**QUIET DRONES**
**Second International e-Symposium**
**on**
**UAV/UAS Noise**
**27th to 30th June 2022**

# Sound source localization and enhancement in 3D space from a flying drone

Lin Wang, Centre for Intelligent Sensing, Queen Mary University of London: lin.wang@qmul.ac.uk
Andrea Cavallaro, Centre for Intelligent Sensing, Queen Mary University of London: a.cavallaro@qmul.ac.uk

## Summary

The ego-noise generated from rotating motors and propellers as well as the movement of the drone impose significant challenges to drone audition, which aims to sense the acoustic environment with onboard microphones mounted on a flying drone. As a state-of-the-art framework for sound processing on drones, time-frequency spatial filtering (TFS) exploits the time-frequency sparsity of the acoustic signals and their correlation at multiple microphones to localize and enhance a target sound in the presence of strong ego-noise. The original TFS framework was proposed with a 2D coordinate system considering azimuth only in the horizontal plane. We extend the TFS framework to a 3D coordinate system for the microphone array considering both azimuth and elevation. We validate the proposed framework with data from a flying drone, and the proposed algorithm significantly outperforms the baseline SRP-PHAT algorithm.

## 1. Introduction

With a drone being able to fly around and hover above a ground terrain, drone audition has found wide applications in search and rescue, aerial filming, monitoring and surveillance, and autonomous human-drone interaction [1-6]. However, acoustic sensing based on the signals captured by airborne microphones is a very challenging task, mainly due to three reasons [7]. First, the rotating motors and propellers generate strong ego-noise that leads to extremely low signal-to-noise ratios (SNR can be lower than -15 dB) at onboard microphones, which are located much closer to motors and propellers than target sound sources around the drone. The ego-noise typically consists of full-band and harmonic components, whose spectrum changes dynamically with the rotating speed of the motors and the flight status of the drone. Second, the wind from the rotating propellers and in the natural environment add a strong noise component and further lower the SNR at onboard microphones. Third, the movement of the drone creates dynamic transmission paths between the target sound sources and onboard microphones, and further increases the challenge of acoustic sensing from the drone.

Microphone arrays have been widely used on ground robots to improve acoustic sensing performance in noisy environments [8]. However, the performance of existing microphone array techniques degrades significantly on drone platforms [9]. In recent years, dedicated methods have been proposed for sound source localization and sound enhancement on drones [11-24]. These methods can be categorized into uni-modal and multi-modal approaches. Uni-modal approaches are based on the microphone signals only [12-16, 19, 21, 27]. To cope with the strong ego-noise, some works optimize microphone array placement and develop algorithms for specific hardware setups [13, 15]. Multi-modal approaches utilize additional sensors to improve acoustic sensing performance. Motor speed sensors can be employed to assist in predicting the ego-noise received at onboard microphones; the prediction is subsequently incorporated into microphone array algorithms for improved robustness to the ego-noise [11, 15]. Onboard cameras can be employed to detect pre-defined sound sources (e.g. human speakers in the application of human-drone interaction) with computer vision algorithms, which are not affected by acoustic noise and thus provides guidance for sound processing [14, 18]. The requirement of additional sensors increases the cost and complexity when applying drone audition in practice.

Time-frequency spatial filtering (TFS) is a recently established framework for sound processing on drones [17-24]. The ego-noise and the target sound (e.g. human speech) typically consist of harmonic components that have concentrated energy at isolated time-frequency bins. Based on this observation, the TFS framework proposed to estimate the directional of arrival (DOA) at each time-frequency bin with the microphone array, based on which a set of spatial filters are formulated to estimate the location of the target sound and to suppress the ego-noise. By exploiting the time-frequency sparsity of the signal (see an example in Fig. 5(b)), TFS effectively improves the acoustic sensing performance in the presence of ego-noise, and achieves state-of-the-art performance for microphone array processing on drones [19, 21]. TFS enables both sound enhancement and sound source localization. The sound enhancement performance was further improved in combination with deep learning [23] and blind source separation [24]. A multi-modal analysis framework was proposed that jointly exploits audio and video to enhance the sounds of multiple targets captured from a drone equipped with a microphone array and a video camera [18]. An audio-visual drone sound recording dataset is made public available to encourage research in the field [22].

A limitation of the current TFS framework is that the algorithm was originally proposed with a 2D circular array and thus works only for a 2D coordinate system considering azimuth only in the plane defined by the array. To encourage a more general application of the algorithm, we extend the TFS framework to a 3D coordinate system that considers both azimuth and elevation. We evaluate the performance of the proposed algorithm with the DREGON dataset [25], which consists of recordings made by a 3D array mounted on a flying quadcopter.

The remaining part of the paper is organized as follows. Sec. 2 formulates the problem. Sec. 3 and Sec. 4 present the time-frequency spatial filtering framework for sound enhancement and sound source localization in 3D space, respectively. Sec. 5 presents experimental results. Finally, we draw conclusions in Sec. 6.

## 2. Problem formulation

Let a microphone array mounted on a quadcopter consist of $I$ microphones arranged in an arbitrary shape. Considering a general 3D coordinate system, the locations of the microphones are denoted as $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_I]$, where $\boldsymbol{r}_m = \left[r_{mx}, r_{my}, r_{mz}\right]^{\mathrm{T}}$ is the position of the $m$-th microphone, and the superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operation. A target sound source in the far field emits sound with a direction of arrival (DOA) $\boldsymbol{\theta}_d = (\alpha_d, \beta_d)$ with respect to the microphone array, where $\alpha_d$ and $\beta_d$ represent the azimuth and elevation, respectively (Fig. 1).

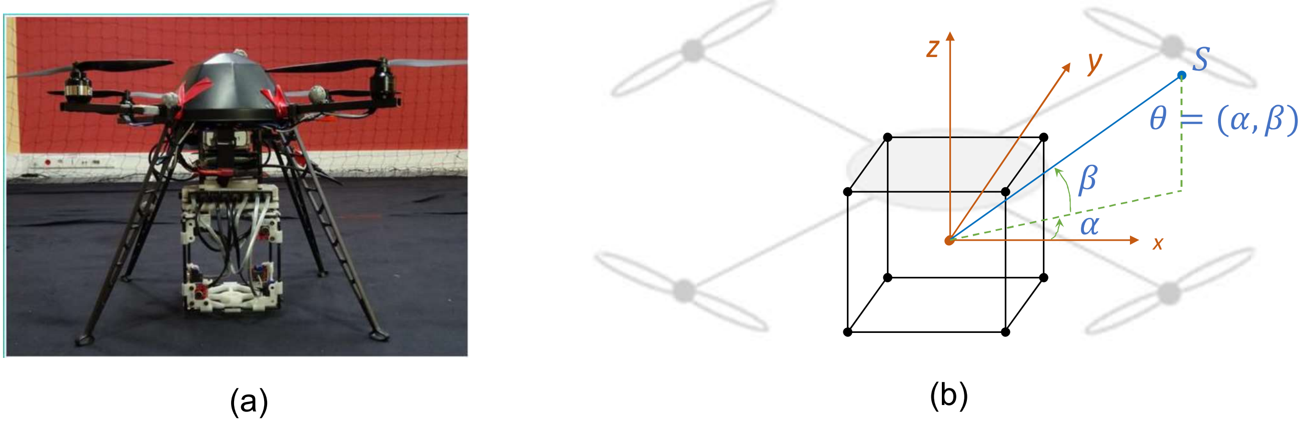

*Figure 1. A microphone array mounted underneath a drone and the 3D coordinate system. (a) Hardware used in the DREGON dataset (image from [25]). (b)The 3D coordinate system.*

The microphone array signal $\boldsymbol{x}(n) = [x_1(n), \cdots, x_I(n)]^{\mathrm{T}}$ consists of the target sound $\boldsymbol{s}(n) = [s_1(n), \cdots, s_I(n)]^{\mathrm{T}}$ and the ego-noise $\boldsymbol{v}(n) = [v_1(n), \cdots, v_I(n)]^{\mathrm{T}}$. This is expressed in the time domain as

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

and in the time-frequency domain as

$$\boldsymbol{X}(k, l) = \boldsymbol{S}(k, l) + \boldsymbol{V}(k, l), \tag{2}$$

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

Given $\boldsymbol{x}(n)$ and $\boldsymbol{R}$, our goal is to estimate the DOA of the target sound $\hat{\theta}_d = (\hat{\alpha}_d, \hat{\beta}_d)$ and to design a spatial filter $\boldsymbol{w}(k, l) = [w_1(k, l), \cdots, w_I(k, l)]^{\mathrm{T}}$ to extract the target sound from the microphone array signal via

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

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

## 3. Time-frequency spatial filtering for Sound enhancement

Given the microphone signal $\boldsymbol{X}(k, l)$, the microphone location $\boldsymbol{R}$, we aim to extract the sound coming from the target direction $\boldsymbol{\theta_d}$. The basic idea of the algorithm is to compute the instantaneous DOA of the sound at each time-frequency bin, which is subsequently utilized to compute the correlation matrix of the target sound and the corresponding spatial filter.

We first estimate the instantaneous DOA of the sound at each time-frequency bin. This is achieved by computing a local spatial likelihood function as

$$\gamma_{TF}(k, l, \boldsymbol{\theta}) = \mathcal{R}\left\{ \sum_{\substack{m_1, m_2 = 1 \\ m_1 \neq m_2}}^{I} \frac{X_{m_1}(k, l) X_{m_2}^*(k, l)}{\left| X_{m_1}(k, l) X_{m_2}(k, l) \right|} e^{j 2\pi f_k \tau(m_1, m_2, \boldsymbol{\theta})} \right\} \tag{4}$$

where $f_k$ denotes the frequency at the $k$-th bin, the superscript $(\cdot)^*$ denotes the complex conjugation, and the operator $\mathcal{R}(\cdot)$ denotes the real component of the argument. The term $\tau(m_1, m_2, \boldsymbol{\theta})$ denotes the delay between two microphones $m_1$ and $m_2$ with respect to the sound coming from a candidate direction $\boldsymbol{\theta} = (\alpha, \beta)$, and can be approximated as

$$\tau(m_1, m_2, \boldsymbol{\theta}) = \frac{|r_{m_1} - r_{\theta}| - |r_{m_2} - r_{\theta}|}{c}, \tag{5}$$

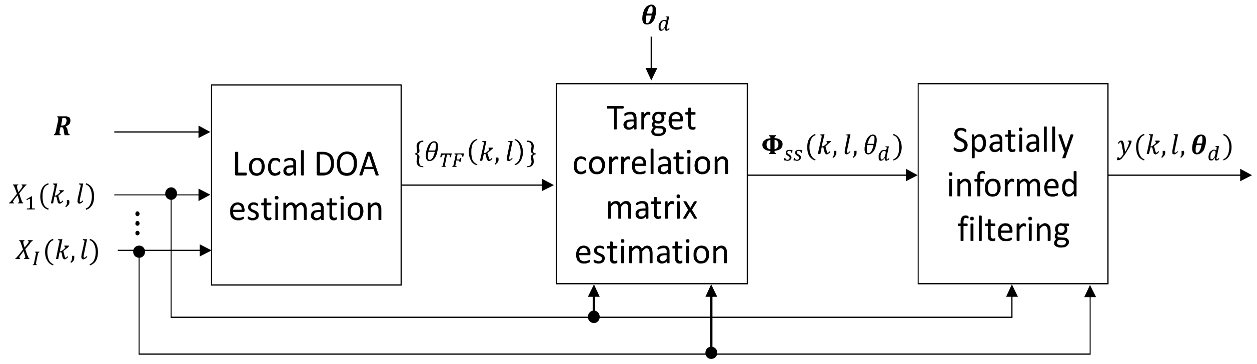
$\boldsymbol{\theta}_d$
$\boldsymbol{R}$
$X_1(k,l)$
$X_I(k,l)$
Local DOA estimation
$\{\theta_{TF}(k,l)\}$
Target correlation matrix estimation
$\boldsymbol{\Phi}_{ss}(k,l,\theta_d)$
Spatially informed filtering
$y(k,l,\boldsymbol{\theta}_d)$

*Figure 2. Time-frequency spatial filtering for sound enhancement, which aims to extract the target sound coming from direction $\boldsymbol{\theta}_d$.*

where $\boldsymbol{r}_{m_1}=[x_{m_1},y_{m_1},z_{m_1}]$ and $\boldsymbol{r}_{m_2}=[x_{m_2},y_{m_2},z_{m_2}]$ denotes the locations of the two microphones; and

$$\boldsymbol{r}_{\boldsymbol{\theta}}=\left[\widetilde{D}\,cos\,(\alpha)\,cos\,(\beta)\,,\ \widetilde{D}\,cos(\alpha)\,sin(\beta)\,,\,\widetilde{D}\,sin(\beta)\right] \tag{6}$$

with $\widetilde{D}\approx 20$ meters representing a sound source in the far field.

The DOA of the sound at each time-frequency bin $\boldsymbol{\theta}_{TF}(k,l)=\{\alpha_{TF}(k,l),\beta_{TF}(k,l)\}$ is then computed as

$$\boldsymbol{\theta}_{TF}(k,l)=\underset{\boldsymbol{\theta}}{argmax}\,\gamma_{TF}(k,l,\boldsymbol{\theta}). \tag{7}$$

Assuming the target sound comes from the direction $\boldsymbol{\theta}_d=(\alpha_d,\beta_d)$, we define a confidence measure to indicate the target sound presence probability at each time-frequency bin, i.e.

$$c_d(k,l,\theta_d)=exp\left(-\frac{|\boldsymbol{\theta}_{TF}(k,l)-\boldsymbol{\theta}_d|}{2\sigma^2}\right), \tag{8}$$

where

$$|\boldsymbol{\theta}_{TF}(k,l)-\boldsymbol{\theta}_d|=\sqrt{(\alpha_{TF}(k,l)-\alpha_d)^2+(\beta_{TF}(k,l)-\beta_d)^2} \tag{9}$$

denotes the distance between $\boldsymbol{\theta}_{TF}$ and $\boldsymbol{\theta}_d$; and $c_d\in[0,1]$. Here we assume the DOA estimate to be Gaussian-distributed with mean $\boldsymbol{\theta}_{\boldsymbol{d}}$ and standard deviation $\sigma$. The higher $c_d$, the closer the local DOA $\boldsymbol{\theta}_{TF}(k,l)$ to the direction $\boldsymbol{\theta}_d$.

Given this confidence measure, we can compute the correlation matrix of the target sound as

$$\boldsymbol{\Phi}_{ss}(k,l,\boldsymbol{\theta}_d)=\frac{1}{L}\sum_{l=1}^{L}c_d(k,l,\boldsymbol{\theta}_d)x^{\mathrm{H}}(k,l)x(k,l), \tag{10}$$

where $c_d$ can be interpreted as the contribution of each time-frequency bin to the target correlation matrix. With this target correlation matrix, we can formulate a spatial filter pointing at direction $\boldsymbol{\theta}_d$. We use a standard Multi-channel Wiener filter (MWF) that is defined as [9]

$$\boldsymbol{w}_{TF}(k,l,\boldsymbol{\theta}_d)=\boldsymbol{\Phi}_{xx}^{-1}(k,l)\Phi_{ss1}(k,l,\boldsymbol{\theta}_d), \tag{11}$$

where $\Phi_{ss1}(k,l,\boldsymbol{\theta}_d)$ is the first column of $\boldsymbol{\Phi}_{ss}(k,l,\boldsymbol{\theta}_d)$, and $\boldsymbol{\Phi}_{xx}(k,l)$ is the correlation matrix of the microphone signal, which can be estimated directly using $\boldsymbol{\Phi}_{xx}(k,l)=\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{x}^{\mathrm{H}}(k,l)\boldsymbol{x}(k,l)$.

Finally, the sound coming from $\boldsymbol{\theta}_d$ is extracted as

$$y_{TF}(k,l,\boldsymbol{\theta}_d)=\boldsymbol{w}^{\mathrm{H}}(k,l,\boldsymbol{\theta}_d)\boldsymbol{x}(k,l). \tag{12}$$

The computation procedure is illustrated in Fig. 2. For sound enhancement, TFS requires to know the target direction $\boldsymbol{\theta}_d$, which can be estimated with the algorithm described in the next section.

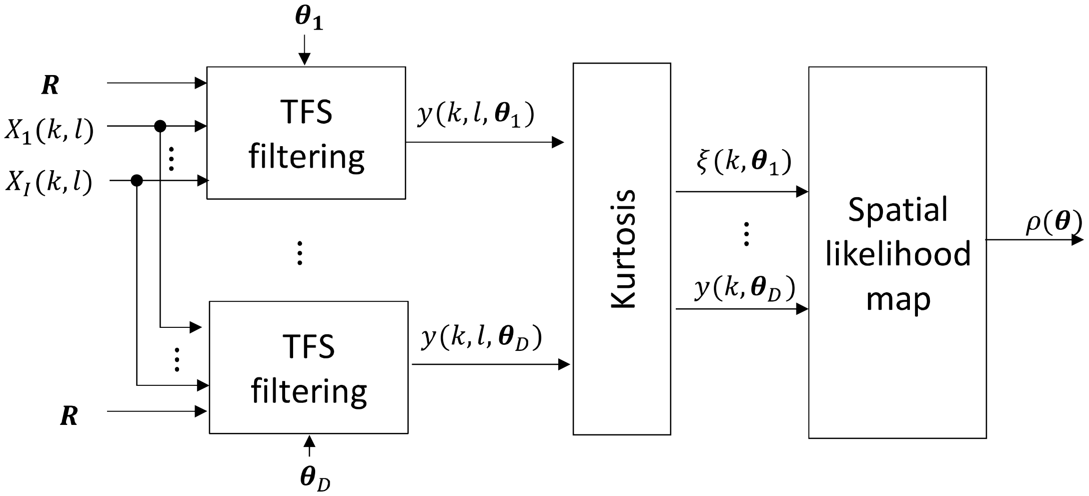

*Figure 3. Time-frequency spatial filtering for sound source localization, which aims to compute a spatial likelihood function $\rho(\boldsymbol{\theta})$ in the search space $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D\}$.*

## 4. Time-frequency spatial filtering for sound source localization

The basic idea of TFS for sound source localization is to formulate a set of spatial filters pointing at candidate directions:

$$\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D\} = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \cdots, (\alpha_D, \beta_D)\}, \tag{13}$$

where $D$ is the total number of candidate directions in a grid search space in azimuth and elevation. We then use the kurtosis of the spatial filtering outputs to indicate the spatial likelihood of the target sound. The target location typically presents a high kurtosis value once the target sound is extracted and the ego-noise is suppressed.

For each candidate direction $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D\}$, we compute a TFS filter and extract the sound coming from the direction $\boldsymbol{\theta}_i$ as

$$y_{TF}(k,l,\boldsymbol{\theta}_i) = \boldsymbol{w}^{\mathrm{H}}(k,l,\boldsymbol{\theta}_i)\boldsymbol{x}(k,l), \tag{14}$$

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

$$\xi(k,\boldsymbol{\theta}_i) = \mathcal{K}(\tilde{\boldsymbol{y}}_{TF}(k,\boldsymbol{\theta}_i)), \tag{15}$$

where $\tilde{\boldsymbol{y}}_{TF}(k,\boldsymbol{\theta}_i)$ denotes the time sequence $|y_{TF}(k,:,\boldsymbol{\theta}_i)|$ and $\mathcal{K}(\cdot)$ denotes the kurtosis value of the sequence. The spatial likelihood of the target sound at $\boldsymbol{\theta}_i$ is represented as the average of the kurtosis value over the whole frequency band, i.e.

$$\rho(\boldsymbol{\theta}_i) = \frac{1}{K}\sum_{k=1}^{K}\xi(k,\boldsymbol{\theta}_i) \tag{16}$$

Repeating this procedure for $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D\}$, we get the spatial likelihood function over the whole search space. The location of the sound source is then estimated as the location with the highest peak, i.e.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_D\}}{\operatorname{argmax}} \{\rho(\boldsymbol{\theta})\}. \tag{17}$$

The whole computation procedure is illustrated in Fig. 3.

## 5. Experimental results

We use the DREGON dataset [25] to validate the performance of the TFS algorithm in 3D scenario. The dataset provides 8-channel recordings made via a cubic microphone array (with side length roughly 10 cm) mounted on the bottom side of a MikroKopter drone, which can fly

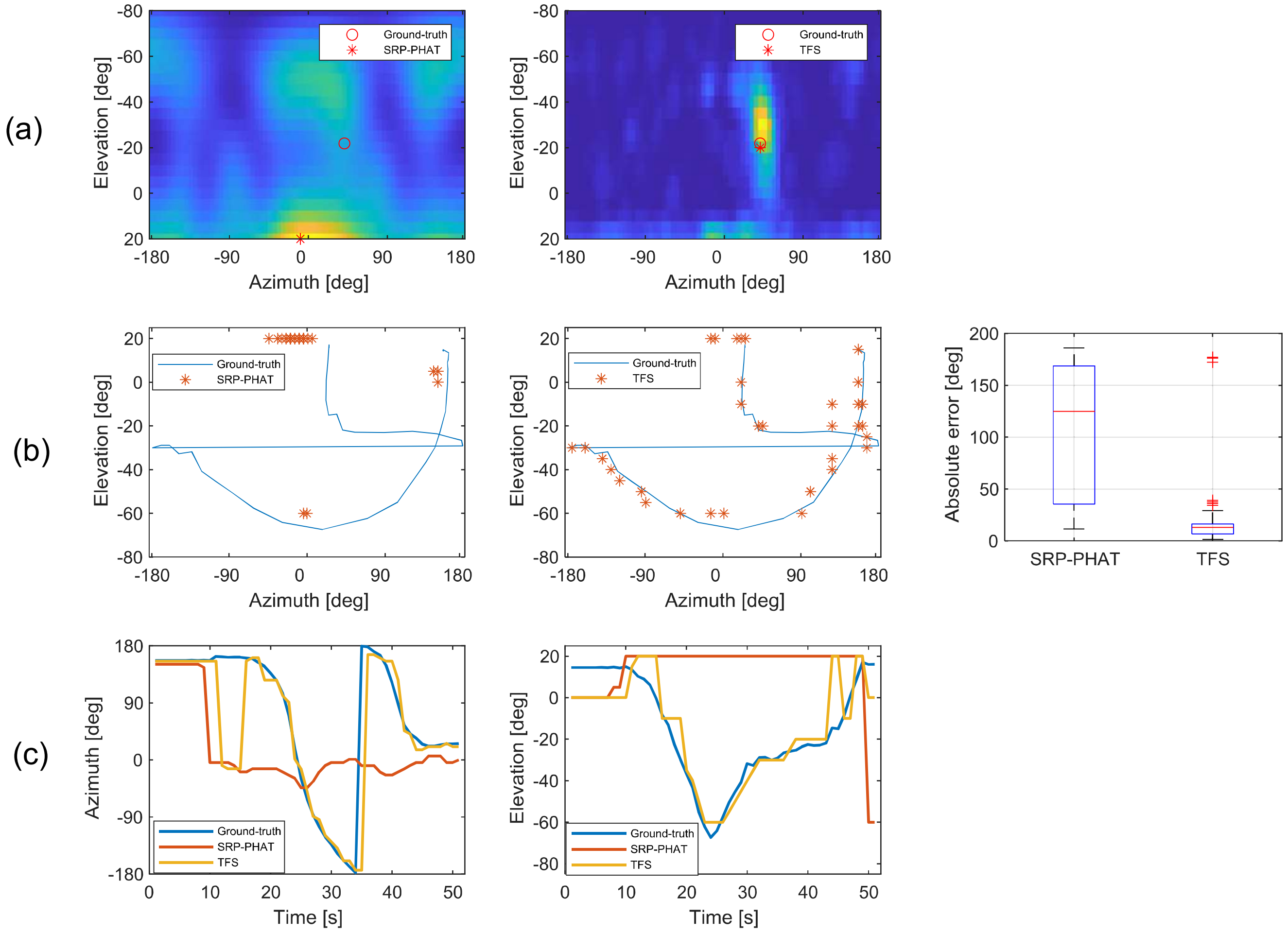

*Figure 4. Sound source localization with SRP-PHAT and TFS. (a) Spatial likelihood map of one sample segment. (b) Scatter plot of the estimation and boxplot of the absolute estimation error. (c) Azimuth and elevation trajectory.*

freely (Fig. 1). A loudspeaker placed on a desk emits speech signals when the drone is flying. The ground-truth location between the sound source and the moving drone was measured with a Vicon motion tracking system. The distance between the drone and the loudspeaker varies between 2 to 4 meters. We use the testing segment “Free Flight – Speech Source at High Volume (Room 1)”. The duration of the recording is about 110 seconds. Based on the description in [25], the SNR of the recording is roughly -12.8 dB.

When applying the TFS algorithm, we set within a space with a grid of $5^\circ$at azimuth $\alpha \in [-179^\circ, 180^\circ]$ and a grid $5^\circ$ at elevation $\beta \in [-90^\circ, 20^\circ]$. This generates 1656 candidate locations in total. We set FFT length 1024 and set $\sigma = 10^\circ$. We employ a block-wise processing scheme to process the signal continuously, i.e. using a processing block of size 2 seconds with half overlap. In this way, we have 51 processing blocks. We apply a medial filter among 3 processing blocks to remove the estimation outliers and to improve the localization accuracy.

We compare with the performance of a baseline algorithm steered response with phase transform (SRP-PHAT) [26], with FFT length 1024. For performance evaluation, we compare the estimated azimuth and elevation with the ground truth, and compute the absolute error as the Euclidean norm of the azimuth and elevation errors.

Fig. 4 shows the sound source localization results by SRP-PHAT and TFS. Fig. 4(a) compares the spatial likelihood map produced by the two algorithms. for one sample segment of 2 seconds. Due to the influence of the ego-noise, SRP-PHAT does not estimate the sound source location correctly. On the other hand, TFS can estimate the sound source location correctly, with a peak clearly observed in the spatial likelihood map. Fig. 4(b) scatterplots the ground-truth location and

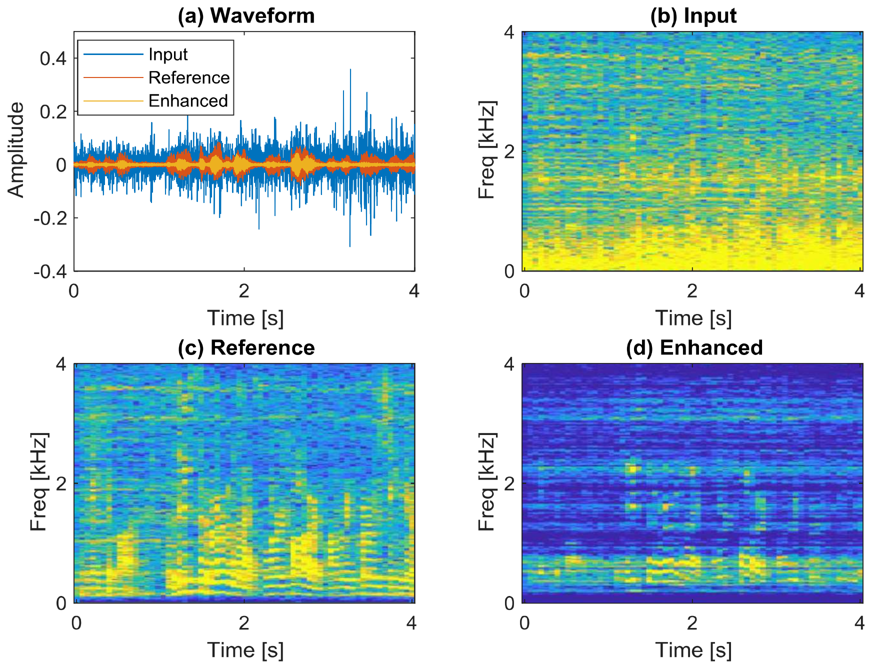

*Figure 5. Sound enhancement results with TFS for a sample segment. (a) Time-domain waveforms of the noisy input, clean reference and enhanced output. (b)-(d) Time-frequency spectrograms of the noisy input, clean reference and enhanced output.*

the estimated locations by the two algorithms in the azimuth-elevation plane. The source locations estimated by SRP-PHAT deviate significantly from the ground-truth while the ones estimated by TFS situate closely to the ground-truth. Fig. 4(b) also boxplots the absolute error across all processing blocks achieved by the two algorithms. SRP-PHAT achieves a median error of $125^\circ$while TFS achieves a median error of $13^\circ$. Finally, Fig. 4(c) visualizes the azimuth trajectory and elevation trajectory estimated by the two algorithms. The azimuth and elevation of the drone vary dynamically during the flight. TFS algorithm can track the trajectory very well in comparison to SRP-PHAT.

Fig. 5 shows sound enhancement results for one sample segment of 4 seconds (19-23th second in the recording). We use the estimated source location at the 21st second (see Fig. 4) as the target location in this segment. For sound enhancement, we choose a processing segment length of 4 seconds and extract the sound from the target location. In Fig. 5, we show the time-domain waveforms and the time-frequency spectrogram of the noisy input, the clean reference (which was provided by an external camera capturing the whole scene), and the enhanced output by TFS. From the spectrogram in Fig. 5(b), the noisy input contains the ego-noise, which consists of full-band and harmonica components, the wind noise, which dominates the low frequency band, and the speech component, which is hardly distinguished. From Fig. 5(d), the speech component is clearly observed after TFS enhancement, although with certain distortion in comparison with the clean reference in Fig. 5(c).

## 6. Conclusions

We presented a time-frequency spatial filtering framework for sound processing on drones. We extended the TFS filtering framework [19, 21] to a more general 3D scenario and validated the performance with a public dataset for sound source localization from a flying drone. Future work includes optimizing the algorithm for real-time computation.

### Acknowledgement

The work is partly supported by the Pilot Research Fund of UK Acoustics Networks (UKAN+).

## References

[1] 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.

[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] Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, “Audio-based search and rescue with a drone: highlights from the IEEE signal processing cup 2019 student competition,” *IEEE Signal Process. Mag.,* vol. 36, no. 5, pp. 138-144, Sep. 2019.

[4] 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.

[5] F. G. Serrenho, J. A. Apolinario, A. L. L. Ramos, and R. P. Fernandes, “Gunshot airborne surveillance with rotary wing UAV embedded microphone array,” *Sensors*, vol.19, no. 4271, pp. 1-26, 2019.

[6] Michez, S. Broset, and P. Lejeune, “Ears in the sky: Potential of drones for the bioacoustic monitoring of birds and bats,” *Drones*, vol. 5, no. 9, pp. 1-19, 2021.

[7] 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 Surv.*, Colorado Springs, USA, 2016, pp. 1-7.

[8] 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.

[9] 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.

[10] A. Schmidt, H. W. Lollmann, and W. Kellermann, “Acoustic self-awareness of autonomous systems in a world of sounds,” *Proceedings of IEEE*, vol. 108, no. 7, pp. 1127-1149, Jul. 2020.

[11] 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.

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

[13] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, “Acoustic source localization from multirotor UAVs,” *IEEE Trans. Industrial Electronics*, vol. 67, no. 10, pp. 8618-8628, 2019.

[14] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, “Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Las Vegas, USA, 2020, pp. 4848-4854.

[15] B. Yen, Y. Hioka, G. Schmid, and B. Mace, “Multi-sensory sound source enhancement for unmanned aerial vehicle recordings,” *Applied Acoustics*, vol. 189, no. 108590, pp. 1-22, 2022.

[16] W. N. Manamperi, T. D. Abhayapala, J. A. Zhang, and P. Samarasinghe, “Drone audition: Sound source localization using onboard microphones,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508-519, 2022.

[17] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, New Orleans, USA, 2017, pp. 496-500.

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

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

[20] L. Wang, R. S. Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Madrid, Spain, 2018, pp. 2511-2516.

[21] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570-4582, Jun. 2018.

[22] L. Wang, R. S. Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Macau, China, 2019, pp. 5320-5325.

[23] L. Wang and A. Cavallaro, "Deep learning assisted time-frequency processing for speech enhancement on drones," *IEEE Trans. Emerging Topics in Computational Intelligence*, vol. 5, no. 6, pp. 871-881, Dec. 2021.

[24] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 58, pp. 2523-2537, Aug. 2020.

[25] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Madrid, Spain, 2018, pp. 5735-5742.

[26] L. Wang, T. 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.

[27] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," Applied Acoustics, vol. 155, pp.423-427, 2019.