TIME-FREQUENCY PROCESSING FOR SOUND SOURCE LOCALIZATION FROM A MICRO AERIAL VEHICLE

Lin Wang, Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London
{lin.wang; a.cavallaro}@qmul.ac.uk

ABSTRACT

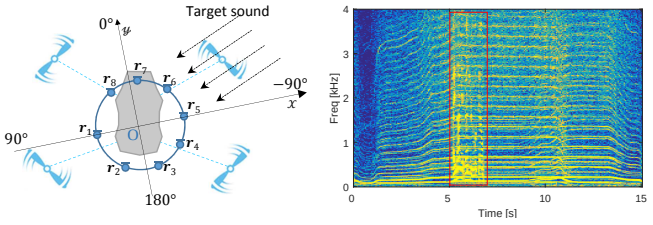
We address the problem of sound source localization with a microphone array mounted on a micro aerial vehicle (MAV). Due to the noise generated by motors and propellers, this scenario is characterized by extremely low signal-to-noise ratios (SNR). Based on the observation that the energy of MAV sound recordings is usually concentrated at isolated time-frequency bins, we propose a time-frequency processing framework to address this problem. We first estimate the direction of arrival of the sound at individual time-frequency bins. Then we formulate a set of spatially informed filters pointing at candidate directions in the search space. The output of the filtering tends to present high non-Gaussianity when the spatial filter is steered towards the target sound source. Finally, by measuring the non-Gaussianity of the spatial filtering outputs we build a spatial likelihood function from which we estimate the direction of the target sound. Experimental results with real-recorded MAV ego-noise show the superiority of the proposed method over the state of the art in performing source localization robustly.

***Index Terms*—** Ego-noise, micro aerial vehicle, microphone array, source localization, time-frequency processing

1. INTRODUCTION

Multirotor micro aerial vehicles (MAV) are increasingly used as mobile sensing platforms equipped with a variety of sensors in a wide range of applications, such as surveillance, broadcasting, and search and rescue [1–8]. When an MAV is equipped with microphones for multichannel recording, sound source localization is an important task that aims to estimate the location of a target sound, such as a human speech or an emergency whistle [8–11]. However, this task is affected by strong ego-noise, which masks the target sound and degrades the recording quality significantly [12]. The motors and propellers, which contribute to the ego-noise, are much closer to the microphones than the emitter of the target sound, thus resulting in extremely low signal-to-noise ratios (SNR), *e.g.* -20 dB [13]. The spectrum of the nonstationary ego-noise depends on the rotation speed of each motor and changes with the varying MAV behavior.

Although microphone array-based source localization has been investigated intensively [14–17], most algorithms are developed for indoor environments with a relatively high SNR (*e.g.*>0 dB). Only a few works have been reported on the challenging MAV-based source localization in extremely low-SNR scenarios (*e.g.*<-15 dB). These works can be classified as unsupervised and supervised. *Unsupervised approaches* perform source localization using microphone signals only. Examples include steered response power with phase transform (SRP-PHAT) [8, 18] and multiple signal classification (MUSIC) [9, 19]. SRP-PHAT, which computes a spatial likelihood map by exploiting the correlation between microphone signals, tends to show degraded performance in MAV-based applications when the target sound is masked by strong ego-noise [8, 18]. MUSIC computes a spatial likelihood map by decomposing the observed signal into orthogonal signal and noise subspaces. The algorithm assumes uncorrelated noise components at the microphones and thus the signal and noise subspaces can be discriminated easily. In MAV-based applications the discrimination becomes difficult since the ego-noise is directional and stronger than the target sound. GEVD-MUSIC (Generalized eigen-value decomposition MUSIC) exploits a noise correlation matrix as additional information to improve robustness to noise. Although several schemes have been proposed to blindly estimate the noise correlation matrix from the microphone signal [9, 19], the estimation is usually inaccurate due to the nonstationarity of the ego-noise. To solve this problem, *supervised approaches* (which need additional sensors to monitor the behavior of the MAV) were proposed to build a noise template database, from which the noise correlation matrix can be estimated corresponding to the motor rotation speed and the MAV behavior [20–22]. The accurate ego-noise estimation enables supervised approaches to improve source localization performance in low-SNR scenarios. However, the need for dedicated monitoring sensors limits the versatility and applicability of supervised approaches.



(a) (b)

Fig. 1. (a) A multirotor MAV with an array of $M = 8$ microphones capturing a target sound. (b) Time-frequency spectrum of the ego-noise and a speech signal (present during 5-7 s and indicated with a red box) recorded by microphone $\boldsymbol{r}_1$ on an operational MAV. The harmonics of the ego-noise and the speech signal generally occupy different time-frequency bins. Our method exploits this time-frequency sparsity.

In this paper we propose an unsupervised approach for sound source localization from a microphone array mounted on an MAV. The ego-noise mainly consists of harmonic components whose energy peaks at isolated harmonic frequencies (Fig. 1). Likewise, the target sound (*e.g.* human speech or emergency whistle) mainly consists of harmonic components. Based on the observation that the harmonics of the ego-noise and the target sound usually occupy

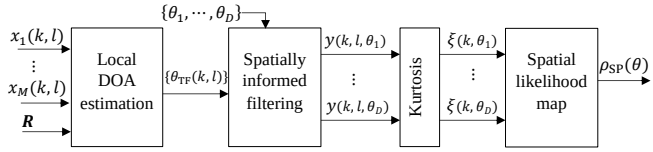


Fig. 2. Block diagram of the proposed method for sound source localization with M microphones mounted on an MAV. For the definition of the variables please see the body of the paper.

different time-frequency bins, *i.e.* at most one source is dominant in one time-frequency bin, we propose a method which exploits the time-frequency sparsity of the MAV sound recording. We first estimate the direction of arrival (DOA) of the sound at each time-frequency bin. Then we formulate a set of spatial filters each pointing at a specific direction. Assuming that the spatial filter corresponding to the target direction can well extract the target sound and that the extracted target sound usually has higher non-Gaussianity (*i.e.* time-frequency sparsity) than the input microphone signal, we build a spatial likelihood function by measuring the non-Gaussianity of the set of spatial filtering outputs.

While the idea of local DOA-based spatial filtering emerged in recent years mainly for indoor speech processing [24, 25], we apply for the first time this technique to MAV-based sound processing and combine it with a non-Gaussianity measure for source localization. With spatial filters suppressing directional ego-noise, the proposed method can obtain improved source localization performance in extremely low-SNR scenarios.

2. PROPOSED METHOD

2.1. Preliminaries

Let a circular array with $M = 8$ microphones be mounted on a multirotor MAV. Let $\boldsymbol{r}_m = [r_{mx}, r_{my}]^{\text{T}}$ be the location of the m-th microphone. The locations of the microphones $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_M]$ are assumed to be known. Let a target sound source located in the far field emit sound with DOA θ_d. The corresponding microphone signal, $\boldsymbol{x}(n) = [x_1(n), \cdots, x_M(n)]^{\text{T}}$, contains the target sound, $\boldsymbol{s}(n) = [s_1(n), \cdots, s_M(n)]^{\text{T}}$, and the ego-noise, $\boldsymbol{v}(n) = [v_1(n), \cdots, v_M(n)]^{\text{T}}$, *i.e.* $\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{v}(n)$. This signal can be written in the short-time Fourier transform (STFT) domain as $\boldsymbol{x}(k,l) = \boldsymbol{s}(k,l) + \boldsymbol{v}(k,l)$, where k and l denote the frequency and frame indices, respectively. Let K and L denote the total number of frequency bins and time frames, respectively.

Since the motors and propellers are closer to the microphones than the target sound source, the MAV sound recording usually presents an extremely-low SNR. We assume a low-reverberant environment without natural wind and that the MAV hovers stably while recording the sound from a static source (*i.e.* the locations of the microphones and the sound source are fixed).

We consider a single target sound from the noisy microphone signal and aim (i) to detect the existence of the target sound and (ii) to estimate its direction. To this end, we compute a spatial likelihood function $\rho(\theta)$ that presents a peak value corresponding to the DOA of the target sound.

In the following, we present the proposed method, which is composed of three main steps (Fig. 2): local DOA estimation, spatially informed filtering and spatial likelihood function computation.

2.2. Local DOA estimation

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 spatial likelihood function [25]

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

where f_k denotes the frequency at the k-th bin; $\tau(m_1,m_2,\theta) = \frac{\|\boldsymbol{r}_{m_2}-\boldsymbol{r}_\theta\|-\|\boldsymbol{r}_{m_1}-\boldsymbol{r}_\theta\|}{c}$ denotes the delay between two microphones m_1 and m_2 with respect to the sound coming from θ, where $\boldsymbol{r}_\theta$ is the location of the far-field sound source with DOA θ and c is the velocity of sound. The term $\text{e}^{j2\pi f_k \tau(m_1,m_2,\theta)}$ represents the inter-channel phase difference, theoretically computed with the delay τ; the term $\frac{x_{m_1}(k,l)x_{m_2}^*(k,l)}{|x_{m_1}(k,l)x_{m_2}(k,l)|}$ represents the inter-channel phase difference measured from x_{m_1} and x_{m_2}, where the superscript '$*$' denotes complex conjugation; the operator $\Re\{\cdot\}$ denotes the real component of the argument. The spatial likelihood γ_{TF} tends to present a high value if these two inter-channel phase differences are consistent with each other. The DOA can thus be estimated as

$$\theta_{\text{TF}}(k,l) = \arg\max_{\theta} \gamma_{\text{TF}}(k,l,\theta). \tag{2}$$

2.3. Spatially informed filtering

The localization results at individual time-frequency bins can be used to construct a spatially informed filter, which extracts the sound coming from a direction θ [24, 25]. To implement this spatial filter, we first detect the time-frequency bins that belong to the desired direction, assuming the DOAs estimated at these time-frequency bins to be Gaussian distributed with mean θ and stand variance σ. The detection is then performed by measuring the closeness of each time-frequency bin to the direction θ:

$$c_d(k,l,\theta) = \exp\left(-\frac{(\theta_{\text{TF}}(k,l)-\theta)^2}{2\sigma^2}\right), \tag{3}$$

where the closeness measure c_d lies in the interval $[0,1]$, with a higher value indicating a higher probability that the (k,l)-th bin belongs to the target sound. Next, we calculate the correlation matrix of the sound signal from the direction θ as

$$\boldsymbol{\Phi}_{ss}(k,l,\theta) = \frac{1}{L}\sum_{l=1}^{L} c_d^2(k,l,\theta)\boldsymbol{x}^{\text{H}}(k,l)\boldsymbol{x}(k,l), \tag{4}$$

where the closeness measure $c_d(k,l,\theta)$ indicates the contribution of the (k,l)-th bin to the correlation matrix, and the superscript $(\cdot)^{\text{H}}$ denotes the Hermitian transpose. Given this estimated target sound correlation matrix, an adaptive beamformer can be formulated easily. We use the multichannel Wiener filter [26]

$$\boldsymbol{w}_{\text{TF}}(k,l,\theta) = \boldsymbol{\Phi}_{xx}^{-1}(k,l)\boldsymbol{\phi}_{ss1}(k,l,\theta), \tag{5}$$

where $\boldsymbol{\phi}_{ss1}(k,l,\theta)$ denotes 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 θ is extracted as

$$y_{\text{TF}}(k,l,\theta) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(k,l,\theta)\boldsymbol{x}(k,l). \tag{6}$$

In this way we can construct multiple spatial filters pointing at a set of predefined candidate directions $\{\theta_1, \cdots, \theta_D\}$ in the search space.

2.4. Spatial likelihood function

We compare the non-Gaussianity of the D spatial filtering outputs. The non-Gaussianity of a sequence can be measured with its statistical kurtosis value, where a higher kurtosis indicates a higher non-Gaussianity [27]. The kurtosis value $\xi(k,\theta)$ is calculated at each frequency bin:

$$\xi(k,\theta) = \mathcal{K}(\tilde{\boldsymbol{y}}_{\text{TF}}(k,\theta)), \tag{7}$$

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. Considering the whole frequency band, we calculate a global spatial likelihood function as

$$\tilde{\rho}_{\text{SP}}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\xi(k,\theta). \tag{8}$$

We normalize (8) into the interval $[0,1]$ by using

$$\rho_{\text{SP}}(\theta) = \mathcal{N}\left(\tilde{\rho}_{\text{SP}}(\theta)\right) = \max\left(\frac{\tilde{\rho}_{\text{SP}}(\theta) - \max(\rho_{\text{TH}}, \text{mean}(\tilde{\rho}_{\text{SP}}))}{\max(\tilde{\rho}_{\text{SP}}) - \min(\tilde{\rho}_{\text{SP}})}, 0\right), \tag{9}$$

where $\min(\cdot)$, $\max(\cdot)$, and $\text{mean}(\cdot)$ denote the minimum, maximum and mean values of the sequence; $\mathcal{N}(\cdot)$ denotes the normalization procedure where ρ_{TH} is a predefined threshold. A sequence with a kurtosis value lower than ρ_{TH} will be detected as noise and hence has its spatial likelihood set to zero. We set $\rho_{\text{TH}} = 5$, which is slightly higher than the kurtosis value of Gaussian noise [27].

3. RESULTS

We compare the proposed spatial filtering-based algorithm (`SP`), with six source localization algorithms, namely the well-known SRP-PHAT algorithm (`SRP`) [14]; three GEVD-MUSIC algorithms: `GMUSIC`, which assumes the noise correlation matrix to be known [19]; `MUSIC`, which uses an identity matrix as the estimate of the noise correlation matrix [19]; `iMUSIC`, which estimates the noise correlation matrix incrementally from the microphone signal [19]; and two additional time-frequency processing-based algorithms we create based on the observations we discussed in this paper: the histogram-based algorithm (`Hist`) and a combined algorithm (`HiSP`). `Hist` builds a histogram using the localization results $\{\theta_{\text{TF}}(\cdot)\}$ at all time-frequency bins [23, 28], *i.e.* $\tilde{\rho}_{\text{Hist}}(\theta) = \mathcal{H}(\{\theta_{\text{TF}}\})$, where $\mathcal{H}(\cdot)$ denotes building a histogram of the values in the sequence. The spatial likelihood function is obtained by normalizing the histogram with $\rho_{\text{Hist}}(\theta) = \mathcal{N}(\tilde{\rho}_{\text{Hist}}(\theta))$. `HiSP` combines `Hist` and `SP` by using $\rho_{\text{HiSP}}(\theta) = \mathcal{N}(\rho_{\text{Hist}}(\theta) + \rho_{\text{SP}}(\theta))$. `SP` performs robustly in low-SNR scenarios, where the kurtosis difference between input and output signals is evident. However, this difference becomes less evident in high-SNR scenarios, decreasing the localization resolution. In contrast, `Hist` has a high localization resolution in high-SNR scenarios but degraded performance in low-SNR scenarios. `HiSP` combines these two complementary algorithms to achieve robust performance for different SNRs.

We evaluate the source localization performance in terms of target sound detection ability, which is measured by the normalized spatial likelihood value at the target direction, *i.e.* $\rho(\theta_d)$. Lying in the interval $[0,1]$, the higher $\rho(\theta_d)$, the more capable the algorithm is to detect the target sound. For this reason, the spatial likelihood function obtained by each algorithm is normalized with (9) by setting $\rho_{\text{TH}} = 0$ (except in `SP` $\rho_{\text{TH}} = 5$). At sampling rate 8 kHz, we set the STFT frame length as 1024 with half overlap. For `SP`, we set $\sigma = 10^\circ$ in (3). The search space $\{\theta_1, \cdots, \theta_D\}$ is set as -180° − 180° with a step of 1°. This search space is also used as the histogram bin for `Hist`. The duration of the test signal is 5 s.

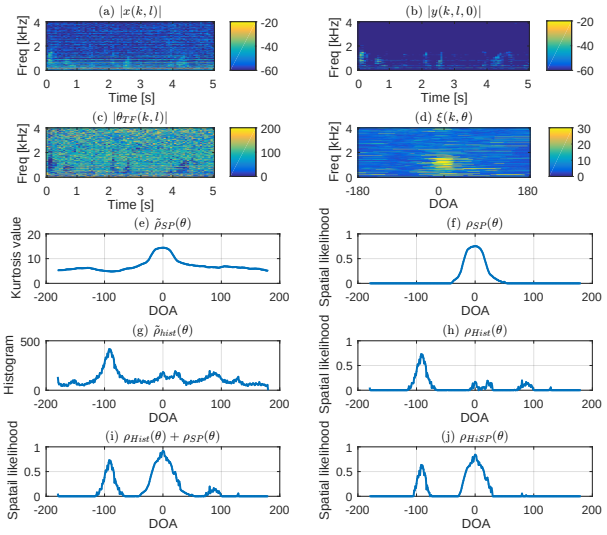


Fig. 3. Intermediate results by `SP`, `Hist` and `HiSP` for a target sound with DOA 0° and input SNR -15 dB. (a)-(b): input and enhanced signals; (c)-(f): results for `SP`; (g)-(h): results for `Hist`; (i)-(j): results for `HiSP`.

We built a hardware prototype with a circular microphone array consisting of eight omnidirectional lavelier microphones fixed above a 3DR IRIS quadcopter [13]. The diameter of the array is 0.2 m. The microphone signal is generated by adding the ego-noise and the target sound at different input SNRs. The ego-noise is recorded in a room with reverberation time of 200 ms [13]. The MAV remains physically static and the motor speed is varied during the recording. The target sound is simulated with the image-source method [29] in a space of size 20m×20m×4m and with reverberation time of 200 ms. A speech source is placed 10 m away and with a varying DOA from -180° to 180°, with an interval of 30°.

Fig. 3 depicts the intermediate processing results by `SP`, `Hist` and `HiSP` for a target sound coming from 0° with input SNR -15 dB. Fig. 3(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. 3(c), performing local DOA estimation can still detect the time-frequency bins that belong to the target sound (*i.e.* at DOA 0°). Fig. 3(d) depicts the kurtosis function $\xi(k,\theta)$, where a high kurtosis value can be observed at DOAs around 0°. Fig. 3(e) depicts $\tilde{\rho}_{\text{SP}}(\theta)$, the kurtosis value averaged across the whole frequency band, and Fig. 3(f) the normalized value $\rho_{\text{SP}}(\theta)$. A peak can be clearly observed around 0°. Fig. 3(g) depicts $\tilde{\rho}_{\text{Hist}}(\theta)$, the histogram of local DOA estimates, and Fig. 3(h) the normalized value $\rho_{\text{Hist}}(\theta)$. A peak can be observed at 0°, which is much weaker than the peak of the ego-noise around -90°. Fig. 3(i) depicts the sum of $\rho_{\text{SP}}(\theta)$ and $\rho_{\text{Hist}}(\theta)$, and Fig. 3(j) the normalized value $\rho_{\text{HiSP}}(\theta)$. The peak of the target sound is also clearly observed. Additionally, Fig. 3(b) depicts the time-frequency spectrum of the spatial filtering output pointing at 0°. The target sound is well extracted while the ego-noise is effectively suppressed.

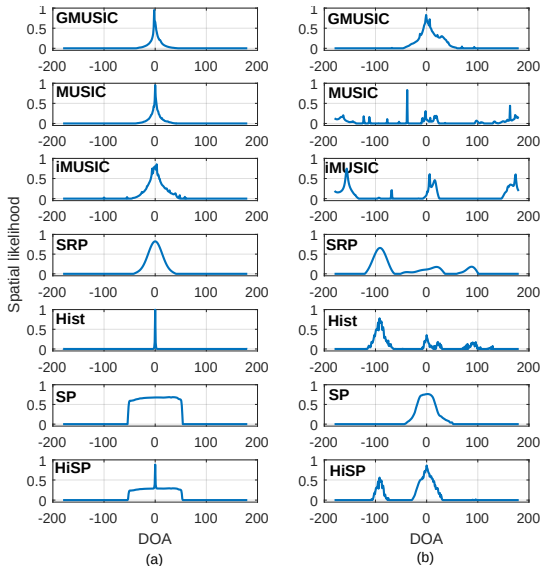


Fig. 4. Spatial likelihood functions obtained by various source localization algorithms for a target sound with DOA 0°: (a) input SNR: ∞; (b) input SNR: -15 dB.

Fig. 4 depicts the spatial likelihood function obtained by the considered algorithms for a target sound coming from 0°, with input SNRs ∞ and -15 dB, respectively. In Fig. 4(a) all the algorithms can estimate the DOA of the target sound correctly when $\text{SNR}_{\text{in}} = \infty$, except that the localization resolution of SP is much lower. In Fig. 4(b) the considered algorithms perform differently for $\text{SNR}_{\text{in}} =$ -15 dB. With prior knowledge of the noise correlation matrix, GMUSIC shows a sole peak at the target direction. In contrast, MUSIC and iMUSIC do not show evident peaks at the target direction, due to inaccurate estimation of the noise correlation matrix. SRP and Hist show strong peaks at ego-noise directions but weak peaks at the target direction. SP shows a sole peak at the target direction. HiSP shows two peaks, including one at the target direction.

Fig. 5(a) depicts the spatial likelihood values at the target direction, $\rho(\theta_d)$, obtained by the considered algorithms when the target sound comes at $\theta_d = 0°$, with a varying input SNR from -25 dB to -5 dB. For all algorithms the obtained $\rho(\theta_d)$ rises with increasing SNR_{in}. GMUSIC performs the best among all the algorithms, achieving a $\rho(\theta_d)$ which is close to 1 for all SNR_{in}. SRP performs the worst. Hist, SP and HiSP evidently outperform MUSIC and iMUSIC. SP obtains a $\rho(\theta_d)$ which is higher than other algorithms, especially when $\text{SNR}_{\text{in}} <$ -10 dB. Hist obtains a $\rho(\theta_d)$ which is close to SP when $\text{SNR}_{\text{in}} \geq$ -10 dB, but drops quickly with decreasing SNR_{in} when $\text{SNR}_{\text{in}} <$ -10 dB. HiSP obtains a $\rho(\theta_d)$ which is between SP and Hist when $\text{SNR}_{\text{in}} <$ -10 dB.

Fig. 6 depicts the spatial likelihood values at the target direction, $\rho(\theta_d)$, obtained by the considered algorithms when varying the DOA of the target sound. We consider two input SNRs: -20 dB and -10 dB. For both scenarios, GMUSIC, SP, and HiSP show almost constant performance for various DOAs. For $\text{SNR}_{\text{in}} =$ -10 dB, Hist shows constant performance for various DOAs. However, its performance degrades significantly for $\text{SNR}_{\text{in}} =$ -10 dB. In addition, Hist shows high spatial likelihood values at DOAs 90° and -90°, the directions where the ego-noise may come from. Hist outperforms SRP especially when $\text{SNR}_{\text{in}} =$ -20 dB.

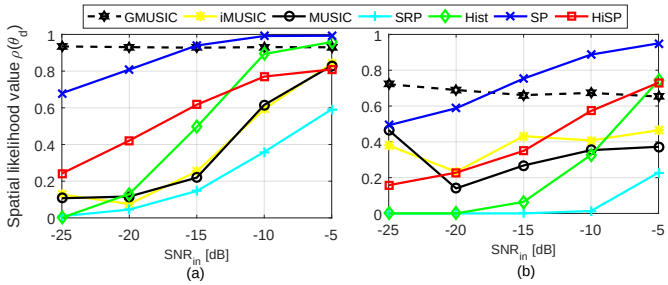


Fig. 5. Spatial likelihood values $\rho(\theta_d)$ at the target direction obtained by various source localization algorithms for a target sound with a varying input SNR. (a) Simulated target sound with DOA $\theta_d = 0°$; (b) Real-recorded target sound with DOA $\theta_d = 160°$.

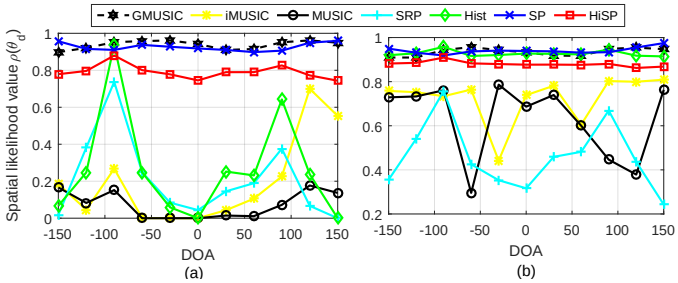


Fig. 6. Spatial likelihood values $\rho(\theta_d)$ at the target direction obtained by various source localization algorithms for a target sound with a varying DOA θ_d. (a) input SNR: -20 dB; (b) input SNR: -10 dB.

Finally, we consider a scenario where the ego-noise and the target sound are recorded separately and added at different input SNRs. A loudspeaker is placed 3 m away from the MAV at the direction 160°, playing speech signals as the target sound [13]. Fig. 5(b) depicts the evaluation results. In comparison to Fig. 5(a), GMUSIC shows degraded performance for real-recorded speech, primarily because the microphones were not calibrated. This also leads to degraded performance of MUSIC and iMUSIC. SRP performs the worst. SP outperforms all the other algorithms except GMUSIC and it is followed by HiSP.

4. CONCLUSION

We proposed a time-frequency processing approach that exploits the time-frequency sparsity of the MAV sound recording for source localization. By estimating the DOA locally at individual time-frequency bins, the proposed spatially informed filter extracts the sound from a desired direction and suppresses other directions effectively. The output signal tends to show a high kurtosis value when the spatial filter corresponds to the target direction. Experimental results demonstrate the advantage of the proposed method in the presence of strong ego-noise.

In our future work we will combine the spatial likelihood value and the mobility of the MAV within a tracker, and investigate the performance of the proposed algorithms in real environments where multiple sound sources as well as natural and motion-induced wind pose additional challenges.

Acknowledgement: This work was supported in part by the ARTEMIS-JU and the UK Technology Strategy Board (Innovate UK) through the COPCAMS Project, under grant 332913.

5. REFERENCES

[1] 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.

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

[3] F. Remondino, L. Barazzetti, F. Nex, M. Scaioni, and D. Sarazzi, "UAV photogrammetry for mapping and 3D modeling - current status and future perspectives," *Int. Archives Photogrammetry, Remote Sensing Spatial Inform. Sci.*, Zurich, Switzerland, 2011, pp. 25-31.

[4] F. Poiesi and A. Cavallaro, "Distributed vision-based flying cameras to film a moving target," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Hamburg, Germany, 2015, pp. 2453-2459.

[5] J. Klapel, *Acoustic Measurements with a Quadcopter: Embedded System Implementations for Recording Audio from Above*, Master Thesis, Norwegian University of Science and Technology, 2014.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. Uemura, O. Sugiyama, R. Kojima, and K. Nakadai, "Outdoor acoustic event identification using sound source separation and deep learning with a quadrotor-embedded microphone array," in *Proc. Int. Conf. Adv. Mechatronics*, Tokyo, Japan, 2015, pp. 329-330.

[11] S. Lana, K. Takahashi, and T. Kinoshita, "Consensus-based sound source localization using a swarm of micro-quadrocopters," in *Proc. Robot. Soc. Japan*, Tokyo, Japan, 2015, pp. 1-4.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] L. Wang, J. D. 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. 1569-1584, Sep. 2016.

[18] 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.

[19] 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.

[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] 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.

[22] 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.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830-1847, Jul. 2004.

[24] 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.

[25] 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.

[26] 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.

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

[28] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193-2206, Oct. 2013.

[29] 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.