Tracking a moving sound source from a multi-rotor drone

Lin Wang, Ricardo Sanchez-Matilla, Andrea Cavallaro

***Abstract*— We propose a method to track from a multi-rotor drone a moving source, such as a human speaker or an emergency whistle, whose sound is mixed with the strong ego-noise generated by rotating motors and propellers. The proposed method is independent of the specific drone and does not need pre-training nor reference signals. We first employ a time-frequency spatial filter to estimate, on short audio segments, the direction of arrival of the moving source and then we track these noisy estimations with a particle filter. We quantitatively evaluate the results using a ground-truth trajectory of the sound source obtained with an on-board camera and compare the performance of the proposed method with baseline solutions.**

I. Introduction

Tracking the time-varying direction of arrival of a sound source with microphones on a small drone is important for human-robot interaction, surveillance, and search and rescue applications [1]–[10]. However, most sound source localization algorithms for robot audition operate with high input signal-to-noise ratios (e.g. indoors) and are not directly applicable to multi-rotor drones [11]–[13]. Robot audition with multi-rotor drones operates under extremely low signal-to-noise ratios (e.g. SNR $<$ -20 dB [8]) because of the natural and motion-induced wind, and the strong and time-varying ego-noise generated by motors and propellers [14]–[18].

Sound source localization approaches for drones can be supervised or unsupervised [15]. Supervised approaches estimate the correlation matrix of the ego-noise [19]–[22] and build a noise template database with pre-recorded sounds to predict the ego-noise based on the drone behavior. The behavior is monitored with additional sensors, which limit the versatility of these approaches. Unsupervised approaches use instead microphone signals only. Steered response power with phase transform (SRP-PHAT) [6], which computes a spatial likelihood map by exploiting the correlation between microphone signals, tends to show degraded performance with drones [23]. Approaches based on Multiple signal classification (MUSIC) [19], [24] need a dedicated microphone array calibration procedure [23]. To improve robustness to noise, Generalized eigenvalue decomposition MUSIC (GEVD-MUSIC) [19] exploits as additional information a noise correlation matrix, whose acquisition is however still an open problem, due to the low SNR and to the non-stationarity of the ego-noise [22].

The authors are with the Centre for Intelligent Sensing, Queen Mary University of London, U.K. E-mail: {lin.wang; ricardo.sanchezmatilla; a.cavallaro}@qmul.ac.uk.

This work was supported in part by ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project under Grant 332913. The support of the UK EPSRC project NCNR (EP/R02572X/1) is also acknowledged.

Time-frequency (TF) processing exploits that the energy of the sound recording is concentrated at isolated time-frequency bins by first estimating the direction of arrival (DOA) of the sound at individual time-frequency bins and then formulating a set of spatially informed filters pointing at candidate directions [23]. The location of the sound sources is estimated by measuring the non-Gaussianity of the spatial filter output. However, this method requires microphones and sound sources to remain static while the correlation matrix of the sound source is estimated. More recently, a spatial filter was combined with computer vision for multi-modal sound source localization (and enhancement) [9].

To the best of our knowledge, existing methods only consider the localization of *static* sources from a multi-rotor drone. A key challenge in tracking *moving* sound sources is to estimate the time-varying direction of arrival in short temporal windows. An approach exists to track a sound source from a moving fixed-wing drone [6], which however produces much lower ego-noise than a multi-rotor drone. In this paper, we propose the first method to track a moving sound source from a multi-rotor drone. We segment the audio streams from a microphone array into blocks (temporal windows) and, in each block, we estimate the location of the source with a time-frequency filter. We then track with a particle filter these noisy estimations. Moreover, we exploit the knowledge that the location of motors and propellers is fixed with respect to the array to predict the spatial characteristics of the ego-noise, and to improve the localization performance when the source moves in front of the drone.

II. Problem definition

Let n be the time index. A target source moves in the far field emitting sound with a time-varying DOA, $\theta(n)$, with respect to the heading of the drone, which hovers stably while recording the sound.

Let the superscript $(\cdot)^{\mathrm{T}}$ denote the transpose operator. The locations of M microphones of an array are $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m, \ldots, \boldsymbol{r}_M]$, where $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\mathrm{T}}$ is the position of the m-th microphone on the 2D coordinate system of the microphone array plane. This position can be measured manually or estimated with microphone array calibration methods [25].

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

Given only $\boldsymbol{x}(n)$ and $\boldsymbol{R}$, our goal is to estimate and track the time-varying DOA of the sound source.

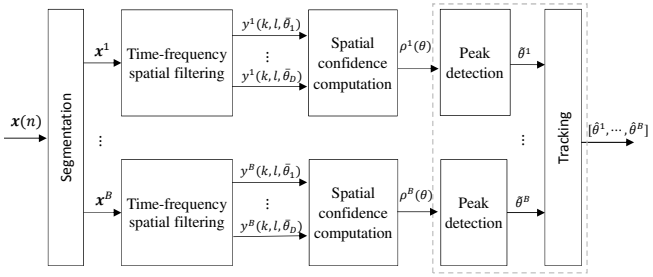


Fig. 1. Block diagram of the proposed moving sound source tracker.

III. Sound source localization and tracking

We extend the original batch method that assumes a static sound source [23] to a block-wise processing scheme in order to capture the motion of the sound source. The proposed method consists of three main steps, namely time-frequency spatial filtering, spatial confidence computation, and peak detection and tracking (see Fig. 1).

A. Time-frequency spatial filtering

We first segment $\boldsymbol{x}(n)$, using a sliding window of size W and skip size $W/2$, as B blocks: $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^b, \ldots, \boldsymbol{x}^B\}$, where the b-th block contains W samples, i.e. $\boldsymbol{x}^b = \left[\boldsymbol{x}(\frac{(b-1)W}{2}+1), \cdots, \boldsymbol{x}(\frac{(b-1)W}{2}+W)\right]$.

We also represent the samples in the b-th block as $\boldsymbol{x}^b(\tilde{n}) = [x_1^b(\tilde{n}), \ldots, x_M^b(\tilde{n})]^{\mathrm{T}}$, where $\tilde{n} \in [1, W]$ denotes the sample index in this block. We transform $\boldsymbol{x}^b(\tilde{n})$ into the short-time Fourier transform (STFT) domain: $\boldsymbol{x}^b(k,l) = [x_1^b(k,l), \ldots, x_M^b(k,l)]^{\mathrm{T}}$, where k and l denote the frequency and frame indexes, respectively. Let K and L denote the total number of frequency bins and time frames, respectively.

Given $\boldsymbol{R}$, we estimate the local DOA of the sound at each time-frequency bin and construct a time-frequency spatial filter[1], $\boldsymbol{w}_{\mathrm{TF}}(b,k,l,\theta)$, that points at a specific direction θ. The sound from direction θ is extracted as

$$y_{\mathrm{TF}}^b(k,l,\theta) = \boldsymbol{w}_{\mathrm{TF}}^{\mathrm{H}}(b,k,l,\theta)\boldsymbol{x}^b(k,l), \tag{1}$$

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

We then define a set of D candidate directions, $\theta \in \{\bar{\theta}_1, \cdots, \bar{\theta}_D\}$ for source localization and formulate the corresponding D spatially informed filters pointing at these directions thus producing $\{y_{\mathrm{TF}}^b(k,l,\bar{\theta}_1), \cdots, y_{\mathrm{TF}}^b(k,l,\bar{\theta}_D)\}$.

B. Spatial confidence

When the spatial filter is steered towards the target sound source, the filter output tends to present high non-Gaussianity, which we measure to build a spatial likelihood function for the estimation of the direction of the target sound [23]. The non-Gaussianity of a sequence can be measured with its statistical kurtosis $\mathcal{K}(\cdot)$, whose value at each frequency bin k and direction θ, $\xi^b(k,\theta)$, is

$$\xi^b(k,\theta) = \mathcal{K}(\tilde{\boldsymbol{y}}_{\mathrm{TF}}^b(k,\theta)), \tag{2}$$

[1]The details to compute the spatial filter can be found in [23].

where $\tilde{\boldsymbol{y}}_{\mathrm{TF}}^b(k,\theta)$ denotes the time sequence $|y_{\mathrm{TF}}^b(k,:,\theta)|$. The higher the kurtosis, the higher the non-Gaussianity [27]. The location of the sound source can thus be estimated by comparing the non-Gaussianity of the D spatial filtering outputs.

Considering the whole frequency band, we calculate a spatial confidence function of block b as

$$\tilde{\rho}^b(\theta) = \frac{1}{K}\sum_{k=1}^{K}\xi^b(k,\theta), \tag{3}$$

which we normalize as

$$\rho^b(\theta) = \frac{\tilde{\rho}^b(\theta)}{\max\limits_{\theta\in\{\bar{\theta}_1,\cdots,\bar{\theta}_D\}}(\tilde{\rho}^b(\theta))}, \tag{4}$$

where $\max(\cdot)$ denotes the maximum value of the sequence.

C. Peak detection and tracking

The spatial confidence function, $\rho^b(\theta)$, usually contains multiple noisy peaks that correspond to the target source and the ego-noise sources. Selecting the location with the highest peak (as done in [23]) may lead to erroneous results. We solve this problem with two steps: peak detection and tracking.

The ego-noise mainly consists of the sound emitted from the motors and the propellers. The motor sound can be interpreted as point sources whose directions are static with respect to the position of the microphones. The propeller sound originates from the swept area of the rotating blades and its direction spreads widely around the directions of the motor sound. When the microphone array is placed at the front of the body of the drone (see Fig. 2), the ego-noise tends to arrive from the side closer to the motors (the back of the array) thus creating a sector with lower ego-noise (the front of the array). Fig. 3(a) shows the SRP-PHAT functions computed at two random frames (each 2048-sample long) [26], where four peaks, corresponding to the four motors, can be observed. Fig. 3(b) shows the histogram of the local DOA estimation at individual time-frequency bins [23]. The histogram has lower values in the sector $[-45°, 45°]$. We name this sector, where we presume that a target sound can be more easily detected, as *noiseless sector* [15].

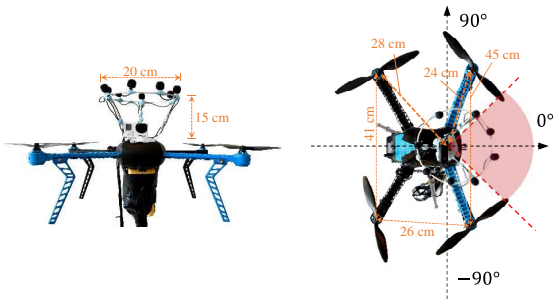
The sound source direction is considered to be the peak with highest confidence in the noiseless sector and it is calculated as

$$\tilde{\theta}^b = \underset{\theta\in[-45°,45°]}{\arg\max}\,\rho^b(\theta), \tag{5}$$

where $\rho^b(\theta)$ is the spatial confidence function of block b. To track $\tilde{\theta}^b$ while filtering out the noisy spatial confidence function received at each block b, we propose to use a particle filter [28], [29]. The particles are defined as

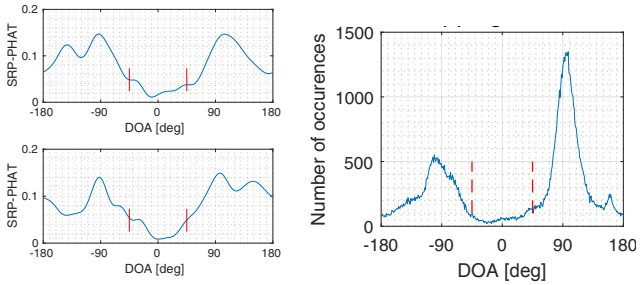
$$\Theta_i^b = \left[\acute{\theta}_i^b, \dot{\theta}_i^b\right], \tag{6}$$

where $\acute{\theta}_i^b$ is the estimated sound direction and $\dot{\theta}_i^b$ is the angular velocity at block b. Each particle has an associated



(a) (b)

Fig. 2. The multi-rotor drone with an 8-microphone circular array and, for tracking performance evaluation, a camera mounted at the center of the array. (a) Front view and (b) top view. The noiseless sector is indicated with a red shadowed area.



(a) (b)

Fig. 3. Localization results using ego-noise only. (a) SRP-PHAT functions at two random frames. (b) Histogram of the DOA estimates at individual time-frequency bins for a 30-second ego-noise segment. The noiseless sector [-45°, 45°] is indicated with red lines.

weight δ_i^b that informs how well a particle represents the actual location of the target. A particle filter typically consists of four steps: prediction, update, state estimation and resampling.

The prediction step operate as

$$\begin{aligned} \acute{\theta}_i^b &= \acute{\theta}_i^{b-1} + \dot{\theta}_i^{b-1} + \mathcal{N}(0;\sigma_p), \\ \dot{\theta}_i^b &= \dot{\theta}_i^{b-1} + \mathcal{N}(0;\sigma_{\dot{p}}), \end{aligned} \tag{7}$$

where $\mathcal{N}(0;\sigma_p)$ and $\mathcal{N}(0;\sigma_{\dot{p}})$ are the Gaussian noise on the source direction and velocity, and σ_p and $\sigma_{\dot{p}}$ are their standard deviations.

The update step calculates the weights of the particles given the observed sound source direction (Eq. (5)) as

$$\delta_i^b = \frac{1}{\sqrt{2\pi}\sigma_u} \mathrm{e}^{-\frac{\left(\tilde{\theta}^b - \acute{\theta}_i^b\right)^2}{2\sigma_u^2}}, \tag{8}$$

where σ_u is the standard deviation that accounts for the observation noise. Next, the state estimation step calculates the direction of arrival of the sound as

$$\hat{\theta}^b = \sum_{i=1}^{N} \delta_i^b \, \acute{\theta}_i^b, \tag{9}$$

where N is the number of particles. Finally, a resampling step discards particles with very low weight and duplicates particles with higher weight.

Fig. 4. Experimental setup. A loudspeaker is carried by a person who walks in front of the drone that is placed on a tripod. The noiseless sector is indicated with yellow marks on the ground.

The tracking results are estimated for each of the B blocks, giving $\left[\hat{\theta}^1, \cdots, \hat{\theta}^b, \cdots, \hat{\theta}^B\right]$.

IV. Dataset

We built a prototype (Fig. 2) composed of a 3DR IRIS quadcopter, an 8-microphone circular array (diameter $d =$ 20 cm), and a GoPro camera [8]. To avoid the self-generated wind blowing downwards from the propellers, the array is fixed 15 cm above the body of the drone. The microphone signals are sampled synchronously with a multichannel audio recorder (Zoom R24). The camera is mounted at the center of the microphone array[2].

We placed the prototype on a tripod at a height of 1.8 m in a park to record speech as sound played by a loudspeaker carried by a person walking in front of the drone (Fig. 4). The drone operates with a constant hovering power, or with a time-varying power between 50% and 150% of the hovering status. The loudspeaker is moving inside the noiseless sector only (indicated by the yellow marks in Fig. 4) or freely in front of the drone. The distance between the loudspeaker and the drone is 2 to 6 m.

We define four scenarios: $\mathcal{S}1$ (the loudspeaker moves in the constrained sector only and the drone is at constant power); $\mathcal{S}2$ (the loudspeaker moves in the constrained sector only and the drone generates ego-noise with time-varying power); $\mathcal{S}3$ (the loudspeaker moves freely and the drone is at constant power); and $\mathcal{S}4$ (the loudspeaker moves freely and the drone generates ego-noise with time-varying power).

We recorded a *natural* and a *composite* dataset, each including all four scenarios (lasting 3 minutes each). The ego-noise and speech are recorded simultaneously in the natural dataset and separately in the composite dataset. The recordings are available at http://www.eecs.qmul.ac.uk/∼andrea/sst.html.

V. Experiments

We evaluate the tracking performance by comparing the trajectory estimated by the sound source tracker, which is updated every half-block interval ($W/2$), with the ground-truth trajectory generated from the video of the on-board

[2]We use the camera to obtain a ground-truth trajectory of the moving sound source. To facilitate this task, we attach a visual marker on the source. Since the microphones and the camera work independently, a calibration procedure is needed to align temporally and geometrically the audio and video signals (for details see [9]).

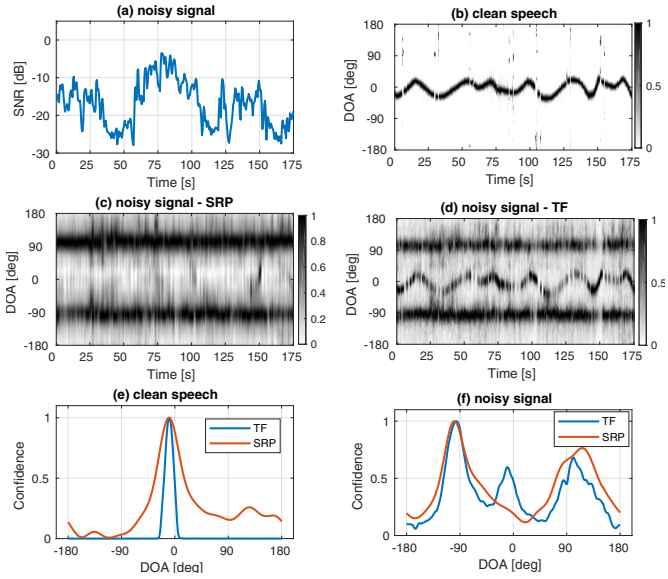


Fig. 5. Source localization results by SRP and TF on the composite sequence in $\mathcal{S}1$, with $W = 1$ s. (a) SNR of the noisy input signal. (b)-(d) The confidence map of the clean speech, the noisy signal processed by SRP, and the noisy signal process by TF, respectively. (e)-(f) The confidence function in the 86-th block (43 s) for the clean speech and the noisy signal, respectively.

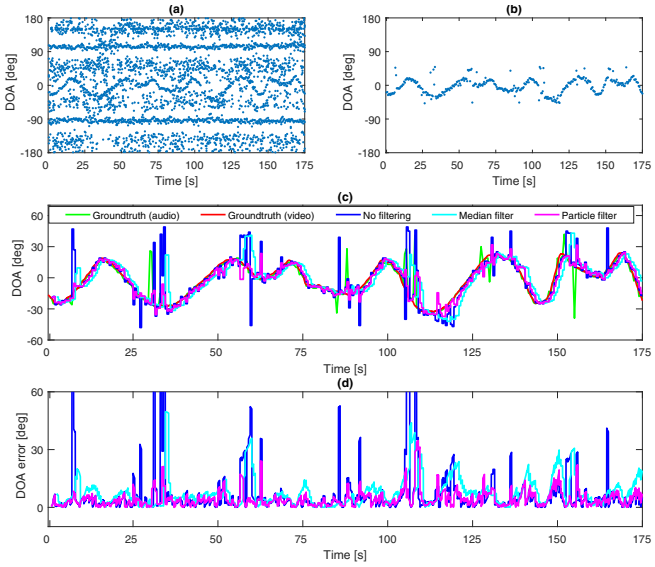


Fig. 6. Tracking results on the composite sequence in $\mathcal{S}1$, with $W = 1$ s. (a) Original peak detection result. (b) Proposed peak detection result. (c) Trajectories generated by different trackers. (d) Tracking errors obtained by different trackers.

camera. We measure the tracking error, bounded at $180°$, at the video frame rate (30 Hz), and calculate the mean and standard deviation across the whole trajectory.

We compare particle filtering (PF) with, as baseline methods, median filtering (MF) and no filtering (NF) on the localization at individual blocks. MF updates the localization at the b-th block as the median value, $\mathcal{M}(.)$, of the localization results across a sequence of W_p blocks:

$$\hat{\theta}^b_{\text{MF}} = \mathcal{M}\left(\tilde{\theta}^{b-W_p+1}, \cdots, \tilde{\theta}^b\right), \tag{10}$$

where W_p is predefined constant. NF uses the localization at the b-th block without any processing:

$$\hat{\theta}^b_{\text{NF}} = \tilde{\theta}^b. \tag{11}$$

We use four block sizes, $W \in \{0.5, 1, 2, 3\}$ s, and in each block we use a STFT of size 1024 and 50% overlap. We set the search area as $[-180°, 180°]$ with an interval of $1°$, i.e. $D = 361$. We set the noiseless sector as $[-45°, 45°]$. For the particle filter we set $N = 1000$, and we use a different set of parameters, empirically chosen, for each block size: for $W = 0.5$ s, $\sigma_p = 3.5°$, $\sigma_u = 10°$, and $W_p = 8$; for $W = 1$ s, $\sigma_p = 5.5°$, $\sigma_u = 4.5°$, and $W_p = 4$; for $W = 2$ s, $\sigma_p = 7.5°$, $\sigma_u = 3.5°$, and $W_p = 2$; for $W = 3$ s, $\sigma_p = 9.5°$, $\sigma_u = 3°$ and $W_p = 1$. $\sigma_{\dot{p}} = 0.05$ for all setups. Unless otherwise specified, $W = 2$ s in the comparisons.

Fig. 5 compares the localization results obtained by the time-frequency (TF) and the steered response power (SRP) approaches [23], with $W = 1$ s, in the composite sequence recorded in $\mathcal{S}1$. Fig. 5(a) depicts the temporal variation of the SNR, computed per processing block [30]. The SNR, which varies significantly across the blocks, is lower than -10 dB in most blocks and can be lower than -25 dB in some blocks (e.g. between 30 s and 60 s). Fig. 5(b)-(d) shows the confidence map computed using clean speech, SRP and TF, respectively. In Fig. 5(b), the trajectory of the clean speech can be observed clearly. In Fig. 5(c), the trajectory of the ego-noise, but not that of the speech, can be observed. In Fig. 5(d), the trajectories of both the ego-noise and the speech can the observed. Fig. 5(e) depicts the confidence function computed by SRP and TF, respectively, at the 86-th block (around 43 s), where both approaches can detect the peak location correctly. Fig. 5(f) shows the confidence function computed by SRP and TF for the noisy signal in the same block. In this low-SNR scenario (-24.9 dB), SRP detects two peaks of the ego-noise only, while TF detects three peaks, including the one from the speech.

Fig. 6 shows intermediate tracking results based on the confidence map in Fig. 5(d). Fig. 6(a) depicts the original peak detection results, where we retain 10 peaks per processing block in the whole circular area [-180°, 180°]. The confidence map contains considerable noise but the trajectory of the speech can still be observed. Fig. 6(b) depicts the proposed peak detection results, where only one peak is detected in the noiseless sector [-45°, 45°]. The proposed method can remove the spurious peaks in Fig. 6(a) effectively. Fig. 6(c) depicts the ground-truth trajectory of the sound source, the trajectory of the clean speech, the trajectory from detection without filtering (i.e. Fig. 6(b)), the tracking results with MF and PF. All the three trackers (NF, MF, and PF) can well capture the trajectory of the moving sound source. Fig. 6(d) depicts the tracking errors: PF has the smallest variations. The mean (standard deviation) localization errors by NF, MF, and PF are 5.8°(9.0°), 5.5°(5.5°) and 4.5°(4.3°), respectively. The proposed peak detection method can produce good localization results,

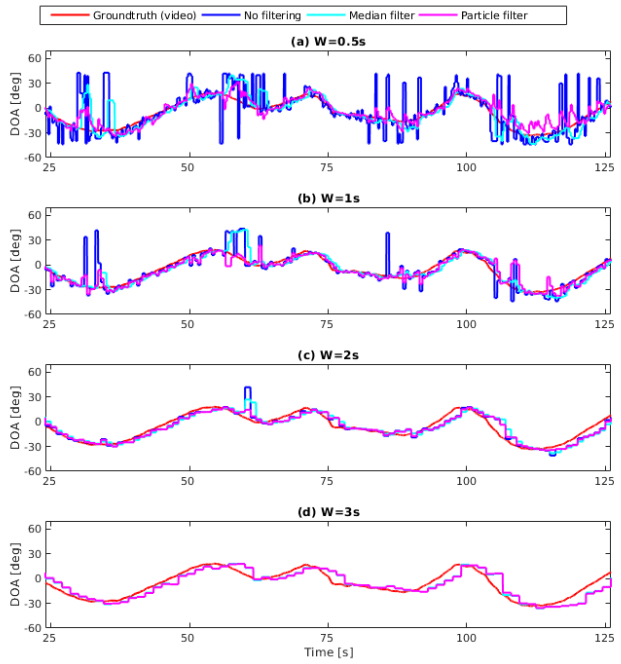


Fig. 7. Tracking results on the composite sequence in $\mathcal{S}1$ with different block sizes $W \in \{0.5, 1, 2, 3\}$ s.

TABLE I

LOCALIZATION ERRORS ON THE COMPOSITE SEQUENCE IN $\mathcal{S}1$ WITH DIFFERENT BLOCK SIZES. EACH CELL SHOWS THE MEAN (STANDARD DEVIATION) ERROR IN DEGREES.

W (s)	No filtering	Median filtering	Particle filtering
0.5	10.6 (15.4)	7.1 (8.9)	6.0 (5.4)
1	5.8 (9.0)	5.5 (5.5)	4.3 (4.5)
2	4.7 (4.7)	5.8 (4.7)	4.7 (3.9)
3	6.2 (4.9)	6.2 (4.9)	6.5 (5.0)

TABLE II

LOCALIZATION ERRORS IN THE FOUR SCENARIOS OF THE COMPOSITE ($\mathcal{C}$) AND NATURAL (N) DATASET ($\mathcal{D}$). EACH CELL SHOWS THE MEAN (STANDARD DEVIATION) ERROR IN DEGREES.

$\mathcal{D}$	Scenario	No filtering	Median filtering	Particle filtering
$\mathcal{C}$	$\mathcal{S}1$	3.5 (4.7)	4.6 (4.5)	3.8 (3.9)
	$\mathcal{S}2$	4.3 (7.8)	4.9 (5.7)	4.4 (4.5)
	$\mathcal{S}3$	7.4 (9.0)	9.3 (9.0)	8.2 (7.8)
	$\mathcal{S}4$	8.0 (17.6)	10.9 (13.8)	8.4 (21.2)
N	$\mathcal{S}1$	8.7 (7.5)	9.9 (7.8)	9.1 (7.4)
	$\mathcal{S}2$	8.8 (8.4)	8.7 (6.5)	8.3 (6.5)
	$\mathcal{S}3$	14.7 (18.9)	15.4 (16.0)	10.3 (8.8)
	$\mathcal{S}4$	16.4 (19.5)	16.4 (16.7)	11.5 (9.3)

with large errors only in a few blocks. The tracker further improves the localization accuracy, with PF slightly outperforming MF.

Table I shows the localization error for each tracker with different block sizes. For all trackers the accuracy improves with the block size until $W = 1$ s, slightly changes with

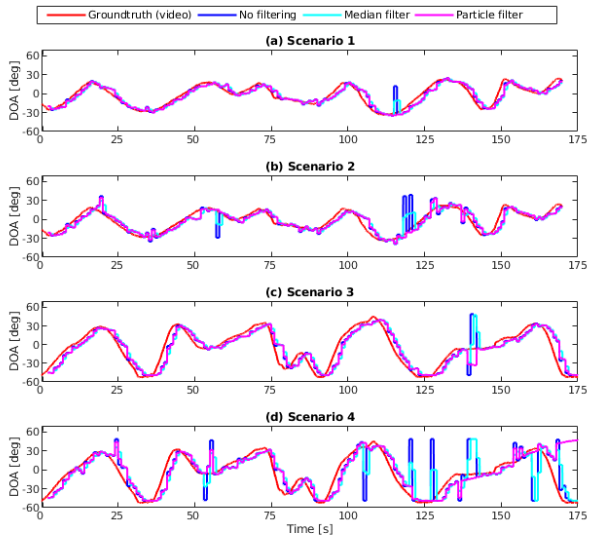


Fig. 8. Tracking results for the four scenarios in the composite dataset.

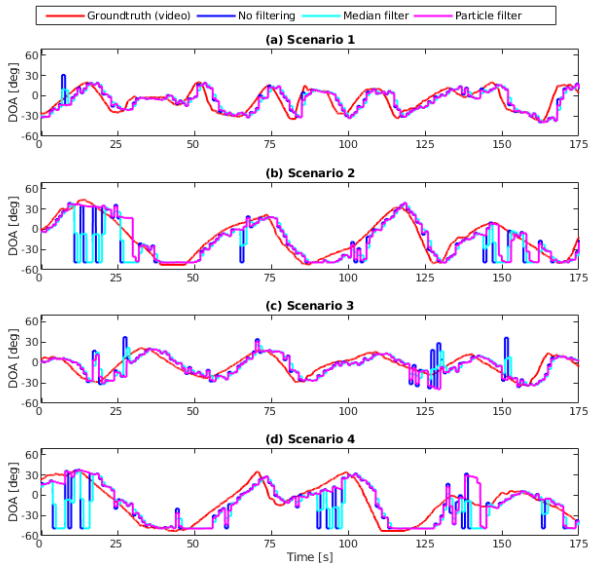


Fig. 9. Tracking results for the four scenarios in the natural dataset.

$W = 2$ s, and then drops with $W = 3$ s. Fig. 7 compares tracking results for different block sizes. The larger the block size, the less noisy the localization results and the smoother the trajectory. However, the larger the block size, the longer the tracking delay, which increases the localization error (see Fig. 7(a) and Fig. 7(d) for $W = 0.5$ s and 3 s, respectively).

Table II shows the localization errors on the four scenarios in the composite and the natural datasets. Fig. 8 and Fig. 9 depict the tracking results for these two datasets. As expected, the trackers perform better in $\mathcal{S}1$ and $\mathcal{S}2$ (loudspeaker moves inside the noiseless sector) than in $\mathcal{S}3$ and $\mathcal{S}4$ (loudspeaker moves freely in front of the drone), because the source localization performance degrades when the speaker moves outside the noiseless sector. The hovering power of the drone does not affect the tracking performance greatly, as shown by the similar performance in $\mathcal{S}1$ and $\mathcal{S}2$,

and in $\mathcal{S}3$ and $\mathcal{S}4$. The composite dataset and natural dataset do not show large differences in localization error.

VI. Conclusion

We tracked a moving sound source from a noisy multi-rotor drone by combining time-frequency filtering, peak detection, and particle filtering. The effectiveness of the proposed method is exemplified with real-recorded experiments with a drone platform and a moving sound source. Future work will extend the algorithm to tracking multiple sound sources in 3D with a flying drone.

References

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

[2] K. Nakadai, M. Kumon, H.G. Okuno, et al. "Development of microphone-array-embedded UAV for search and rescue task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Vancouver, Canada, 2017, pp. 5985-5990.

[3] 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.

[4] J. R. Cauchard, K. Y. Zhai, and J. A. Landay, "Drone and me: an exploration into natural human-drone interaction," in *Proc. 2015 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing*, Osaka, Japan, 2015, pp. 361-365.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] P. Misra, A. A. Kumar, P. Mohapatra, and P. Balamuralidhar, "Aerial drones with location-sensitive ears," *IEEE Communications Mag.* vol. 56, no. 7, pp. 154-160, Jul. 2018.

[11] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K., Yamada, T. Nakamura, and H. Tsujino, "Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 929-932.

[12] 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.

[13] 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.

[14] 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.

[15] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-roto drone," *IEEE Sensor J.*, vol. 18, no. 11, pp. 4570-4582, Jun. 2018.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] K. Hoshiba, K. Nakadai, M. Kumon, and H. G. Okuno, "Assessment of MUSIC-based noise-robust sound source localization with active frequency range filtering," *J. Robotics Mechatronics*, vol. 30, no. 3, pp. 426-435, 2018

[25] T. K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 10, pp. 1623-1636, 2015.

[26] 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, 2016.

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

[28] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174-188, Feb. 2002.

[29] B. N. Vo, S. Singh and A. Doucet, "Sequential Monte Carlo implementation of the PHD filter for multi-target tracking", *Proc. Int. Conf. Info. Fusion*, Queensland, Australia, 2003, pp. 792-799.

[30] 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.