing	Recording microphones	Monitoring sensors
Localization	S	template-based	[15]	AR-Drone		8-mic circular array d = 30 cm	flight status sensors motor rotation sensors
	U	SRP-PHAT	[4], [20]		prototype	4-mic tetrahedron array d = 10 cm	
		GEVD-MUSIC	[3], [21]	AR-Drone		8-mic circular array d = 30 cm	
			[8]	Mini Surveyor		12-mic spherical array d = 10 cm	
Enhancement	S	template-based	[14]		prototype	1 recording mic	motor rotation sensor
		reference-based	[5], [18]	AR-Drone		1 recording mic	4 reference mics
			[19]	nano drone		1 recording mic	piezoelectric sensors
	U	fixed beamforming	[24], [25]	Zion PG560		16-mic octagonal array d = 2 m	
			[26]	prototype		1 shortgun mic 4 unidirectional mics	
		blind source separation	[9], [12]	3DR Iris		8-mic circular array d = 20 cm	
		time-frequency filtering	[12]				
Localization & Enhancement	U	DOA-weighted kurtosis-histogram measures + time frequency filtering	**proposed**	3DR Iris		8-mic circular array d = 20 cm	

for noise-robust sound source localization [15]. The predicted noise correlation matrix has also been used to design a multi-channel beamformer to suppress the ego-noise of a ground robot [16], [17]. However, application to flying MAVs has not been reported yet. Reference microphones installed close to the propellers can also be used to pick up motor noises that are then adaptively cancelled [5], [18], [19]. This approach usually requires the use of insulation materials to prevent the reference microphones from picking up the target sound. Supervised approaches usually perform robustly under strong ego-noise. However, the need for dedicated monitoring sensors limits the versatility of these approaches.

Unsupervised approaches perform acoustic sensing using microphone signals only. Due to the nonstationarity of the ego-noise and the extremely low SNR, it is a challenging task to estimate the direction of the target sound from the noisy recording. Steered response power with phase transform (SRP-PHAT) [4], [20] and multiple signal classification (MUSIC) [3], [21] have been applied to MAV-based source localization. SRP-PHAT exploits the correlation of microphone signals and computes a spatial likelihood map with peaks that correspond to the locations of the target sound sources [22]. SRP-PHAT has been widely used for source localization in high-SNR scenarios. However, for MAV-based applications with low SNRs, the performance of SRP-PHAT degrades considerably because the coherence of the target sound is masked by strong ego-noise. MUSIC is a subspace-based high-resolution localization algorithm that is widely employed for robot audition [11], [23]. Using eigenvector decomposition, MUSIC decomposes an observed noisy signal into the signal subspace and noise subspace, which are orthogonal to each other, and then computes a spatial spectrum of the locations, i.e. the MUSIC spectrum, with peaks at the locations corresponding to the target sound sources. The standard MUSIC algorithm assumes noise signals to be uncorrelated between microphones to easily discriminate signal and noise subspaces. In practice the discrimination is difficult especially when the noise is directional and stronger than the target sound. GEVD-MUSIC exploits as additional information a noise correlation matrix to improve robustness to noise. Although several approaches have been proposed to blindly estimate the noise correlation matrix from the microphone signal [3], [8], [21], the nonstationarity of the ego-noise makes the estimation inaccurate.

Delay-and-sum (fixed) beamforming is a typical unsupervised approach to enhance sounds from a desired location by coherently delaying and summing multi-channel microphone signals [24]. Relying only on the array geometry and the target sound location, fixed beamforming is robust to low SNRs and MAV movement. However, to get satisfactory noise reduction performance it usually needs a large-size array with many microphones, e.g. an octagonal array with 16 microphones and 2 m diameter [24], [25]. A single-channel post-filter is thus employed to further enhance the beamforming output [26].

Another technique for ego-noise reduction is blind source separation (BSS) [9], [12]. BSS can reduce the ego-noise more effectively than fixed beamforming and does not require the knowledge of the locations of the microphones and the target sound sources. However, the performance of BSS degrades in a dynamic scenario with moving microphones. Moreover, due to permutation ambiguities [27], [28], the target sound is usually extracted into one of the several output channels, whose channel index is unknown. The detection of this target channel is still an open problem [12].

Time-frequency processing, a recently proposed approach, exploits the sparsity of the acoustic signal in the time-frequency domain to design a spatial filter that enhances the signal from a desired direction [12]. While this approach works robustly under ego-noise, it requires the knowledge of the target direction. Moreover, its performance is sensitive to the direction of the target sound. In this paper, we propose a method that address both issues.

III. Problem Formulation

Let a circular array with M microphones be mounted on a multi-rotor MAV. Let the locations of the microphones be $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_M]$, where $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\mathrm{T}}$ is the position of the m-th microphone in a 2D coordinate system and the superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operator.

In MAV-based applications, the distance between the target sound source and the microphone array is usually much larger than the array size. It is thus reasonable to assume a far-field model for the target sound source, whose sound wave impinges on the array in the form of planar waves [29]. Let a target source lie in the far field emitting sound with DOA θ_d. We assume a low-reverberant environment without natural wind and that the relative positions of microphones and sound source are fixed (e.g. the MAV hovers stably while recording the sound from a static source).

The microphone signal, $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^{\mathrm{T}}$, contains both the target sound, $\boldsymbol{s}(n) = [s_1(n), \cdots, s_M(n)]^{\mathrm{T}}$, and the ego-noise, $\boldsymbol{v}(n) = [v_1(n), \cdots, v_M(n)]^{\mathrm{T}}$, i.e.

$$\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{v}(n), \tag{1}$$

or, written in the short-time Fourier transform (STFT) domain:

$$\boldsymbol{x}(k,l) = \boldsymbol{s}(k,l) + \boldsymbol{v}(k,l), \tag{2}$$

where k and l are the frequency and frame indices, respectively. Let K and L be the total number of frequency bins and time frames, respectively.

Given only $\boldsymbol{x}(n)$ and $\boldsymbol{R}$, our goal is to estimate the DOA of the target sound $\hat{\theta}_d$ and to design a spatial filter $\boldsymbol{w}(k,l) = [w_1(k,l), \cdots, w_M(k,l)]^{\mathrm{T}}$ that extracts the target sound from the noisy recording via

$$y(k,l) = \boldsymbol{w}^{\mathrm{H}}(k,l)\boldsymbol{x}(k,l), \tag{3}$$

where the superscript $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose.

To achieve the goal, we built a prototype composed of a 3DR IRIS quadcopter and an 8-microphone circular array with diameter $d = 0.2$ m [9]. The array is placed on the top side of the MAV body (at a distance of 0.15 m) in order to avoid the self-generated wind blowing downwards from the propellers. We consider two configurations when mounting the array on the MAV (Fig. 1). The first configuration (Array-C) is more compact and the array is mounted close to the middle of the MAV body, centring the four motors. In the second configuration (Array-F), the array is mounted close to the front end of the MAV body, but it is still inside the critical collision protection area [30].

The main challenge is the extremely low SNR at the microphones. Fig. 2 gives an example of the input SNR at the microphones for a human talking aloud at a varying distances, from 2 m to 6 m from the MAV. The input SNR was measured with the prototype shown in Fig. 1(c), with the MAV fixed on a tripod at a height of 1.8 m and operating at 50%, 100% and 150% of the hovering power. When the MAV is operating at the hovering status, the input SNR can be lower than -20 dB: this is extremely challenging for state-of-the-art sound enhancement and source localization algorithms.

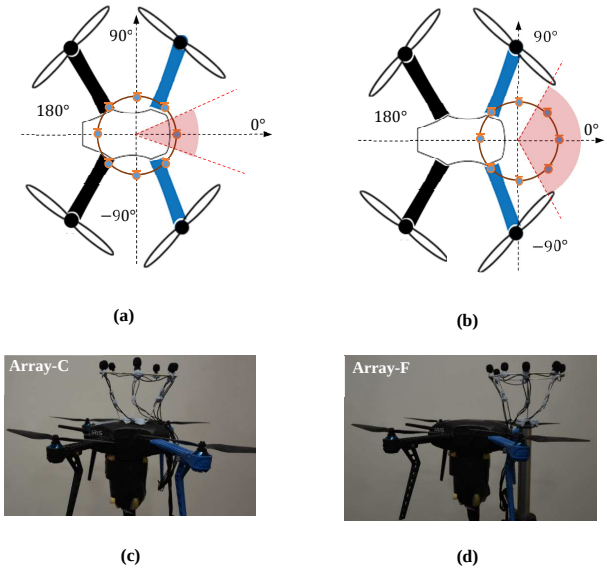


Fig. 1. Two array placements on the hardware prototype. (a)(c) Array-C: the array is placed close to the centre of the MAV body. (b)(d) Array-F: the array is placed close to the front end of the MAV body. The noiseless sector is illustrated with a shadowed area.

IV. Ego-noise: Spectral and Spatial Characteristics

Since the relative positions of microphones and ego-noise sources (i.e. motors and propellers) are fixed, prior knowledge on the ego-noise would be helpful for choosing an appropriate sound processing algorithm. To this end, we investigate the spectral and spatial characteristics of the ego-noise.

Fig. 3 depicts the time-frequency spectrum of a segment of ego-noise recorded with a microphone randomly chosen from Array-C. The sampling rate is 8 kHz and the time-frequency spectrum is obtained by applying STFT with a window size

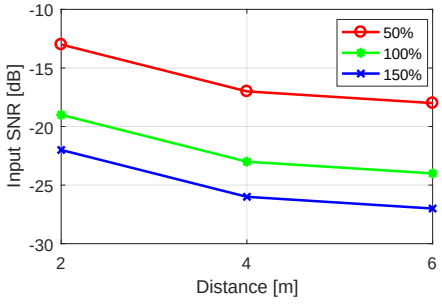


Fig. 2. Input SNR at onboard microphones for a human speaker talking aloud at a varying distance (from 2 m to 6 m) to the MAV. The MAV is fixed on a tripod and operating stably at 50%, 100%, and 150% of the hovering power.

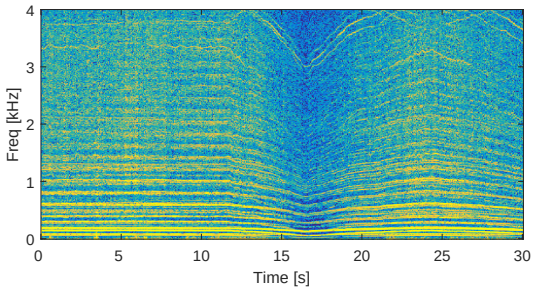


Fig. 3. Time-frequency spectrum of a segment of ego-noise lasting 30 seconds. The rotation speed of the motors remains constant in the first 12 seconds and then varies randomly in the following 18 seconds.

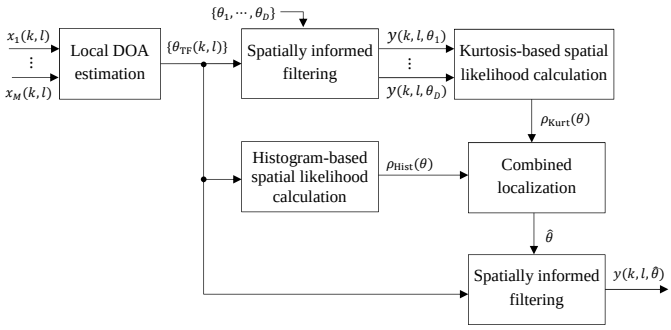


Fig. 5. Block diagram of the proposed method.

1024 and half overlap. The rotation speed of the motors is constant in the first 12 seconds and then varies randomly in the following 18 seconds. The ego-noise mainly consists of multiple narrow-band harmonic noise (the mechanical sound of the rotating motors) and broadband noise (the rotating propellers cutting air) [9]. The fundamental frequency of the harmonic noise typically varies with the motor rotation speed, leading to nonstationary spectrum. The harmonic noise presents evident time-frequency sparsity with energy peaks at isolated frequency bins. The ego-noise typically shows high correlation at these harmonic frequencies (a detailed analysis is presented in [9]). This time-frequency sparsity can be exploited to design a time-frequency spatial filter that enhances the sound from a desired direction [31].

In both cases (Array-C and Array-F) the microphones are close to the noise sources (i.e. motors and propellers) and thus present similarly low SNR. However, the ego-noise does show different spatial characteristics for the two array placements. To verify this, we build a histogram of the *local* DOA estimates at individual time-frequency bins (see Sec. V-A for the details of local DOA estimation). Fig. 4 compares the DOA histograms for different noise segments recorded by Array-C and Array-F. Interestingly, for each noise segment, the whole DOA circle $[-180°, 180°]$ can be divided into two sectors: a noisy sector and a noiseless sector, with the latter one indicated with red arrows in Fig. 4. For each noisy segment, the DOA histogram shows high values in the noisy sector, and shows low values in the noiseless sector. Let us compare the two segments recorded by Array-C, i.e. in Fig. 4(a) and (c). The histogram of the ego-noise has high values in the noisy sector, with roughly four peaks. While the shape of the histogram differs for the two segments, the locations of the four peaks remain almost unchanged. The ego-noise consists of the motor and propeller sounds. The directions of the motor sound correspond to the four peaks of the DOA histogram and remain unchanged with time. The propeller sound originates from the swept area of the rotating blades and its direction spreads widely within the noisy sector and around the four DOA peaks. The variation of the DOA histogram plot across segments implies that the DOA of the propeller sound at individual time-frequency bins changes as consequence of the MAV behaviour.

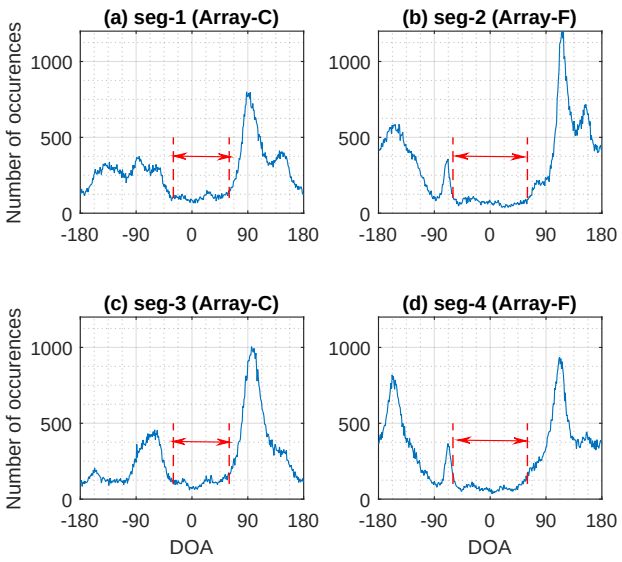


Fig. 4. Histogram of the DOA estimates at individual time-frequency bins for a segment of ego-noise lasting 30 seconds. (a)(c) Two different segments for Array-C. (b)(d) Two different segments for Array-F. The noiseless sector is indicated with red arrows.

The histogram of the ego-noise has instead low values in the noiseless sector. These values remain almost constant for the two signal segments. Similar observations can be made for the two ego-noise segments recorded by Array-F, i.e. in Fig. 4(b) and (d). One contrast is that the width of the noiseless sector for Array-F is larger than that for Array-C. In Fig. 1(a) the ego-noise tends to arrive from the directions around Array-C, while in Fig. 1(b) the ego-noise tends to arrive from the back side of Array-F (i.e. the side closer to the motors), thus leading to a larger noiseless sector. Considering the low probability of the ego-noise from the noiseless sector, we presume that a target sound could get detected more easily and accurately if it arrives from the noiseless sector.

V. Proposed Method

We propose a method for joint source localization and sound enhancement based on the spectral and spatial characteristics of the ego-noise discussed in the previous section. The proposed method estimates the DOA *locally* at individual time-frequency bins and performs source localization with a DOA-weighted combination of a histogram and a kurtosis measure.

The source localization result then leads to a spatially informed time-frequency spatial filter (Fig. 5).

A. Local DOA estimation

The observation that the target sound (e.g. human speech) and the ego-noise usually present energy peaks sparsely in the time-frequency domain allows us to estimate the DOA at individual time-frequency bins. Given the microphone signal $\boldsymbol{x}(k,l)$ and the microphone locations $\boldsymbol{R}$, the DOA of the sound at each time-frequency bin can be estimated by building a local generalized cross correlation (GCC) function [12], [32]

$$\gamma_{\text{TF}}(k,l,\theta) = \Re\left\{\sum_{\substack{m_1,m_2=1\\ m_1\neq m_2}}^{M} \frac{x_{m_1}(k,l)x^*_{m_2}(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|} \mathrm{e}^{j2\pi f_k \tau(m_1,m_2,\theta)}\right\}, \tag{4}$$

where f_k denotes the frequency at the k-th bin, the superscript $(\cdot)^*$ denotes the complex conjugation, and the operator $\Re\{\cdot\}$ denotes the real component of the argument. The term $\tau(m_1,m_2,\theta) = \frac{\|\boldsymbol{r}_{m_2}-\boldsymbol{r}_\theta\|-\|\boldsymbol{r}_{m_1}-\boldsymbol{r}_\theta\|}{c}$ denotes the delay between two microphones m_1 and m_2 with respect to the sound coming from θ, where c is the velocity of sound and $\boldsymbol{r}_\theta$ is the location of the far-field sound source from direction θ, and can be approximated as $r_\theta = [D\sin\theta, D\cos\theta]^T$, where $D \gg d$ is set as 10 m for ease of computation. The local DOA of the sound at the (k,l)-th bin can then be determined as

$$\theta_{\text{TF}}(k,l) = \underset{\theta\in(-180^\circ,180^\circ]}{\arg\max}\ \gamma_{\text{TF}}(k,l,\theta). \tag{5}$$

B. Histogram-based spatial likelihood

We use the local localization results at all time-frequency bins, $\{\theta_{\text{TF}}\}$, to build a histogram-based spatial likelihood function:

$$\tilde{\rho}_{\text{Hist}}(\theta) = \mathcal{H}(\{\theta_{\text{TF}}\}), \tag{6}$$

where $\mathcal{H}(\cdot)$ denotes the histogram. We normalize (6) as

$$\rho_{\text{Hist}}(\theta) = \mathcal{N}(\tilde{\rho}_{\text{Hist}}(\theta)) = \frac{\tilde{\rho}_{\text{Hist}}(\theta)}{\max(\tilde{\rho}_{\text{Hist}})}, \tag{7}$$

where $\max(\cdot)$ is the maximum value of the sequence, and $\mathcal{N}(\cdot)$ is the normalization procedure.

C. Kurtosis-based spatial likelihood

A target sound usually shows a higher non-Gaussianity, as measured by its statistical kurtosis value [33], than an ego-noise. In extremely low-SNR scenarios, the microphone signal is dominated by the ego-noise and thus presents a lower non-Gaussianity. If a spatial filter can extract the target sound by suppressing the ego-noise, the output tends to show a higher non-Gaussianity. Based on this assumption we formulate multiple spatial filters pointing at a set of candidate directions, $\theta \in \{\theta_1,\cdots,\theta_D\}$, and use the kurtosis of the spatial filtering outputs to indicate the spatial likelihood of the target sound.

We use a time-frequency approach to design the spatial filter [12], which is based on the localization results at individual time-frequency bins. To formulate a spatial filter pointing at direction θ, we first measure the closeness of each time-frequency bin (k,l) to the direction θ. Assuming the DOA estimates to be Gaussian-distributed with mean θ and standard deviation σ, the closeness measure is defined as

$$c_d(k,l,\theta) = \exp\left(-\frac{(\theta_{\text{TF}}(k,l)-\theta)^2}{2\sigma^2}\right), \tag{8}$$

where the scalar $c_d(\cdot)\in[0,1]$. The higher $c_d(\cdot)$, the higher the probability that the sound at the (k,l)-th bin arrives from direction θ. Next, we calculate an $M\times M$ target correlation matrix of the direction θ as

$$\boldsymbol{\Phi}_{ss}(k,l,\theta) = \frac{1}{L}\sum_{l=1}^{L} c_d^2(k,l,\theta)\boldsymbol{x}^{\text{H}}(k,l)\boldsymbol{x}(k,l), \tag{9}$$

where $c_d(\cdot)$ is the contribution of the (k,l)-th bin to the correlation matrix [31]. With this target correlation matrix, we formulate a standard Multichannel Wiener filter (MWF) [34]

$$\boldsymbol{w}_{\text{TF}}(k,l,\theta) = \boldsymbol{\Phi}_{xx}^{-1}(k,l)\boldsymbol{\phi}_{ss1}(k,l,\theta), \tag{10}$$

where $\boldsymbol{\phi}_{ss1}(k,l,\theta)$ is the first column of $\boldsymbol{\Phi}_{ss}(k,l,\theta)$, and $\boldsymbol{\Phi}_{xx}(k,l)$ is the correlation matrix of the microphone signal, which can be estimated directly using $\boldsymbol{\Phi}_{xx}(k,l) = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}(k,l)^{\text{H}}\boldsymbol{x}(k,l)$. The sound coming from the direction θ is extracted as

$$y_{\text{TF}}(k,l,\theta) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(k,l,\theta)\boldsymbol{x}(k,l). \tag{11}$$

We calculate the kurtosis value $\xi(k,\theta)$ of the time sequence in each frequency bin:

$$\xi(k,\theta) = \mathcal{K}(\tilde{\boldsymbol{y}}_{\text{TF}}(k,\theta)), \tag{12}$$

where $\tilde{\boldsymbol{y}}_{\text{TF}}(k,\theta)$ denotes the time sequence $|y_{\text{TF}}(k,:,\theta)|$ and $\mathcal{K}(\cdot)$ denotes the kurtosis value of the sequence. Averaging the whole frequency band, the spatial likelihood function is obtained as

$$\tilde{\rho}_{\text{Kurt}}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\xi(k,\theta), \tag{13}$$

which is further normalized as

$$\rho_{\text{Kurt}}(\theta) = \mathcal{N}\left(\tilde{\rho}_{\text{Kurt}}(\theta)\right). \tag{14}$$

D. Source localization and time-frequency spatial filtering

The measures discussed in Sections V-B and V-C are complementary. The *kurtosis-based measure* can detect the target sound in extremely low-SNR scenarios. However, when the ego-noise is nonstationary, the spatial filter tends to present a high kurtosis when pointing at the ego-noise direction and extracting the time-varying harmonic components. This produces spurious peaks on the spatial likelihood function and leads to ambiguities on determining the target sound direction. The *histogram-based measure* relies mainly on the spatial information and is robust to the nonstationarity of the acoustic signal. However, the performance of this measure degrades in

low-SNR scenarios, when the ego-noise masks the histogram peak from the target sound.

Because of their complementarity, we combine the measures to improve target sound localization and define a new spatial likelihood function as:

$$\rho_{\text{HisK}}(\theta) = \alpha\rho_{\text{Hist}}(\theta) + (1-\alpha)\rho_{\text{Kurt}}(\theta), \tag{15}$$

where $\alpha \in [0, 1]$.

We additionally exploit the fact that the ego-noise arrives rarely from the directions inside the noiseless sector, and propose a DOA weighting scheme to further improve the robustness to low SNRs. This is achieved by defining a new spatial likelihood function as

$$\rho_{\text{wHisK}}(\theta) = \beta(\theta)\rho_{\text{HisK}}(\theta), \tag{16}$$

where the weighting function $\beta(\theta)$ is:

$$\beta(\theta) = \begin{cases} 1, & \theta_L \le \theta \le \theta_H \\ \beta_T, & \text{otherwise} \end{cases} \tag{17}$$

where θ_L (θ_H) is the lower (upper) bound of the noiseless sector and $\beta_T < 1$ de-emphasizes the spatial likelihood values outside the noiseless sector.

We estimate the DOA of the target source as the location corresponding to the peak of the spatial likelihood function:

$$\hat{\theta}_d = \underset{\theta\in(-180^\circ, 180^\circ]}{\arg\max}\ \rho_{\text{wHisK}}(\theta). \tag{18}$$

Since the histogram values of the ego-noise in the noiseless sector are much lower than those in the noisy sector, this DOA weighting scheme can detect the target sound coming from the noiseless sector. When the target sound comes from the noisy sector, the scheme will not change the localization result, i.e. no improvement in the localization performance. In this case, the target sound can still be correctly detected if the SNR is sufficiently high (e.g. > 0 dB).

Similarly to (11), the target sound from direction $\hat{\theta}_d$ is extracted as

$$y_{\text{TF}}(k, l, \hat{\theta}_d) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(k, l, \hat{\theta}_d)\boldsymbol{x}(k, l). \tag{19}$$

VI. Experiments

A. Experimental setup

We compare the histogram measure `Hist` in (7), the kurtosis measure `Kurt` in (14), the combined measure `HisK` in (15), and the DOA-weighted measure `wHisK` in (16). For each measure, we estimate the DOA and then implement a spatial filter (18) pointing at the estimated direction. As a reference, we additionally implement a spatial filter `Target` pointing at the target direction, which is assumed to be known.

For all the algorithms, we set the STFT frame length as 1024 with half overlap. The working frequency is between 300 Hz and 3700 Hz at the sampling rate 8000 Hz. We set $\sigma = 10^\circ$ in (8), $\alpha = 0.33$ in (15), and $\beta_T = 0.2$ in (17).

Based on the observations in Fig. 4, we set the noiseless sector to be $\theta_L = -30^\circ$ and $\theta_H = 30^\circ$ for Array-C, and set $\theta_L = -60^\circ$ and $\theta_H = 60^\circ$ for Array-F. The search space $\{\theta_1, \cdots, \theta_D\}$ is set as $[-180^\circ, 180^\circ]$ with a step of 2°. This search space is also used as the histogram bins for `Hist`.

B. Data

The recording is made in a room of size 6×5×3m with reverberation of around 200 ms. The prototype used for the recording [9] is fixed on a tripod at a height of 1.8 m. The array consists of eight omnidirectional lavelier microphones. We consider the two specific array placements as described in Fig. 1. A loudspeaker is placed 3 m away from the MAV and at a height of 1.3 m, playing speech signals as the target sound. The ego-noise and the target sound are recorded separately and then added together at a varying input SNR from -30 dB to 5 dB, with a step of 5 dB. The locations of the MAV and the loudspeaker are fixed during the recording. The speed of the motors is varied during the recording of the ego-noise. The signals from the array are sampled simultaneously with a Zoom R24 multi-channel audio recorder, at a sampling rate of 44.1 kHz (downsampled to 8 kHz before processing).

For each array placement we produce two datasets. Dataset-1 is produced with recorded ego-noise and simulated speech. The speech is simulated with the image-source method [35] in a space of size 20×20×4m, with reverberation time 200 ms. The speech source is placed 10 m away, emitting sound at a varying DOA from -180° to 180°, with an interval of 10°. In such a distance the sound arrives at the microphones similarly to a plane wave. Dataset-2 is produced under a realistic scenario with the ego-noise and the speech recorded separately. The speech is recorded at two DOAs 110° and -20°, respectively.

C. Evaluation measures

We quantify the source localization and sound enhancement performance when the target sound arrives with a varying DOA $\theta_d \in [-180^\circ, 180^\circ]$ and a varying input SNR $\in [-30, 5]$ dB. For each combination of input SNR and DOA, we implement $I = 40$ realizations (segments), each lasting 6 seconds.

The *source localization performance* is evaluated with a correct ratio R_c. For a testing segment with a target direction θ_d and an estimation $\hat{\theta}_d$, the localization is regarded as correct if the estimation error is sufficiently small, e.g. $|\theta_d - \hat{\theta}_d| < 5^\circ$. Suppose among the I testing segments there are I_c segments with correct estimation results, the correct ratio is defined as

$$R_c = \frac{I_c}{I}. \tag{20}$$

The *sound enhancement performance* is evaluated with the signal-to-noise ratio (SNR) and signal-to-distortion ratio (SDR) measures, assuming the target $\boldsymbol{s}(n)$ and the noise component $\boldsymbol{v}(n)$ at the microphones to be known [36]. Given a spatial filter $\boldsymbol{w}(n)$, which is a time-domain version of $\boldsymbol{w}(k, l)$, the spatial filtering procedure is written as

$$\begin{aligned} y(n) &= \boldsymbol{w}(n) * \boldsymbol{x}(n) = \sum_{p=0}^{L_w-1} \boldsymbol{w}(p)\boldsymbol{x}(n-p) \\ &= y_s(n) + y_v(n) = \boldsymbol{w}(n) * \boldsymbol{s}(n) + \boldsymbol{w}(n) * \boldsymbol{v}(n), \end{aligned} \tag{21}$$

where '$*$' denotes the convolutive filtering procedure and L_w is the length of the filter $w(n)$; $y_s(n)$ and $y_v(n)$ are, respectively,

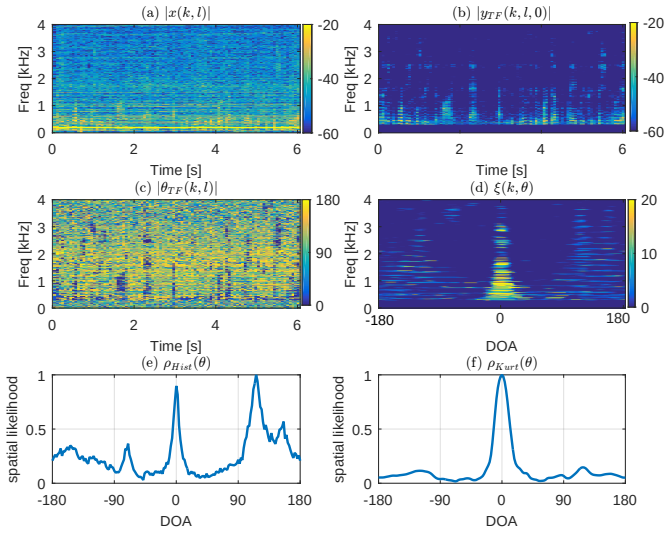


Fig. 6. Intermediate processing results by `Hisk` and `Kurt` for a target sound with DOA 0° and input SNR -10 dB. The ego-noise is recorded with Array-F. (a)-(b): Time-frequency spectra of the input and enhanced signals. The output SNR is 11.9 dB. (c) Local DOA estimation results at individual time-frequency bins. (d) Kurtosis map of the spatial filtering outputs for each frequency and DOA. (e) Spatial likelihood function by `Hist`. (f) Spatial likelihood function by `Kurt`.

the target and noise components at the output. The SNR is calculated in target-sound-active periods $\mathbb{N}_s$ [36]

$$\text{SNR} = 10\log_{10}\frac{\sum_{n'\in\mathbb{N}_s} y_s^2(n')}{\sum_{n'\in\mathbb{N}_s} y_v^2(n')}. \tag{22}$$

The SNR improvement between the input and output signals is calculated as

$$\text{SNR}_{\text{imp}} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \tag{23}$$

The SDR is defined given a reference signal $s_r(n)$ and a processed target signal $y_s(n)$:

$$\text{SDR} = 10\log_{10}\frac{\sum_{n'\in\mathbb{N}_s} s_r^2(n')}{\sum_{n'\in\mathbb{N}_s}\left(s_r(n') - y_s(n')\right)^2}. \tag{24}$$

We use the clean target sound at the first microphone as the reference signal. For each SNR_{in} and DOA, we calculate the averaged SNR_{imp} and SDR across the I testing segments.

D. Source localization results

Fig. 6 depicts the intermediate processing results by `Hist` and `Kurt` for a target sound arriving with $\theta_d = 0°$ and $\text{SNR}_{\text{in}} = -10$ dB. The ego-noise is recorded with Array-F while the speech signal is generated by simulation. Fig. 6(a) depicts the time-frequency spectrum of the input signal at one microphone, where the target sound is severely masked by the ego-noise. However, as shown in Fig. 6(c), performing local DOA estimation can still detect the time-frequency bins that belong to the target sound (i.e. at DOA 0°). Fig. 6(e) depicts the normalized spatial likelihood ρ_{Hist} based on the histogram of local DOA estimates. While the target sound presents a peak at 0°, the ego-noise also presents a peak at 110° with a higher spatial likelihood value, thus leading to an erroneous DOA estimation. Fig. 6(d) depicts the kurtosis map $\xi(k,\theta)$, where a high kurtosis value can be observed at DOAs around 0°. Fig. 6(f) depicts the normalized spatial likelihood ρ_{Kurt} by averaging the kurtosis values across the whole frequency band. A single peak can be clearly observed at around 0°.

We compare the spatial likelihood functions obtained by the four measures (`Hist`, `Kurt`, `HisK` and `wHisK`) for a target sound in the presence of two types of ego-noise: stationary and nonstationary, which are obtained when the motors are operating at a constant and a time-varying speed, respectively. Both types of ego-noise are recorded with Array-F. Fig. 7 illustrates the time-frequency spectra of the two types of noise. The simulated target sound arrives with different DOAs (0° and −150°) and input SNRs (-10 dB and -15 dB). Note that the two DOAs 0° and −150° belong to the noiseless and noisy sectors, respectively.

Fig. 8 shows the evaluation results for the stationary ego-noise in the upper block. For $\theta_d = 0°$ and $\text{SNR}_{\text{in}} = -15$ dB in Fig. 8(a), `Hist` gives a wrong estimate at 110° due to the low SNR. The other three measures all give a correct estimate at 0°. For $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -15$ dB in Fig. 8(b), `Hist` gives a wrong estimate at 110° due to the low SNR. `Kurt` and `HisK` both give a correct estimate at −150°. Although `wHisK` de-emphasizes the spatial likelihood value in the noisy sector, it still gives a correct estimate at −150°. For $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -10$ dB in Fig. 8(c), `Hist` gives a correct estimate at −150° due to the rise of the SNR. The other three measures also give a correct estimate at −150°.

The lower block of Fig. 8 shows the evaluation results for the nonstationary ego-noise. For $\theta_d = 0°$ and $\text{SNR}_{\text{in}} = -15$ dB in Fig. 8(d), `Hist` presents multiple peaks with the highest one at 0°. Due to the nonstationarity of the ego-noise, `Kurt` presents multiple peaks but with the highest one at 110°. When combining these two, `HisK` gives a wrong estimate at 110°. By de-emphasizing the value at 110°, `wHisK` gives a correct estimate at 0°. For $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -15$ dB in Fig. 8(e), `Hist` gives a correct estimate at −150° while `Kurt` gives a wrong estimate at 110°. When combining these two, `HisK` gives a wrong estimate at 110°. `wHisK` equally de-emphasizes the peaks at 110° and −150°, and thus does not change the estimation result, i.e. giving a wrong estimate at 110°. For $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -10$ dB in Fig. 8(f), `Hist` gives a correct estimate at −150°. `Kurt` presents multiple peaks at −150° and 110°, but with the highest one at −150°. Combining the two, `HisK` also gives a correct estimate. `wHisK` de-emphasizes the peaks at −150° and −110° equally, and thus does not change the estimation result, i.e. giving a correct estimate at −150°.

Fig. 9 comprehensively compares the localization performance by the four measures for Array-C and Array-F. The simulated target sound arrives with a varying DOA $\theta_d \in [-180°, 180°]$ and a varying $\text{SNR}_{\text{in}} \in \{-20, -15, -10\}$ dB. The localization correct ratio is computed using 40 testing segments, containing both stationary and nonstationary ego-noise.

Fig. 9(a) presents the evaluation results for Array-C, whose noiseless sector is $[-30°, 30°]$. `Hist` performs the worst in low SNRs with $\text{SNR}_{\text{in}} \leq -15$ dB, but might outperform `Kurt`

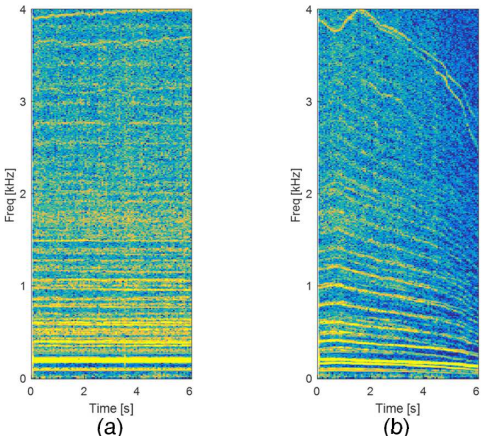


Fig. 7. Time-frequency spectra of (a) the stationary ego-noise generated when the motors are operating with a constant speed and (b) the non-stationary ego-noise generated when the motors are operating with a time-varying speed. The recording is made with Array-F.

when $\text{SNR}_{\text{in}} = -10$ dB. `HisK` performs worse than `Kurt` when $\text{SNR}_{\text{in}} = -20$ dB, similarly when $\text{SNR}_{\text{in}} = -15$ dB, and better when $\text{SNR}_{\text{in}} = -10$ dB. When the target sound comes from the noiseless sector, `wHisK` performs obviously the best with a correct ratio close to 1 for all input SNRs. When the target sound comes from the noisy sector, `wHisK` slightly outperforms `HisK` for all input SNRs. Meanwhile, `wHisK` performs worse than `Kurt` when $\text{SNR}_{\text{in}} = -20$ dB, similarly when $\text{SNR}_{\text{in}} = -15$ dB, and better when $\text{SNR}_{\text{in}} = -10$ dB.

Similar observations can be made for Array-F in Fig. 9(b). However, Array-F has a wider noiseless sector and thus a wider area, i.e. $[-60°, 60°]$, with better localization performance. When $\text{SNR}_{\text{in}} = -20$ dB, Array-F even performs slightly better than Array-C in the noiseless sector. When $\text{SNR}_{\text{in}} = -15$ dB, Array-F performs similarly as Array-C in both noiseless and noisy sectors. When $\text{SNR}_{\text{in}} = -10$ dB, Array-F performs similarly as Array-C in the noiseless sector but performs better in the noisy sector.

For all input SNRs, a performance rise is clearly observed for all algorithms at around 90° (in Fig. 9(a)) or 110° (in Fig. 9(b)). This is one of the directions that the ego-noise mainly comes from and detected as the target sound.

As summary of the source localization experiment, `Hist` shows degraded performance in low SNRs while `Kurt` shows degraded performance in the presence of nonstationary noise. `HisK` provides a trade-off between the two, while `wHisK` can substantially improve the robustness to low SNRs when the target sound arrives from the noiseless sector. Array-F has a wider noiseless sector than Array-C and thus overall better localization performance.

Finally, we investigate the variation of the localization performance with respect to the two parameters α and β_T, which are used in (15) and (17), respectively. We vary $\alpha \in [0, 1]$ with a 0.05 step and vary $\beta_T \in [0, 1]$ with a 0.1 step. For each pair of α and β_T we compute the average correct ratio for the 36 DOAs of the target sound around the circle. Fig. 10 depicts the variation of the average correct ratio with the two parameters at various input SNRs (-20 dB and -15 dB) and for

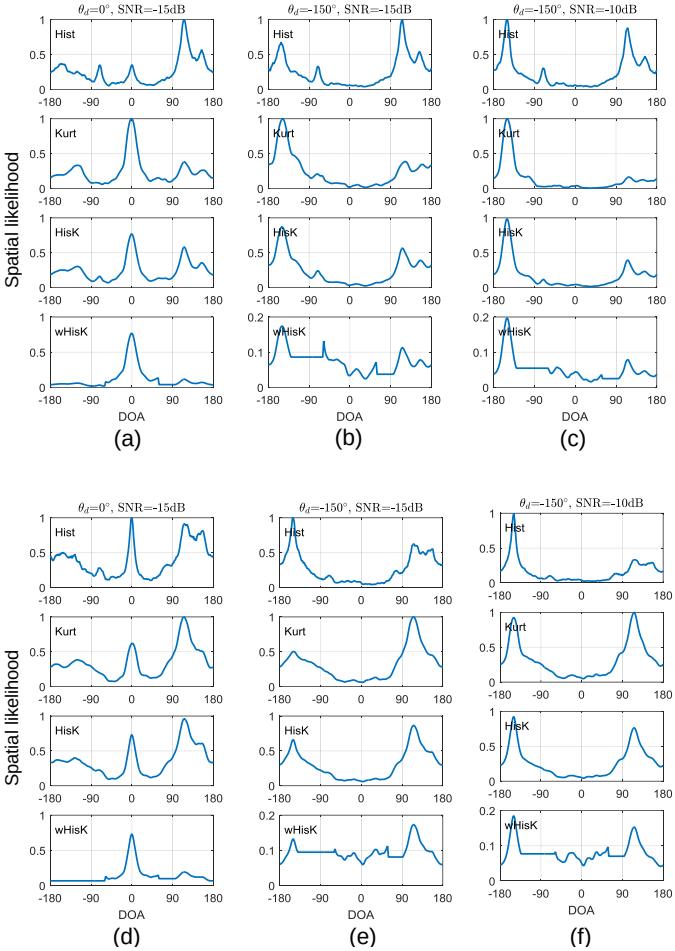


Fig. 8. Spatial likelihood functions obtained by `Hist`, `Kurt`, `HisK` and `wHisK` for different SNR_{in} and θ_d. The upper block corresponds to the stationary ego-noise. The lower block corresponds to the non-stationary ego-noise. (a)(d) $\theta_d = 0°$ and $\text{SNR}_{\text{in}} = -15$ dB. (b)(e) $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -15$ dB. (c)(f) $\theta_d = -150°$ and $\text{SNR}_{\text{in}} = -10$ dB.

the two array placements (Array-C and Array-F). As shown in each panel of Fig. 10 (with various array placement and input SNR), the obtained average correct ratio tends to have a higher value in the area $\alpha \in [0.1, 0.4]$ and $\beta_T \in [0.1, 0.4]$, which is consistent with our default choice in the experiment: $\alpha = 0.33$ and $\beta_T = 0.2$.

E. Sound enhancement results

We estimate the DOA of the target sound with the considered source localization algorithms and then implement a time-frequency spatial filter pointing at the estimated direction. Fig. 11 compares the noise reduction performance for the two array placements by polar-plotting the SNR improvement with respect to a varying target DOA $\theta_d \in [-180°, 180°]$ at different $\text{SNR}_{\text{in}} \in \{-20, -15, -10\}$ dB (Dataset-1).

In Fig. 11(a), `Target` constructs the spatial filter by assuming the target direction to be known and thus provides a benchmark for all noise reduction algorithms. For both array placements, the spatial filter responds non-uniformly to a varying target direction θ_d. Array-C obtains a higher SNR

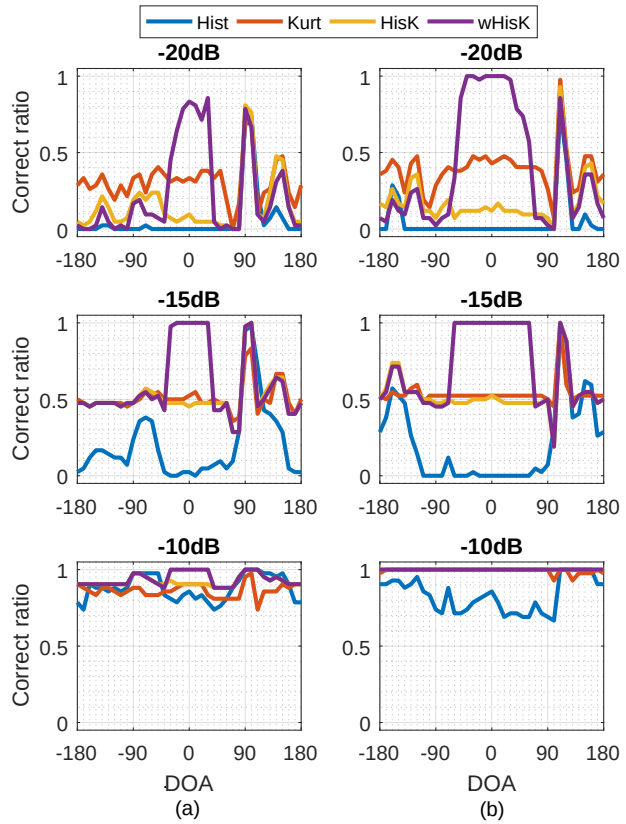


Fig. 9. Localization performance in terms of correct ratio for (a) Array-C and (b) Array-F. The target sound arrives with a varying DOA $\theta_d \in [-180°, 180°]$ and a varying input SNR $\in \{-20, -15, -10\}$ dB.

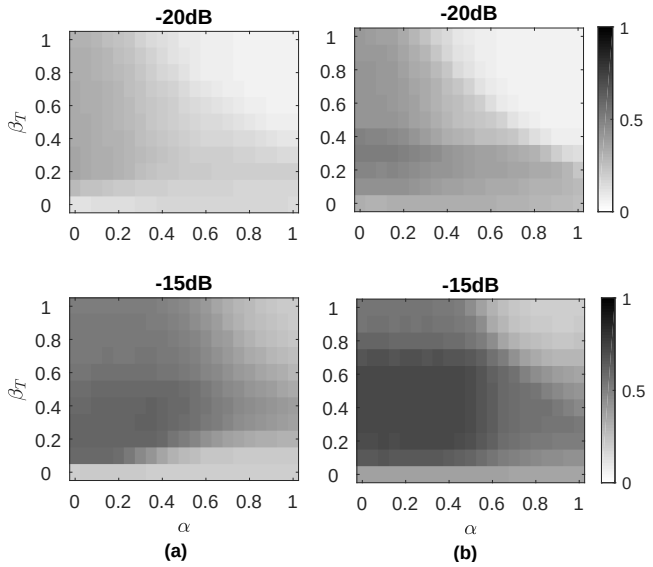


Fig. 10. Average correct ratio versus the parameters α and β_T at various input SNRs -20 dB and -15 dB for (a) Array-C and (b) Array-F.

improvement for directions inside the noiseless sector, i.e. $\theta \in [-30°, 30°]$, than directions outside. Similarly, Array-F obtains a higher SNR improvement for $\theta \in [-60°, 60°]$. Array-F obtains an even higher SNR improvement than Array-C in the noiseless sector. This is because, as observed in Fig. 4, Array-F has a slightly lower ego-noise histogram value in the noiseless sector than Array-C.

In Fig. 11(b), `Hist` almost shows no SNR improvement when $\text{SNR}_{\text{in}} \leq -15$ dB, because it cannot estimate the target direction correctly. Interestingly, an obvious SNR improvement is observed at $90°$ for Array-C and at $110°$ for Array-F. This is one of the directions that the ego-noise comes from. When the target sound comes from this direction, its DOA can be correctly detected (see Fig. 9). However, the spatial filter would extract both the target sound and the ego-noise, achieving a quite limited SNR improvement. When $\text{SNR}_{\text{in}} = -10$ dB, `Hist` achieves a higher localization accuracy for all directions and thus a higher SNR improvement. However, its performance is still worse than `Target`. Array-C and Array-F perform similarly in this case.

In Fig. 11(c), `Kurt` achieves higher SNR improvement than `Hist` when $\text{SNR}_{\text{in}} \leq -15$ dB, because `Kurt` can better localize the target sound. However, due to localization errors, `Kurt` still performs worse than `Target`. When $\text{SNR}_{\text{in}} = -10$ dB, the performance of `Kurt` improves significantly as the localization accuracy rises. Array-F achieves a localization correct ratio close to 1 (Fig. 9(b)) and thus performs similarly to `Target`. Array-C achieves a localization correct ratio lower than 1 (Fig. 9(a)) and thus performs slightly worse than `Target`.

In Fig. 11(d), `HisK`, as a combination of `Hist` and `Kurt`, performs similarly to `Hist` when $\text{SNR}_{\text{in}} = -20$ dB and performs similarly to `Kurt` when $\text{SNR}_{\text{in}} \geq -15$ dB.

In Fig. 11(e), the polar curve of the SNR improvement by `wHisK` looks very interesting especially when $\text{SNR}_{\text{in}} \leq -15$ dB. The performance of `wHisK` varies considerably between noisy and noiseless sectors. In the noiseless sector, `wHisK` achieves a localization correct ratio close to 1 (Fig. 9) and thus performs similarly to `Target` for all input SNRs. In the noisy sector, `wHisK` performs similarly as `HisK`: it cannot localize the target sound correctly and only improves the SNR limitedly. Array-F has a wider noiseless sector and thus overall better sound enhancement performance. Especially, when $\text{SNR}_{\text{in}} = -20$ dB, Array-F achieves a higher localization correct ratio than Array-C in the noiseless sector (Fig. 9), and thus also achieves a higher SNR improvement than Array-C. When $\text{SNR}_{\text{in}} = -10$ dB, `wHisK` performs similarly to `Target` for Array-F and slightly worse for Array-C.

As summary of the simulated experiment, the observations made in Fig. 11 verify `wHisK` as a promising method for MAV sound processing in extremely low-SNR scenarios. When $\text{SNR}_{\text{in}} = -10$ dB, the four algorithms `Hist`, `Kurt`, `HisK` and `wHisK` perform similarly. When $\text{SNR}_{\text{in}} \leq -15$ dB, `wHisK` significantly outperforms the other three algorithms if the target sound arrives from the noiseless sector. With an appropriate array placement, the sound enhancement performance can be further optimized. For instance, Array-F outperforms Array-C with a wider noiseless sector and also a higher SNR improvement inside this sector.

Finally, we compare the performance of `wHisK` and `Target` with a real-recorded target sound coming from $110°$ and $-20°$, respectively (Dataset-2). Fig. 12 depicts the SNR

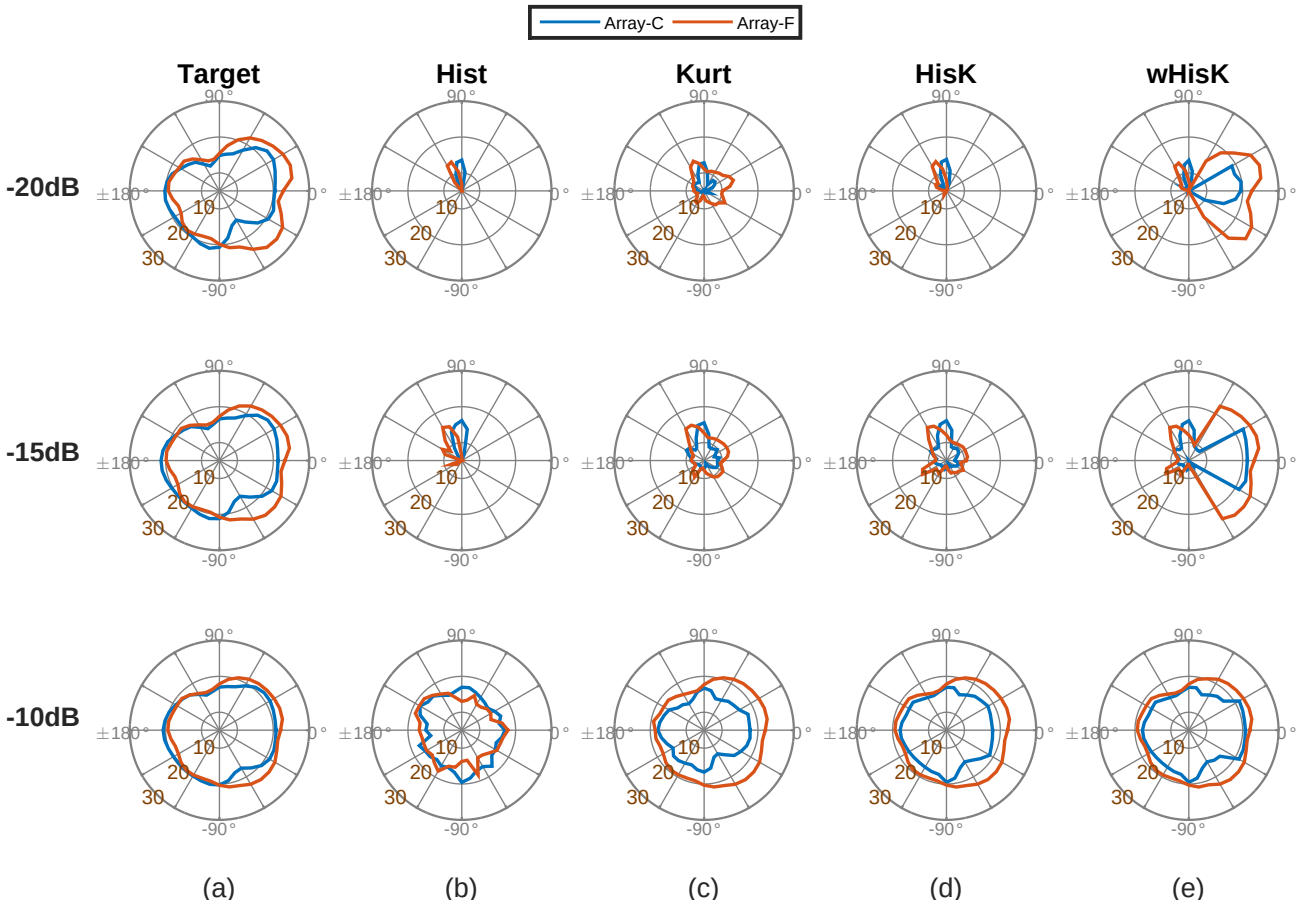


Fig. 11. Polar plots of SNR improvement with respect to a varying DOA of the target sound with input SNRs -20 dB, -15 dB and -10 dB. Five algorithms are considered. (a) `Target`. (b) `Hist`. (c) `Kurt`. (d) `HisK`. (e) `wHisK`. The radius of the polar plot denotes the SNR improvement in dB.

improvement and SDR obtained by the two algorithms when the input SNR varies from -30 dB to 5 dB.

For both array placements, `Target` always achieves higher SNR improvement than `wHisK`, especially when $\text{SNR}_{\text{in}} \leq -15$ dB. The difference becomes bigger as SNR_{in} decreases. When $\theta_d = 110°$, the target sound comes from the noisy sector for both array placements. `Target` performs similarly for Array-C and Array-F when $\text{SNR}_{\text{in}} \geq -10$ dB, but performs slightly better for Array-C when $\text{SNR}_{\text{in}} \leq -15$ dB. `wHisK` performs similarly for the two array placements when $\text{SNR}_{\text{in}} \geq -10$ dB, but performs slightly better for Array-F when $\text{SNR}_{\text{in}} \leq -15$ dB. When $\theta_d = -20°$, the target sound comes from the noiseless sector for both array placements. `Target` performs similarly for Array-C and Array-F when $\text{SNR}_{\text{in}} \geq -20$ dB, and performs slightly better for Array-F when $\text{SNR}_{\text{in}} \leq -25$ dB. `wHisK` performs similarly for the two array placements when $\text{SNR}_{\text{in}} \geq -15$ dB, and performs significantly better for Array-F when $\text{SNR}_{\text{in}} \leq -20$ dB. The SDR obtained by the two algorithms improves slowly with increasing input SNR. For $\theta_d = 110°$, `Target` slightly outperforms `wHisK`. The two algorithms both perform slightly better for Array-F. For $\theta_d = -20°$, `Target` and `wHisK` perform almost equally for the two array placements.

As summary of the real-recorded experiment, `wHisK` performs better for Array-F, especially when the target sound comes from the noiseless sector and when the input SNR is low, e.g. $\leq$ 15 dB. The time-frequency spatial filtering however distorts the target sound, as verified by the SDR between 0 dB and 5 dB.

F. Comparison with the state of the art

We compare the source localization and sound enhancement performance of the proposed method with state-of-the-art algorithms. For source localization, we consider the proposed algorithm `wHisK`, `SRP-PHAT` and `GEVD-MUSIC` [21]. These three algorithms compute a spatial likelihood function and estimate the source location as the one that maximizes the spatial likelihood function, similarly to (18). For sound enhancement, we consider the proposed algorithm `wHisK`, fixed beamforming (`FB`) [24], and `BSS` [12]. `wHisK` enhances the direction it estimates while `FB` assumes the direction of the target sound to be known.

We first compare the source localization and sound enhancement performance of the considered algorithms for a simulated target sound with a DOA varying from -180° to 180° with an interval of 10°, and at two input SNRs -20 dB and -10 dB (Dataset-1). The recording is made with Array-F. Fig. 13 depicts the correct ratio (source localization) and the SNR improvement (sound enhancement) obtained by the considered algorithms. For the source localization performance shown in

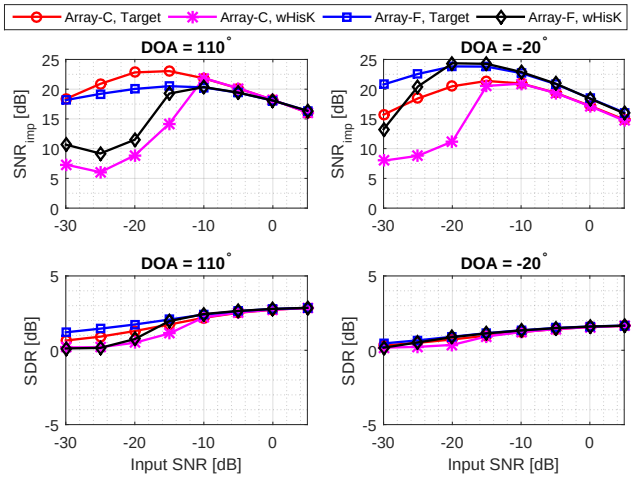


Fig. 12. Sound enhancement performance in terms of SNR improvement and SDR by `Target` and `wHisK`. The target sound is real-recorded with DOAs 110° and -20°. The input SNR varies from -30 dB to 5 dB.

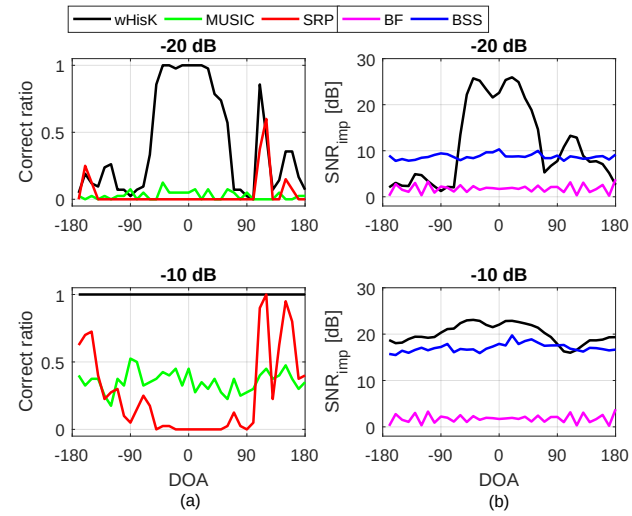


Fig. 13. Source localization and sound enhancement performance by the considered algorithms for a simulated target sound with a varying DOA $\theta_d \in [-180, 180]$ at various input SNRs (-20 dB and -10 dB). The recording is made with Array-F. (a) Source localization performance in terms of correct ratio. (b) Sound enhancement performance in terms of SNR improvement.

Fig. 13(a), `wHisK` obviously outperforms `MUSIC` and `SRP` for both input SNRs. When $\text{SNR}_{\text{in}} = -20$ dB, `wHisK` achieves high correct ratios inside the noiseless sector [-60°, 60°], and low correct ratios inside the noisy sector. `MUSIC` and `SRP` fail when $\text{SNR}_{\text{in}} = -20$ dB, with correct ratios close to 0 for most target DOAs. An exceptional peak is observed for `SRP` around 110°. This is because that an ego-noise is dominant at this direction and is detected as the target sound. When $\text{SNR}_{\text{in}} = -10$ dB, `wHisK` can detect the target sound with correct ratios close to 1 for all target DOAs. `MUSIC` and `SRP` achieve higher correct ratios as the input SNR increases, but still much lower than those by `wHisK`.

For the sound enhancement performance shown in Fig. 13(b), `wHisK` achieves a much higher SNR improvement in the noiseless sector [-60°, 60°] than in the noisy sector when $\text{SNR}_{\text{in}} = -20$ dB. The performance of BSS remains

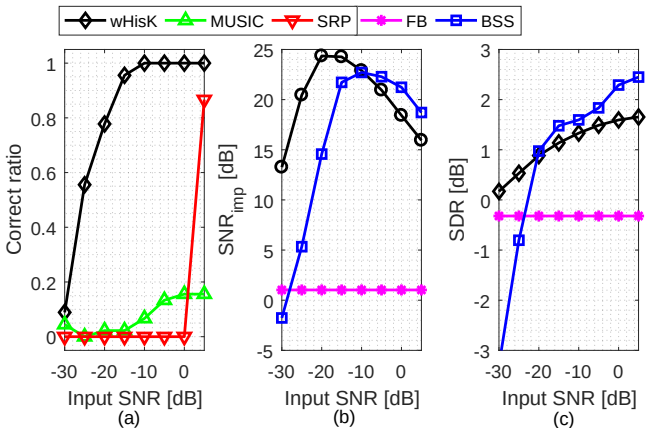


Fig. 14. Source localization and sound enhancement performance by the considered algorithms for a real-recorded target sound coming from -20°. The recording is made with Array-F. The input SNR varies from -30 dB to 5 dB. (a) Source localization performance in terms of correct ratio. (b)(c) Sound enhancement performance in terms of SNR improvement and SDR, respectively.

stable as the target DOA varies. `wHisK` clearly outperforms `BSS` in the noiseless sector, and but performs worse than `BSS` in the noisy sector. When $\text{SNR}_{\text{in}} = -10$ dB, `wHisK` achieves a slightly higher SNR improvement in the noiseless sector than in the noisy sector. `wHisK` performs slightly better than `BSS` for most target DOAs. For both input SNRs, `FB` performs significantly worse than `wHisK` and `BSS`, improving the SNR limitedly for most target DOAs.

We then compare the source localization and sound enhancement performance of the considered algorithms for a real-recorded target sound coming from −20° with a varying input SNR from -30 dB to 5 dB with an interval of 5 dB (Dataset-2). The recording is made with Array-F. Fig. 14 shows the experimental results in terms of source localization and sound enhancement. In Fig. 14(a) `wHisK` significantly outperforms `GEVD-MUSIC` and `SRP-PHAT` in all testing scenarios. `GEVD-MUSIC` outperforms `SRP-PHAT` when $-15\text{dB} \leq \text{SNR}_{\text{in}} \leq 0\text{dB}$, while `SRP-PHAT` performs better in high SNRs with $\text{SNR}_{\text{in}} \geq 5$ dB. The poor performance of `GEVD-MUSIC` is mainly due to the inaccurate estimate of the noise correlation matrix and the lack of calibration of the microphones [13]. In Fig. 14(b) `wHisK` significantly outperforms `BSS` in low-SNR scenarios with $\text{SNR}_{\text{in}} \leq -15$ dB, while `BSS` performs slightly better in high-SNR scenarios. The fixed beamformer only improves the SNR limitedly and the improvement remains constant with respect to the varying input SNR. In Fig. 14(c) `wHisK` achieves higher SDR than `BSS` in low-SNR scenarios with $\text{SNR}_{\text{in}} \leq -20$ dB, while `BSS` achieves slightly higher SDR in high-SNR scenarios.

VII. Conclusion

We proposed a time-frequency processing method to localize and enhance a target sound captured by an MAV by exploiting the spectral and spatial characteristics of the ego-noise. The spatial filter presents a high directivity towards a noiseless sector even in extremely low-SNR scenarios.

The proposed method significantly outperforms the competing source localization algorithms. The proposed method also outperforms the competing sound enhancement algorithms especially in low-SNR scenarios. We also showed how to further widen noiseless sectors and to achieve higher SNR improvement with different positionings of the array.

The benefits of the proposed method increase when the MAV turns its looking direction [37] so that the target sound comes from a noiseless sector. Moreover, multiple arrays could be used to further widen the noiseless sector. For instance, mounting one array at the front and one at the back side of the MAV would enable the perception of sounds coming from both sides of the MAV. In addition, since the direction of the motor noise remains unchanged with respect to the microphone array, how to exploit this information to further improve the source localization performance is an interesting future research topic. Future work includes extending the proposed algorithm to real 3D environments with natural wind and multiple sound sources.

REFERENCES

[1] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, pp. 460-466, May 2015.

[2] K. Daniel, S. Rohde, N. Goddemeier, and C. Wietfeld, "Cognitive agent mobility for aerial sensor networks," *IEEE Sensors J.*, vol. 11, no. 11 pp. 2671-2682, Jun. 2011.

[3] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 3288-3293.

[4] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 4737-4742.

[5] S. Yoon, S. Park, Y. Eom, and S. Yoo, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, USA, 2015, pp. 26-29.

[6] J. Cacace, R. Caccavale, A. Finzi, and V. Lippiello, "Attentional multimodal interface for multidrone search in the Alps" in *Proc. IEEE Int. Conf. Sys. Man, Cybernetics*, Budapest, Hungary, 2016, pp. 1178-1183.

[7] T. Latif, E. Whitmire, T. Novak, and A. Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3444-3453, May 2016.

[8] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H.G. Okuno, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, pp. 1-16, 2017.

[9] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles", in *Proc. Int. Conf. Adv. Video Signal-Based Surveillance*, Colorado Springs, USA, 2016, pp. 1-7.

[10] G. Sinibaldi and L. Marino, "Experimental analysis on the noise of propellers for small UAV," *Appl. Acoust.*, vol. 74, no. 1, pp. 79-88, Jan. 2015.

[11] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, 2015, pp. 5610-5614.

[12] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors J.*, vol. 17, no. 8, pp. 2447-2455, Apr. 2017.

[13] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, USA, 2017, pp. 1-5.

[14] 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 *Proc. Acoust.*, 2012, pp. 1-6.

[15] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Tokyo, Japan, 2013, pp. 3943-3948.

[16] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, 2011, pp. 3517-3522.

[17] G. Ince, K. Nakadai, and K. Nakamura, "Online learning for template-based multi-channel ego noise estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vilamoura-Algarve, Portugal, 2012, pp. 3282-3287.

[18] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, USA, 2016, pp. 219-222.

[19] R. P. Fernandes, E. C. Santos, A. L. L. Ramos, and J. A. Apolinario Jr., "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proc. Int. Conf. Transformative Sci. Eng. Business Social Innovation*, Fort Worth, USA, 2015, pp. 536-541.

[20] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound,", *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 820-827, 2016.

[21] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and L. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Chicago, USA, 2014, pp. 1902-1907.

[22] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079-1093, Jun. 2016.

[23] S. Argentieri, P. Danes, and P. Soueres, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech Lang.* vol. 34, no. 1, pp. 87-112, 2015.

[24] T. Ishiki and M. Kumon, "A microphone array configuration for an auditory quadrotor helicopter system," in *Proc. IEEE Int. Symp. Safety, Security, Rescue Robot.*, Toyako-cho, Japan, 2014, pp. 1-6.

[25] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Hamburg, Germany, 2015, pp. 6143-6148.

[26] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Xi'an, China, 2016, pp. 1-5.

[27] L. Wang, "Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation," *Digit. Signal Process.*, vol. 31, pp. 79-92, 2014.

[28] L. Wang, J. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1573-1588, Sep. 2016.

[29] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, pp. 2641-2673, Dec. 2003.

[30] L. Zeng and G. M. Bone, "Mobile robot collision avoidance in human environments," *Int. J. Adv. Robot. Sys.*, vol. 10, pp. 1-14, 2013.

[31] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: a flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31-42, Mar. 2015.

[32] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182-2196, Dec. 2014.

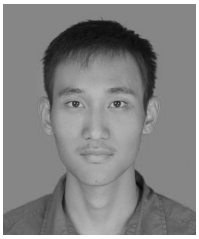
[33] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York, USA: John Wiley & Sons, 2004.

[34] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230-2244, Sep. 2002.

[35] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, 1979.

[36] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1493-1508, Sep. 2015.

[37] R. Sanchez-Matilla, L. Wang, and A. Cavallaro, "Multi-modal localization and enhancement of multiple sound sources from a micro aerial vehicle," *Proc. ACM Multimedia*, Silicon Valley, USA, 2017, pp. 1591-1599.



Lin Wang received the B.S. degree in electronic engineering from Tianjin University, China, in 2003; and the Ph.D degree in signal processing from Dalian University of Technology, China, in 2010. From 2011 to 2013, he was an Alexander von Humboldt Fellow at the University of Oldenburg, Germany. Since 2014, he has been a postdoctoral researcher in the Centre for Intelligent Sensing at Queen Mary University of London. His research interests include video and audio compression, microphone array, blind source separation, 3D audio processing, and machine learning (https://sites.google.com/site/linwangsig/).

Andrea Cavallaro received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2002. He was a Research Fellow with British Telecommunications in 2004. He is a Professor of Multimedia Signal Processing and the Director of the Centre for Intelligent Sensing at Queen Mary University of London. He has authored more than 200 journal and conference papers, one monograph on Video Tracking (Wiley, 2011), and three edited books, Multi-Camera Networks (Elsevier, 2009), Analysis, Retrieval and Delivery of Multimedia Content (Springer, 2012), and Intelligent Multimedia Surveillance (Springer, 2013). Prof. Cavallaro is Senior Area Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology and the IEEE MultiMedia Magazine. He is vice-chair of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee, and an elected member of the IEEE Circuits and Systems Society Visual Communications and Signal Processing Technical Committee. He is a former elected member of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee, Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON IMAGE PROCESSING, and Associate Editor and Area Editor of IEEE Signal Processing Magazine, and Guest Editor of eleven special issues of international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was Technical Program Chair of IEEE AVSS 2011 and 2018, EUSIPCO 2008, and WIAMIS 2010. He received the Royal Academy of Engineering Teaching Prize in 2007, three Student Paper Awards at IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009.