Multi-Modal Localization and Enhancement of Multiple Sound Sources from a Micro Aerial Vehicle

Ricardo Sanchez-Matilla*
Centre for Intelligent Sensing
Queen Mary University of London
London, United Kingdom
ricardo.sanchezmatilla@qmul.ac.uk

Lin Wang*
Centre for Intelligent Sensing
Queen Mary University of London
London, United Kingdom
lin.wang@qmul.ac.uk

Andrea Cavallaro
Centre for Intelligent Sensing
Queen Mary University of London
London, United Kingdom
a.cavallaro@qmul.ac.uk

ABSTRACT

The ego-noise generated by the motors and propellers of a micro aerial vehicle (MAV) masks the environmental sounds and considerably degrades the quality of the on-board sound recording. Sound enhancement approaches generally require knowledge of the direction of arrival of the target sound sources, which are difficult to estimate due to the low signal-to-noise-ratio (SNR) caused by the ego-noise and the interferences between multiple sources. To address this problem, we propose a multi-modal analysis approach that jointly exploits audio and video data to enhance the sounds of multiple targets captured from an MAV equipped with a microphone array and a video camera. We first perform audio-visual calibration via camera resectioning, audio-visual temporal alignment and geometrical alignment to jointly use the features in the audio and video streams, which are independently generated. The spatial information from the video is used to assist sound enhancement by tracking multiple potential sound sources with a particle filter. Then we infer the directions of arrival of the target sources from the video tracking results and extract the sound from the desired direction with a time-frequency spatial filter, which suppresses the ego-noise by exploiting its time-frequency sparsity. Experimental results with real outdoor data verify the robustness of the proposed multi-modal approach for multiple speakers in extremely low-SNR scenarios.

KEYWORDS

audio-visual sensing; ego-noise reduction; micro aerial vehicles; microphone array; multi-modal localization; enhancement of multiple sound sources; multiple object tracking

1 INTRODUCTION

Multi-rotor micro aerial vehicles (MAV) with audio sensing capabilities could localize, recognize and enhance the sound emitted from an aerial or ground object [1, 16, 20, 35]. However, a strong ego-noise is generated by rotating motors and propellers, which are closer to the on-board microphones than ground or aerial target sources [24]. The strong ego-noise leads to an extremely low signal-to-noise ratio (e.g. SNR < -15 dB), which masks the target sounds. By exploiting the spectral and spatial characteristics of the acoustic signals, microphone-array algorithms can suppress the ego-noise and enhance the target sounds [25]. However, to steer the spatial filter, these algorithms typically require the direction of arrival (DOA) of the target sound, which is difficult to estimate from the microphone signal due to the extremely low SNR, the non-stationarity of the ego-noise, and multiple active sound sources. While video-assisted sound enhancement has already been investigated [14, 22, 30], existing works address indoor environments and static audio-visual sensors. To the best of our knowledge, our work is the first to combine audio and visual modalities for the challenging problem of sound enhancement from an MAV.

In this paper, we integrate the audio and visual modalities to enhance target sounds captured by an array of microphones mounted on an MAV. We first synchronize the audio and video streams and geometrically align the spatial information estimated from the two streams. We then robustly estimate the position of potential sound-emitting objects (*e.g.* human speakers) from the video stream. Finally, we design a time-frequency spatial filter which, based on the location provided by the video, extracts the target sound from the audio streams captured by multiple microphones. By exploiting the complementarity of the two modalities, the proposed audio-visual sensing system works in extremely low SNR scenarios and can isolate, track and enhance the sounds from a time-varying number of speakers. A demonstration of the results is available online [1]

The paper is organized as follows. Sec. 2 reviews the related work. Sec. 3 formulates the problem. Sec. 4 describes the audio-visual calibration procedure. Sec. 5 presents the proposed video-assisted sound enhancement method. Experimental results are discussed in Sec. 6 and conclusions are drawn in Sec. 7.

2 RELATED WORK

The ego-noise of a flying MAV leads to extremely low SNRs, non-stationarity and varying dynamics. These are considerable challenges for noise reduction and sound source localization algorithms.

Beamforming is a widely-used microphone-array technique, which enhances the sound from a specific direction by coherently delaying and summing the microphone signals based on the transmitting delays from the sound source to the microphones [9]. The performance of a fixed beamformer is usually limited by the size of the microphone array and the number of the microphones. Blind

*The first two authors contributed equally to this work.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/https://doi.org/10.1145/3123266.3123412

[1] http://cis.eecs.qmul.ac.uk/projects/multimodalmav/

source separation (BSS) has recently been used for ego-noise reduction [24]. BSS treats the target and noise signals equally and separates the individual sources from the mixed signals captured by the microphone array [23]. The application of BSS to MAV-based ego-noise reduction is straightforward as the locations of the microphones and the target sources are not needed. However, BSS suffers from the inherent permutation ambiguities, which are difficult to address in low-SNR scenarios [29].

Time-frequency spatial filtering has emerged recently for MAV sound enhancement [25]. Based on the observation that the ego-noise and the target sound usually have concentrated energy at sparsely isolated time-frequency bins, the time-frequency approach estimates the DOA of the sound at each bin and then combines the localization results from all the bins for noise reduction. While the time-frequency approach can suppress the ego-noise effectively, similarly to beamforming, the design of the spatial filter also requires the DOA of the sound.

Classical microphone-array sound source localization approaches include steered response power (SRP) and multiple signal classification (MUSIC) [28, 29]. The performance of both approaches degrades significantly in low-SNR scenarios [15]. Recently, it was proposed that combining time-frequency spatial filtering with a kurtosis measure would lead to noise-robust sound source localization [26]. However, this approach assumes a single target and thus cannot handle a multi-source scenario.

Video-based object detection and tracking can provide spatial information about the objects in the field of view of the camera. Features used to represent object models include intensities [3, 5, 6], edges [3, 6] and textures [31]. The performance of these models can be compromised in challenging scenarios with low contrast or crowds. Color attributes can be used as an explicit color representation [10] and inter-object occlusions can be used as clues to improve the detection of partly occluded objects [7]. Recently, deep learning techniques have been proposed where image regions with objects of interest produce a high response of a pool of filters [18, 32, 33].

Multi-object trackers can estimate the trajectory of the targets by temporally associating sets of noisy detections generated at each frame. This association compensates for false-positive and false-negative detections using spatio-temporal relationships [2, 17]. The probability hypothesis density (PHD) filter [11, 12] estimates the state of multiple targets by building a positive and integrable function over a multi-dimensional state, usually known as *posterior*. This probabilistic filter can cope with clutter, spatial noise and missing detections while effectively filtering the state estimation using current and past information only. The posterior can be estimated using Bayesian recursion. As this iterative process is computationally intractable, the first order posterior function can be approximated using a sequential Monte Carlo method with weighted samples. This approximation is known as probability hypothesis density particle filter (PHD-PF) [21] and the weighted samples are known as particles.

3 PROBLEM FORMULATION

Fig. 1 depicts the audio-visual sensing platform consisting of a circular microphone array and a camera mounted on the MAV. The microphone array is placed on top of the MAV in order to avoid

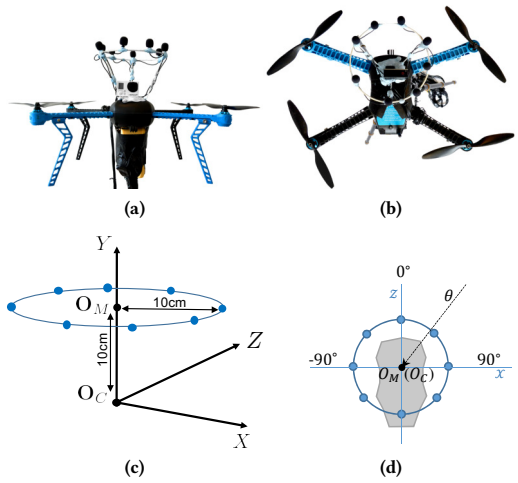

Figure 1: The audio-visual sensing platform consisting of a microphone array and a camera mounted on the MAV. (a) Side and (b) top view of the real object. (c) 3D and (d) 2D geometrical representation.

the influence of the wind from the propellers. The microphone array contains $M = 8$ microphones whose signals are sampled synchronously with a multichannel analogue-to-digital converter. The center of the camera overlaps that of the microphone array to ease audio-visual calibration. The audio and video acquisition devices work independently of each other.

Fig. 2 illustrates the coordinate systems of the microphone array and the camera. We use the pinhole model for the camera [36]. A real-world object P is projected onto the image plane $\boldsymbol{p}$, with $\boldsymbol{O}_C$, $\boldsymbol{O}_I$ and F being the center of the camera, the principal point (center) in the image and the focal length, respectively. We only consider the DOA of the sound on the 2D horizontal plane. The horizontal angles of the object with respect to the microphone array and the camera are indicated as θ_a and θ_v, respectively.

We consider an unknown number of speakers, N, who might talk or remain silent in front of the camera. The locations of the microphones are known to be $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M]$, where $\boldsymbol{r}_m = [r_{mu}, r_{mv}]^{\mathrm{T}}$ is the location of the m-th microphone. The superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operator. The video $\mathbb{I} = \{\boldsymbol{I}_k\}_{k=1}^{K}$, where $\boldsymbol{I}_k$ is the k-th frame and K is the total number of video frames, has frame rate f_c. The microphone signal $\boldsymbol{x}(n) = [x_1(n), \ldots, x_M(n)]^{\mathrm{T}}$ contains the sound from the N speakers and the ego-noise, *i.e.*

$$\boldsymbol{x}(n) = \sum_{j=1}^{N} \boldsymbol{s}_j(n) + \boldsymbol{v}(n), \tag{1}$$

where $\boldsymbol{s}_j(n) = [s_{1j}(n), \ldots, s_{Mj}(n)]^{\mathrm{T}}$ denotes the sound from the j-th speaker, $\boldsymbol{v}(n) = [v_1(n), \ldots, v_M(n)]^{\mathrm{T}}$ denotes the ego-noise and n is the digital audio sequence index.

We aim to design a set of spatial filters $\{\boldsymbol{w}_1(n), \cdots, \boldsymbol{w}_N\}$ that can extract the N target sounds from the noisy microphone recordings,

$$y_j(n) = \boldsymbol{w}_j(n) * \boldsymbol{x}(n) = \sum_{i=1}^{M} \sum_{p=1}^{L_P} w_{ji}(p) x_i(n-p), \quad j = 1, \cdots, N \tag{2}$$

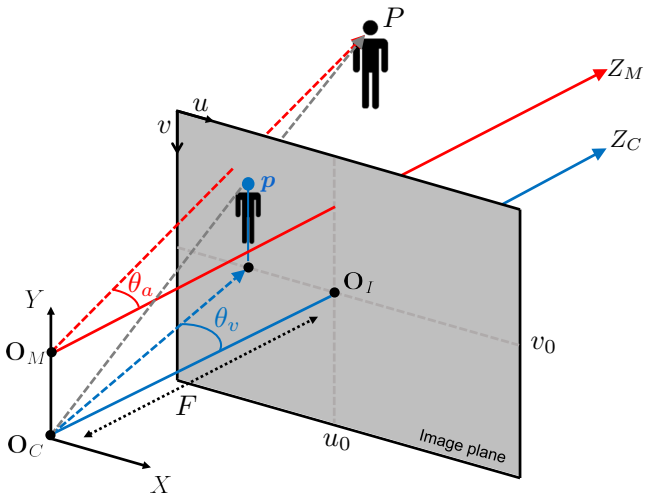

Figure 2: Schematic illustration of the coordinate systems of the microphone array and the camera (pinhole model). The 3D position P of a real-world object is projected on the image plane as $\boldsymbol{p}$. θ_a and θ_v are the angles, on the 2D horizontal planes, of the object with respect to the microphone array and the camera. O_M and O_C are the centers of the microphone array and camera, respectively, $O_I = (u_0, v_0)$ is the principle point (center) of the image and F is the focal length of the camera.

where $\boldsymbol{w}_j = [w_{j1}(n), \cdots, w_{jM}(n)]$ denotes the spatial filter corresponding to the j-th target, L_P is the length of the filter, and the operator '$*$' denotes the spatial filtering procedure [23].

The proposed work can be decomposed into three elements, namely, audio-visual calibration (Sec. 4), visual object detection and tracking (Sec. 5.1), and spatially informed audio enhancement (Sec. 5.2). The first step calibrates the locations of the camera and microphones and aligns the audio and video streams so that they can be correctly associated. The second step works on the video stream by estimating the location of potential sound emitting objects. The third step works on the audio stream by designing a time-frequency spatial filter to enhance the sound from the video-informed directions. The block diagram of the proposed multi-modal source localization and sound enhancement pipeline is shown in Fig. 3.

4 AUDIO-VISUAL CALIBRATION

Calibration of the microphone array and the camera is needed so that the features from the audio and video streams can be jointly exploited. The calibration procedure consists of camera resectioning, audio-visual temporal alignment and geometrical alignment.

4.1 Camera resectioning and audio-visual temporal alignment

To compensate for the deformation produced by the lens and to infer the real-world location of objects from the image, we use *camera resectioning* to estimate the intrinsic and distortion parameters [8, 36]. We first record a calibration video of a checkerboard at

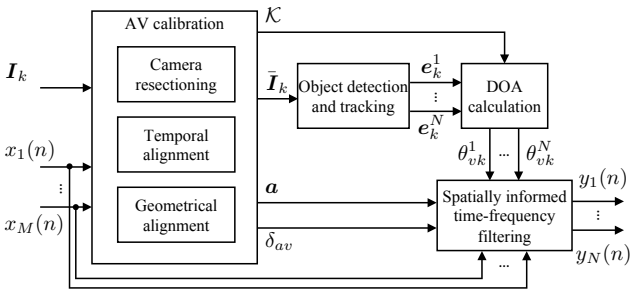

Figure 3: Block diagram of the proposed multi-modal source localization and sound enhancement pipeline, which consists of three main steps: audio-visual (AV) calibration, visual object detection and tracking, and spatially informed audio enhancement.

different locations and then estimate the camera parameters with the MATLAB Camera Calibration Toolbox [13]. The radial and tangential lens distortion parameters are represented by $\boldsymbol{\xi}$ and the intrinsic matrix is defined as

$$\mathcal{K} = \begin{bmatrix} F_u & c_s & u_0 \\ 0 & F_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $F = \frac{F_u + F_v}{2}$ is the camera focal length measured in pixels (see Fig. 2), u_0 and v_0 indicate the location of the principle point (optical center) in the image, and c_s is the skew axis coefficient. The parameter $\mathcal{K}$ will be used in Sec. 4.2 for audio-visual geometrical alignment.

The parameter $\boldsymbol{\xi}$ is used to undistort the frames as

$$\bar{\boldsymbol{I}}_k = \mathcal{D}(\boldsymbol{I}_k, \boldsymbol{\xi}), \tag{4}$$

where $\mathcal{D}(\cdot)$ represents the undistortion procedure [8].

We then estimate the unknown time offset between the microphone array and the video camera, δ_{av}, to *temporally align* the audio and video streams. As our camera has its own built-in microphone, we only need to detect the time offset between the audio sequences from the array microphone and the camera microphone. We present a calibration sound (*e.g.* clapping) to estimate the offset between the two audio sequences.

If we represent the two segments of sequences as $s_a(n)$ and $s_v(n)$, where $n \in \mathcal{N}_c$, both containing the calibration sound, then the time offset δ_{av} is determined by maximizing the correlation between the two segments:

$$\delta_{av} = \underset{\delta \in [\delta_L, \delta_H]}{\arg\max} \sum_{n \in \mathcal{N}c} s_a(n) s_v(n - \delta), \tag{5}$$

where δ_L and δ_H denote predefined minimum and maximum delays, respectively. The parameter δ_{av} will be used in Sec. 5.2 when temporally associating the spatial information from audio and video streams.

4.2 Audio-visual geometrical alignment

When an object emits a sound, the angle on the 2D microphone array plane (*i.e.* its DOA) can be estimated from the microphone-array signals or from the visual signal, *e.g.* θ_a and θ_v in Fig. 2. Since the

microphone array and the video camera have their own coordinate systems, it is important to know the relationship between the θ_a and θ_v to infer the DOA of the sound from the corresponding object in the image.

If the camera and the microphone array are placed with their centers and coordinates aligned (see Fig. 2), θ_a should be equal to θ_v. However, it is difficult to satisfy this condition by mounting the two devices on the MAV manually. We address this displacement error numerically, assuming that these two DOAs are linearly related as

$$\theta_a = a_1\theta_v + a_2, \tag{6}$$

where $\boldsymbol{a} = [a_1, a_2]^{\mathrm{T}}$ are unknown constants.

To estimate a_1 and a_2, we record the sound from a speaker at Q different locations with both the microphone array and the camera while the MAV is muted. Let us use the sound from the q-th location as an example. For the audio, the DOA of the sound, θ_a^q, can be estimated from the microphone signal with the classical SRP-PHAT algorithm [26]. For the video, we manually label the sound emitting point (speaker's mouth) in the image, *e.g.* $\boldsymbol{p}_q = (u_q, v_q)$, and then estimate the DOA as

$$\theta_v^q = \arctan\left(\frac{u_q}{F}\right). \tag{7}$$

We estimate the DOAs of the speaker from the audio as $\boldsymbol{\theta}_a = [\theta_a^1, \ldots, \theta_a^Q]^{\mathrm{T}}$ and from the video as $\boldsymbol{\theta}_v = [\theta_v^1, \ldots, \theta_v^Q]^{\mathrm{T}}$. The parameter $\boldsymbol{a}$ can be estimated from $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_v$ using least-square fitting. This parameter will be used in Sec. 5.2 when a sound event in the audio and video streams is geometrically associated.

5 VIDEO-ASSISTED SOUND ENHANCEMENT

5.1 Visual object detection and tracking

As the video information is not affected by the strong ego-noise, we propose to exploit this modality to obtain the spatial information of the objects which potentially emit sound, *e.g.* a person. We first detect people in each frame and then track their location over time with a multiple-object tracker.

For person detection, we employ the Aggregate Channel Features (ACF) algorithm [5], a supervised object detector which can robustly detect quasi-rigid objects from images, *e.g.* faces, pedestrians or cars. In each undistorted video frame $\tilde{\boldsymbol{I}}_k$, the object detector generates a set of candidate detections represented as $\mathbb{D}_k = \{\mathbf{d}_k^i\}_{i=1}^{|\mathbb{D}_k|}$, where $|\cdot|$ indicates the cardinality operator. Each individual detection can be represented as

$$\mathbf{d}_k^i = \left(u_k^i, v_k^i, w_k^i, h_k^i\right), \tag{8}$$

where (u_k^i, v_k^i) is the center, (w_k^i, h_k^i) are the width and height of the detection on the image plane, respectively, and $i \in [1, |\mathbb{D}_k|]$ is the detection index. These detections can be inaccurate, generating false-positive or false-negative errors, and do not have any identity information.

For object tracking, we employ the early association probability hypothesis density particle filter (EA-PHD-PF) [19], which estimates the trajectory of multiple objects from noisy detections. Through four processing steps (*i.e.* prediction, early association, update and resampling) for each undistorted video frame $\tilde{\boldsymbol{I}}_k$, the algorithm approximates a state probability function using a set of particles $\mathbb{E}_k = \left\{\dot{\boldsymbol{e}}_k^i\right\}_{i=1}^{|\mathbb{E}_k|}$, where each particle $\dot{\boldsymbol{e}}_k^i = \{\lambda_k^i, \pi_k^i, \boldsymbol{e}_k^i\}$ is associated to the identity information λ_k^i, the posteriori probability π_k^i, and the state information

$$\boldsymbol{e}_k^i = \left(u_k^{\tilde{\lambda}}, v_k^{\tilde{\lambda}}, \dot{u}_k^{\tilde{\lambda}}, \dot{v}_k^{\tilde{\lambda}} w_k^{\tilde{\lambda}}, h_k^{\tilde{\lambda}}\right)_{\tilde{\lambda}=\lambda_k^i} \tag{9}$$

where $u_k^{\tilde{\lambda}}, v_k^{\tilde{\lambda}}, w_k^{\tilde{\lambda}}, h_k^{\tilde{\lambda}}$ are defined similarly as in Eq. 8 and $\dot{u}_k^{\tilde{\lambda}}$ and $\dot{v}_k^{\tilde{\lambda}}$ are the horizontal and vertical velocities, respectively, and $\lambda_k^i \in \{1, \cdots, \Lambda_k\}$ where Λ_k is the number of identities that are detected by the tracker. Finally, the state of each target with identity λ is estimated as

$$\boldsymbol{e}_k^{\lambda} = \frac{1}{\sum_i \pi_k^i} \sum_i \pi_k^i \boldsymbol{e}_k^i, \tag{10}$$

where $i \in \mathbb{I}_{\tilde{\lambda}}$ and $\mathbb{I}_{\lambda}$ denotes a set of indexes with $\lambda_k^i = \tilde{\lambda}$.

Based on Fig. 2, the DOA of each identified object in the frame $\tilde{\boldsymbol{I}}_k$ is estimated as

$$\theta_{vk}^{\lambda} = \arctan\left(\frac{u_k^{\lambda}}{F}\right),\ \lambda = 1, \cdots, N, \tag{11}$$

where $N = \Lambda_k$, and F is the focal length obtained as in Sec. 4.1.

5.2 Spatially informed audio enhancement

Given the potential sound emitting objects detected by the video tracker, we could design a set of spatial filters to extract the sounds from those visually informed directions. This is a challenging task due to the existence of strong ego-noise with extremely low SNR, *e.g.* < -15 dB. For this aim, we employ a time-frequency (T-F) spatial filtering, a recently emerged MAV sound enhancement approach [25]. This approach can extract the sound from the desired DOA from the strong ego-noise by exploiting the time-frequency sparsity of the acoustic signals. The visually-informed audio enhancement approach consists of five steps.

In the first step, we geometrically transform the video trajectory of each potential sound source, $\Theta_{vk}^{\lambda} = \left\{\theta_{vk}^{\lambda}\right\}_{k=1}^{K}$, to the audio reference system as

$$\theta_{ak}^{\lambda} = a_1\theta_{vk}^{\lambda} + a_2, \tag{12}$$

where a_1 and a_2 are the geometrical alignment parameters obtained in Sec. 4.2.

Second, we transform the time-domain signal $\boldsymbol{x}(n)$ into the time-frequency domain as $\boldsymbol{x}(\omega, l)$ via short-time Fourier transform (STFT) with frame length N_ω and shift size $N_s = \frac{N_\omega}{2}$, where ω and l are the frequency and audio frame indexes, respectively.

Suppose we have a segment of signal $l \in [l_b, l_e]$, corresponding to a time segment of $n \in [n_b, n_e]$ where $n_b = l_b N_s$ and $n_e = l_e N_s$. The DOA of the target sound in this segment is estimated as the median value among all video-informed estimates, *i.e.*

$$\theta_d = \text{median}\left\{\theta_{ak}^{\lambda}\right\}_{k\in[(n_b+\delta_{av})/f_c,(n_e+\delta_{av})/f_c]}, \tag{13}$$

where δ_{av} is the time offset between the audio and video streams, as obtained in Sec. 4.1.

Third, given the microphone signal $\boldsymbol{x}(\omega, l)$ and location of the microphones $\boldsymbol{R}$, we build a spatial likelihood function

$$\gamma_{\text{TF}}(\omega, l, \theta) = \Re\left\{ \sum_{\substack{m_1, m_2 = 1 \\ m_1 \neq m_2}}^{M} \frac{x_{m_1}(\omega, l) x_{m_2}^*(\omega, l)}{|x_{m_1}(\omega, l) x_{m_2}(\omega, l)|} e^{j2\pi f_\omega \tau(m_1, m_2, \theta)} \right\}, \quad (14)$$

where the superscript $(\cdot)^*$ denotes complex conjugation, the operator $\Re\{\cdot\}$ denotes the real component of the argument, and $\tau(m_1, m_2, \theta) = \frac{\|\boldsymbol{r}_{m_2} - \boldsymbol{r}_\theta\| - \|\boldsymbol{r}_{m_1} - \boldsymbol{r}_\theta\|}{c}$ denotes the time difference of arrival between the sound at two microphones m_1 and m_2, and c denotes the sound velocity in the air. The term $e^{j2\pi f_\omega \tau(m_1, m_2, \theta)}$ is the inter-channel phase difference theoretically computed with the delay τ; the term $\frac{x_{m_1}(\omega, l) x_{m_2}^*(\omega, l)}{|x_{m_1}(\omega, l) x_{m_2}(\omega, l)|}$ is the inter-channel phase difference measured from x_{m_1} and x_{m_2}. The spatial likelihood γ_{TF} is high when these two inter-channel phase differences are consistent with each other. The DOA can thus be estimated as

$$\theta_{\text{TF}}(\omega, l) = \underset{\theta \in (-180°, 180°]}{\arg\max}\ \gamma_{\text{TF}}(\omega, l, \theta). \quad (15)$$

Fourth, we detect the time-frequency bins that belong to the target sound, assuming that the time-frequency bins belonging to the target sound have their DOA estimates normally distributed around the mean θ_d, with standard deviation σ_d. The detection is performed by measuring the closeness of each time-frequency bin to the target sound:

$$c_d(\omega, l, \theta_d) = \exp\left(-\frac{(\theta_{\text{TF}}(\omega, l) - \theta_d)^2}{2\sigma_d^2} \right), \quad (16)$$

where $c_d(\cdot) \in [0, 1]$. The higher $c_d(\cdot)$, the higher the probability that the (ω, l)-th bin is dominated by the target sound. We then calculate the correlation matrix of the noisy microphone signal and of the target sound, *i.e.*

$$\boldsymbol{\Phi}_{xx}(\omega, l, \theta_d) = \frac{1}{l_e - l_b + 1} \sum_{l=l_b}^{l_e} \boldsymbol{x}(\omega, l) \boldsymbol{x}^{\text{H}}(\omega, l), \quad (17)$$

$$\boldsymbol{\Phi}_{ss}(\omega, l, \theta_d) = \frac{1}{l_e - l_b + 1} \sum_{l=l_b}^{l_e} c_d^2(\omega, l, \theta_d) \boldsymbol{x}(\omega, l) \boldsymbol{x}^{\text{H}}(\omega, l), \quad (18)$$

where the closeness measure $c_d(\omega, l, \theta_d)$ indicates the contribution of the (ω, l)-th bin to the correlation matrix, and the superscript $(\cdot)^{\text{H}}$ denotes the Hermitian transpose. Given this estimated target correlation matrix, an adaptive beamformer can be formulated easily. We use the multichannel Wiener filter [4]

$$\boldsymbol{w}_{\text{TF}}(\omega, l, \theta_d) = \boldsymbol{\Phi}_{xx}^{-1}(\omega, l) \boldsymbol{\phi}_{ss1}(\omega, l, \theta_d), \quad (19)$$

where $\boldsymbol{\phi}_{ss1}(\omega, l, \theta_d)$ is the first column of $\boldsymbol{\Phi}_{ss}(\omega, l, \theta_d)$. The sound coming from θ_d is extracted as

$$y_{\text{TF}}(\omega, l, \theta_d) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(\omega, l, \theta_d) \boldsymbol{x}(\omega, l). \quad (20)$$

Finally, we transform $y_{\text{TF}}(\omega, l, \theta_d)$ in the time-frequency back to the time domain, obtaining $y_{\text{TF}}(n, \theta_d)$, *i.e.* we can extract the sound from N potential speakers sequentially and represented them as $y_1(n), \cdots, y_N(n)$.

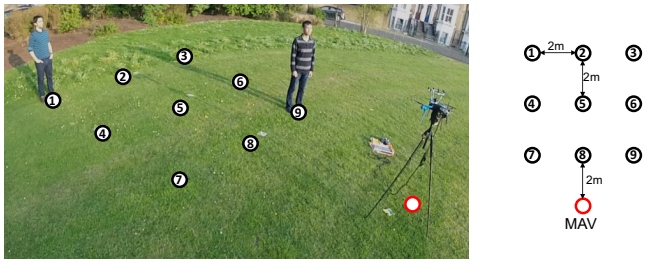

Figure 4: Experimental setup illustrating the projection of the position of the MAV on the ground (red circle) and the nine landmarks (black circles).

6 EVALUATION

6.1 Experimental setup

Dataset. We built a prototype that is composed of a 3DR Iris quadcopter, a GoPro camera, and a microphone array (Fig. 1). We used this prototype to record two datasets: the *evaluation* and the *demonstration* dataset. Fig. 4 depicts the recording setup, where people move among nine predefined landmarks in a park and the MAV is fixed on a tripod. In the evaluation dataset, we record the ego-noise and the speech separately in order to investigate the performance comprehensively. When recording the ego-noise, the MAV operates at 50%, 100% and 150% of the power level at hovering state. When recording the clean speech, we have two people moving randomly along the landmarks. At each location, they talk sequentially for about 40 s each and then move simultaneously to their next location. We mix the two speech signals and the ego-noise to generate the microphone signal. In the demonstration dataset, we record the ego-noise and the speech simultaneously. We have three people moving along the landmarks randomly. At each location, the three people randomly choose to talk alone or simultaneously, each for about 10 s, and then simultaneously move to their next location. The MAV operates at the power level of hovering state during the whole recording.

Algorithms for comparison. We compare the proposed multimodal (*audio-visual*) method, which steers the time-frequency spatial filter at the directions provided by the visual module, against a mono-modal (*audio-only*) method, which estimates the direction of the sound from the microphone signal [26] and then steers the time-frequency spatial filter at it. In addition, we compare time-frequency filtering with a traditional delay-and-sum beamformer [25].

Implementation details and parameters. The GoPro camera is set to record at a wide field of view, at 1920x1080 resolution and $f_c = 30\ Hz$. The audio processing employs a segment-wise processing scheme, which divides the audio signals into non-overlapped segments of 6 s long and processes them sequentially. The STFT frame length is set to 1024 with half overlap. The standard deviation in (16) is set to $\sigma_d = 10°$.

Evaluation measures. We are interested in evaluating the noise reduction performance in terms of signal-to-noise ratio (SNR), the separation between competing speakers in terms of signal-to-interference ratio (SIR) [27], and also the enhanced speech quality in terms of Perceptual Evaluation of Speech Quality (PESQ).

Given a spatial filter $\boldsymbol{w}(n)$ and the microphone signal $\boldsymbol{x}(n) = \sum_{\lambda=1}^{N} \boldsymbol{s}_\lambda(n) + \boldsymbol{v}(n)$ with its constituent components assumed to be known, the spatial filtering output can be written as

$$\begin{aligned} y(n) &= \boldsymbol{w}(n) * \boldsymbol{x}(n) = \sum_{\lambda=1}^{N} \boldsymbol{w}(n) * \boldsymbol{s}_\lambda(n) + \boldsymbol{w}(n) * \boldsymbol{v}(n) \\ &= \sum_{\lambda=1}^{N} y_{s_\lambda}(n) + y_v(n). \end{aligned} \tag{21}$$

The SNR and SIR for the λ-th source are calculated in target-sound-active periods $\mathbb{N}_{s_\lambda}$ as

$$\mathrm{SNR}_\lambda = 10 \log_{10} \frac{\sum_{n' \in \mathbb{N}_{s_\lambda}} y_{s_\lambda}^2(n')}{\sum_{n' \in \mathbb{N}_{s_\lambda}} \left(y_v^2(n') + \sum_{\lambda' \neq \lambda} y_{s_{\lambda'}}^2(n') \right)}, \tag{22}$$

$$\mathrm{SIR}_\lambda = 10 \log_{10} \frac{\sum_{n' \in \mathbb{N}_{s_\lambda}} y_{s_\lambda}^2(n')}{\sum_{n' \in \mathbb{N}_{s_\lambda}} \left(\sum_{\lambda' \neq \lambda} y_{s_{\lambda'}}^2(n') \right)}. \tag{23}$$

Finally, PESQ $\in [0, 4.5]$ is a widely-used measure to assess the overall quality of the processed speech $s_e(n)$ relative to the referenced clean speech $s_o(n)$ [34]. The higher PESQ, the better the speech quality. We represent the PESQ operator as $Q\{s_e, s_o\}$. The PESQ of the λ-th source is calculated by comparing the enhanced signal $y_{s_\lambda}(n)$ with the clean signal in the first microphone $s_{1\lambda}(n)$, *i.e.*

$$\mathrm{PESQ}_\lambda = Q\{y_{s_\lambda}, s_{1\lambda}\}. \tag{24}$$

6.2 Discussion

We first evaluate the sound enhancement performance of the time-frequency spatial filter assuming that the DOA of the speaker is known. Fig. 5 depicts the sound enhancement results, in terms of SNR and PESQ, for a single speaker with a varying distance (2 m, 4 m and 6 m) from the MAV, which operates at three different power levels. The locations of the speaker are ⑧, ⑤ and ①. For each evaluation case, we choose 5 segments of noisy data (each lasting 6 s) and calculate the average performance measure. The input SNR varies depending on the distance between the speaker and the MAV and also on the operation power of the MAV. In all evaluation cases, the input SNR varies between −20 dB and −30 dB, which indicates an extremely challenging scenario for sound enhancement. The PESQ values of the input microphone signals are all below 1.5. The time-frequency spatial filter tends to perform better for higher input SNRs. In all evaluation cases, the spatial filter can improve the SNR remarkably by up to 20 dB, and improve the PESQ by up to 1. In comparison to time-frequency filtering, the fixed beamforming performs much worse in all evaluation scenarios, even if the DOA is given.

Fig. 6 depicts the sound enhancement results, in terms of SIR, SNR and PESQ, for two speakers talking concurrently at a varying distance (2 m, 4 m and 6 m) from the MAV, which operates at hovering power level. The locations of the two speakers are (⑦, ⑨), (④, ⑥) and (①, ③). For each evaluation case, we choose 5 segments of noisy data (each lasting 6 s) and calculate the averaged performance measure. For each speaker, the input SIR is around 0 dB while the input SNR is below -20 dB, and the PESQ is below 1. The spatial filter can extract a target speaker by suppressing the

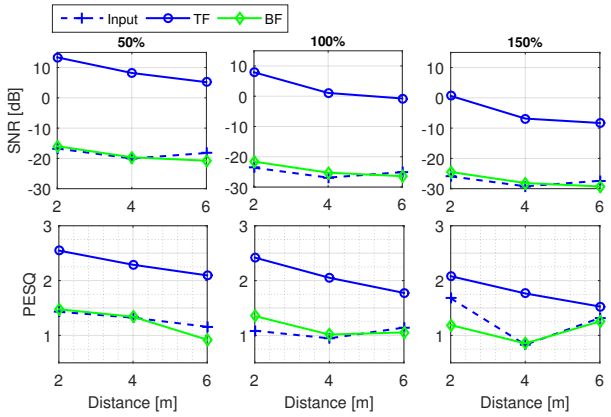

Figure 5: Sound enhancement by time frequency filtering (TF) and fixed beamforming (BF) for a single speaker with a varying distance from the MAV, which operates at 50%, 100%, and 150% of the hovering power level. Note that the DOA of the speaker is assumed to be known.

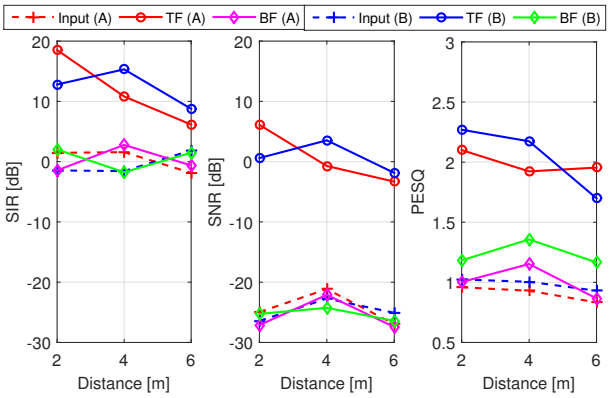

Figure 6: Sound enhancement by time frequency filtering (TF) and fixed beamforming (BF) for two speakers talking concurrently at varying distances from the MAV, which operates at the hovering power level. Note that the DOAs of the two speakers are assumed to be known.

interfering speaker and the ego-noise simultaneously. The spatial filter can isolate the two speakers well by improving their input SIR by up to 10 dB. The spatial filter can improve the input SNR by up to 20 dB and improve the PESQ value by up to 1.5. The fixed beamforming performs much worse than time-frequency filtering in all evaluation scenarios.

We then evaluate the sound enhancement performance of the two types of spatial filters (audio-only and audio-visual) when processing the evaluation sequence continuously in a segment-by-segment style. Fig. 7 presents the processing results for a single speaker whose location varies with time and the MAV is operating at the hovering power level. Fig. 7(a) depicts the trajectory (⑦→④→①→③) of the speaker as well as their voice activity

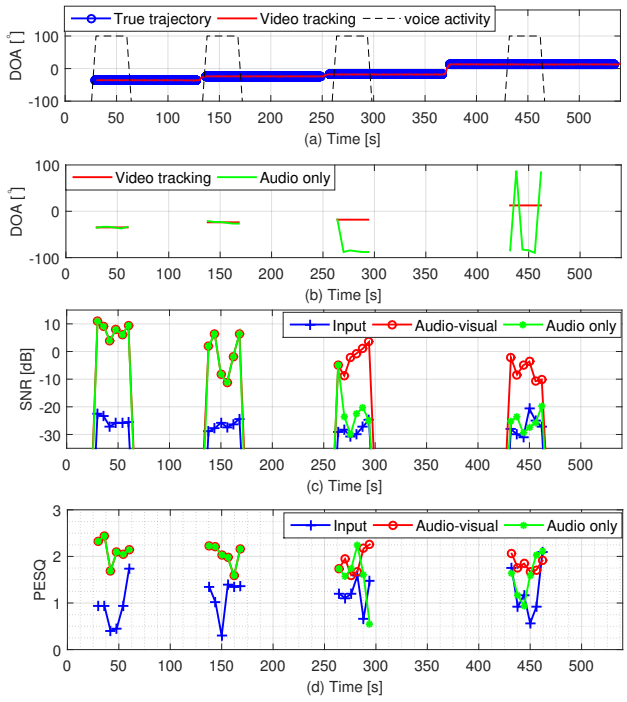

Figure 7: Results for a single speaker. (a) Trajectory and voice activity of the speaker. (b) Estimated DOAs from audio and video. (c) SNR of the enhanced sound by two spatial filters (audio-only and audio-visual). (d) PESQ of the enhanced sound by two spatial filters.

(manually labeled). The video tracker can capture the location of the speaker accurately. Fig. 7(b) compares the DOA estimation results between the video tracker and the audio-only localizer. The audio localization results are consistent with the video tracking results when the speaker is close to the MAV and the input SNR is relatively high (*i.e.* in the first two positions). The audio localization results deviate from the video tracking results significantly when the speaker is farther from the MAV and the input SNR becomes lower (*i.e.* in the last two positions). Fig. 7(c) presents the SNR, which is calculated in speech-active periods only, obtained by the two spatial filters. For all four positions, the input SNR is below -20 dB and decreases as the distance increases. For the first two positions, audio-only and audio-visual spatial filters perform similarly and improve the SNR by up to 20 dB. For the last two positions, the audio-visual spatial filter can still improve the SNR up to 20 dB, while the audio-only spatial filter fails. This behavior is expected since the results are consistent with the DOA estimation shown in Fig. 7(b). Consequently, as shown in Fig. 7(d), the audio-visual spatial filter can improve the PESQ value of the input signal by up to 1 for all four positions, while the audio-only spatial filter works only for the first two positions.

Fig. 8 and Fig. 9 present the processing results when two speakers are in the scene. The trajectory of the two speakers are speaker A: ⑦ → ④ → ① → ③ and speaker B: ⑨ → ⑧ → ⑤ → ①. Fig. 8(a) depicts the trajectories of the video tracking results for the

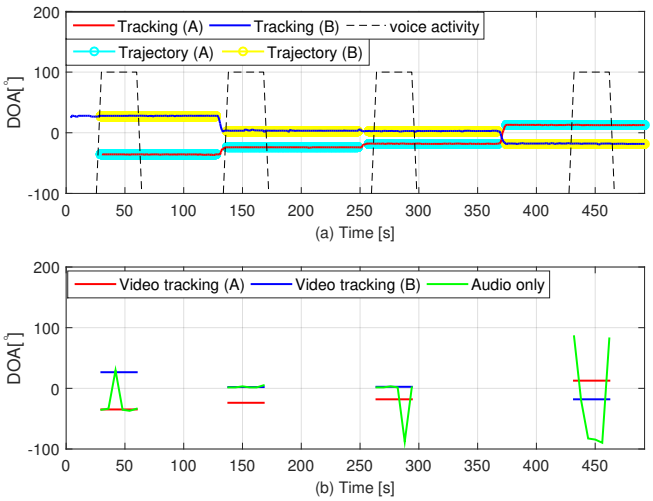

Figure 8: Results when two speakers are talking concurrently. (a) Trajectory and voice activity of the speakers. (b) Estimated DOAs from audio and video.

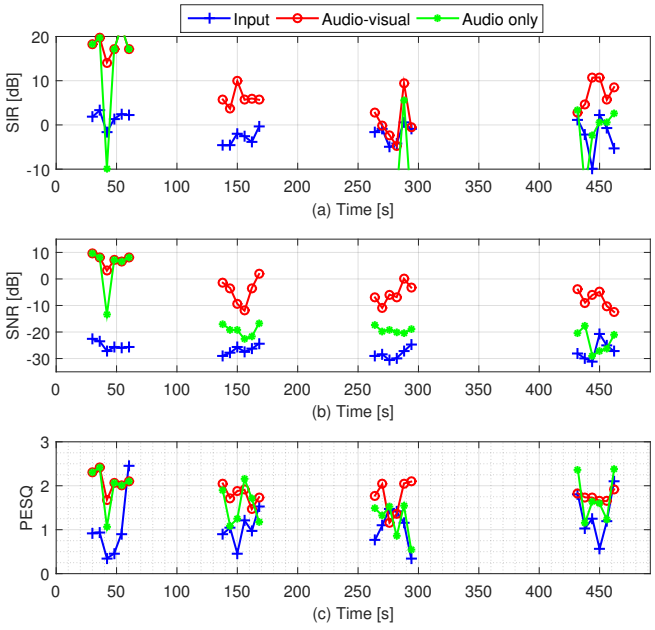

Figure 9: Results for speaker A when two speakers are talking concurrently. (a) SIR of the enhanced sound by two spatial filters (audio-only and audio-visual). (b) SNR of the enhanced sound by two spatial filters. (c) PESQ of the enhanced sound by two spatial filters.

two speakers. It can be clearly observed that the video tracker can capture the location of both speakers accurately. Fig. 8(b) compares the DOA estimation results of the video tracker and the audio-only localizer. The audio localizer has only one output and cannot handle the ambiguities in the multi-speaker scenario. The audio

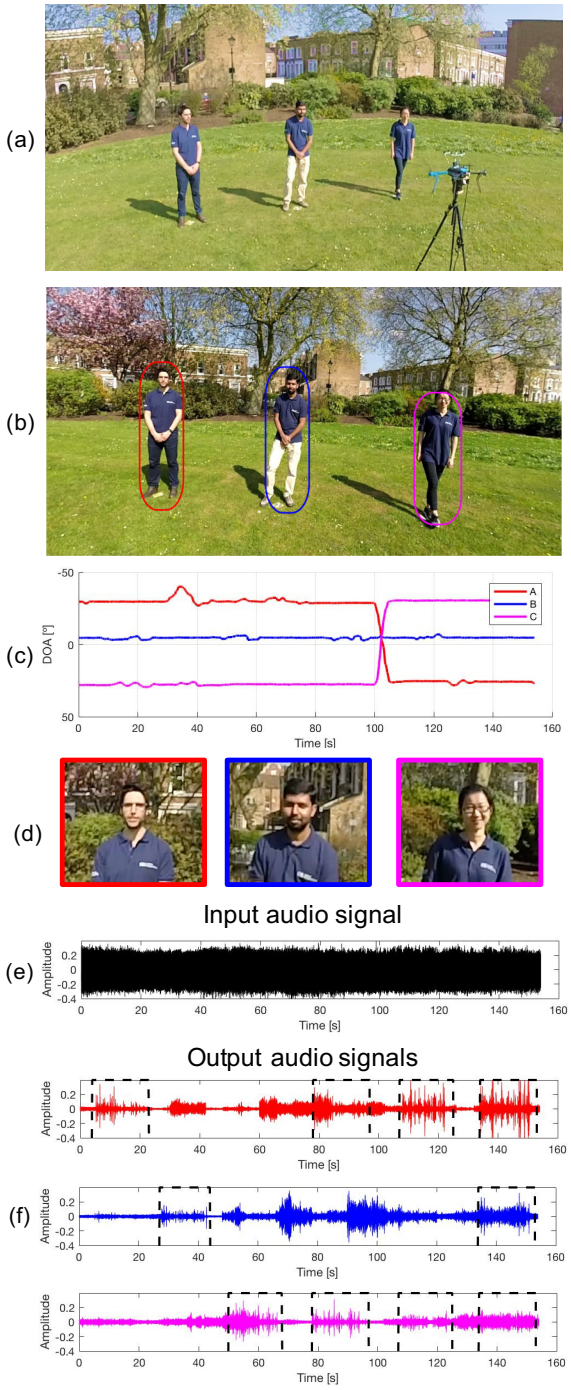

Figure 10: Sample results on the demonstration dataset. The proposed method tracks the DOAs of three people and extracts the sound of each speaker from the noisy microphone signal. (a) Recording setup. (b) Undistorted video frame. (c) Video tracking results. (d) Crops of the upper-body of the speakers. (e) Waveforms of the original microphone signal. (f) Enhanced audio stream for each speaker. The voice activity of each speaker is manually labeled.

localization result either detects only one speaker (*i.e.* in the first three positions) or deviates from both speakers (*i.e.* in the fourth position). We obtain similar sound enhancement results for the two speakers and thus only show the results for speaker A achieved by two spatial filters (audio-visual and audio-only) in Fig. 9. The audio-visual spatial filter clearly outperforms the audio-only one in terms of SIR, SNR and PESQ.

Finally, Fig. 10 shows the results for the demonstration dataset where the trajectories of the three speakers are A: ④ → ④ → ④ → ④ → ⑥ → ⑥, B: ⑤ → ⑤ → ⑤ → ⑤ → ⑤ → ⑤ and C: ⑥ → ⑥ → ⑥ → ⑥ → ④ → ④. The voice activity of each speaker during the recording is manually labeled. Based on the DOA informed by the tracker, the time-frequency spatial filter can extract the sound of each speaker from the noisy microphone signals. The visual tracker can robustly track each speaker even under the severe visual occlusions that happen at around 100 s (see Fig. 10(c)).

7 CONCLUSIONS

We explored the combination of audio and visual modalities to enhance sounds captured from an MAV. The visual module employs a multi-object tracker that locates potential sound emitting objects, whereas the audio module employs a time-frequency spatial filtering technique to enhance the sound from the directions provided by the video module. We have shown that by exploiting the two modalities the proposed method can isolate the sound of individual speakers in extremely low-SNR scenarios.

In future work, we will extend the proposed method to cope with flying MAVs with the additional challenge introduced by the movement of the camera and the microphones.

REFERENCES

[1] M. Basiri, F. Schill, P. U. Lima, and D. Floreano. 2012. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vilamoura-Algarve, Portugal, 4737–4742.

[2] W. Choi. 2015. Near-online multi-target tracking with aggregated local flow descriptor. In *Proc. of International Conference on Computer Vision*. Santiago, Chile, 3029–3037.

[3] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. of Computer Vision and Pattern Recognition*. San Diego, CA, USA, 886–893.

[4] S. Doclo and M. Moonen. 2002. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing* 50, 9 (Sept. 2002), 2230–2244.

[5] P. Dollar, R. Appel, S. Belongie, and P. Perona. 2014. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (Aug. 2014), 1532–1545.

[6] P. F. Felzenszwalb, R.B. Girshick, and D. Ramanan. 2010. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (Sept. 2010), 1627–1645.

[7] A. García-Martín, R. Sanchez-Matilla, and J. M. Martínez. 2017. Hierarchical detection of persons in groups. *Signal, Image and Video Processing* (2017), 1–8.

[8] J. Heikkila and O. Silven. 1997. A four-step camera calibration procedure with implicit image correction. In *Proc. of Computer Vision and Pattern Recognition*. 1106–1112.

[9] T. Ishiki and M. Kumon. 2015. Design model of microphone arrays for multirotor helicopters. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*. Hamburg, Germany, 6143–6148.

[10] F.S. Khan, R.M. Anwer, J. van de Weijer, A.D. Bagdanov, M. Vanrell, and A.M. Lopez. 2012. Color attributes for object detection. In *Proc. of Computer Vision and Pattern Recognition*. Providence, RI, USA, 3306–3313.

[11] R. Mahler. 2002. A theoretical foundation for the Stein-Winter probability hypothesis density (PHD) multitarget tracking approach. In *Proc. of MSS National Symposium on Sensor and Data Fusion*. San Diego, CA, USA.

[12] R. Mahler. 2003. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transaction on Aerospace and Electronic Systems* 39, 4 (Oct. 2003), 1152–1178.

[13] MathWorks. 2013. MATLAB camera calibration toolbox. (2013). https://uk.mathworks.com/help/vision/ref/cameracalibrator-app.html

[14] S. M. Naqvi, W. Wang, M. S. Khan, M. Barnard, and J. A. Chambers. 2012. Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking. *IET Signal Processing* 6, 5 (July 2012), 466–477.

[15] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and L. Nakadai. 2014. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Chicago, USA, 1902–1907.

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

[17] H. Possegger, T. Mauthner, P.M. Roth, and H. Bischof. 2014. Occlusion geodesics for online multi-object tracking. In *Proc. of Computer Vision and Pattern Recognition*. Columbus, OH, USA, 1306–1313.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: unified, real-time object detection. In *Proc. of Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, 779–788.

[19] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. 2016. Online multi-target tracking with strong and weak detections. In *Proc. of European Conference on Computer Vision*. Amsterdam, The Netherlands, 84–99.

[20] S. Uemura, O. Sugiyama, R. Kojima, and K. Nakadai. 2015. Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array. In *IEEE International Conference on Advanced Intelligent Mechatronics*. Tokyo, Japan, 329–330.

[21] B.-N. Vo, S. Singh, and A. Doucet. 2003. Sequential Monte Carlo implementation of the PHD filter for multi-target tracking. In *Proc. of Information Fusion*, Vol. 2. Queensland, Australia, 792–799.

[22] J. Wang, Y. Liang, and J. Wilder. 1998. Visual-information-assisted microphone array processing in a high-noise environment. In *Proceedings of SPIE*. Boston, MA, USA, 198–203.

[23] L. Wang. 2014. Multi-band multi-centroid clustering based permutation alignment for frequency-domain blind speech separation. *Digital Signal Processing* 31, 1 (2014), 79–92.

[24] L. Wang and A. Cavallaro. 2016. Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles. In *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*. Colorado Springs, USA, 152–158.

[25] L. Wang and A. Cavallaro. 2017. Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles. *IEEE Sensors Journal* 17, 8 (Feb 2017), 2447–2455.

[26] L. Wang and A. Cavallaro. 2017. Time-frequency processing for sound source localization from a micro aerial vehicle. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. New Orleans, LA, USA, 496–500.

[27] L. Wang, Gerkmann, and S. Doclo. 2015. Noise power spectral density estimation using MaxNSR blocking matrix. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 9 (Sept. 2015), 1493–1508.

[28] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro. 2016. An iterative approach to source counting and localization using two distant microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 6 (June 2016), 1079–1093.

[29] L. Wang, J. D. Reiss, and A. Cavallaro. 2016. Over-determined source separation and localization using distributed microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 9 (Sept. 2016), 1569–1584.

[30] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers. 2005. Video assisted speech source separation. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Philadelphia, PA, USA, 425–428.

[31] X. Wang, T.X. Han, and S. Yan. 2009. An HOG-LBP human detector with partial occlusion handling. In *Proc. of International Conference on Computer Vision*. Kyoto, Japan, 32–39.

[32] F. Yang, W. Choi, and Y. Lin. 2016. Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. of Computer Vision and Pattern Recognition*. Las Vegas, NV, USA, 2129–2137.

[33] Y. Yang, G. Shu, and M. Shah. 2013. Semi-supervised learning of feature hierarchies for object detection in a video. In *Proc. of Computer Vision and Pattern Recognition*. Portland, OR, USA, 1650–1657.

[34] H. Yi and P. C. Loizou. 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 1 (Jan. 2008), 229–238.

[35] S. Yoon, S. Park, Y. Eom, and S. Yoo. 2015. Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters. In *IEEE International Conference on Consumer Electronics*. Las Vegas, USA, 26–29.

[36] Z. Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 11 (Nov. 2000), 1330–1334.