CIS centre for intelligent sensing

Queen Mary University of London

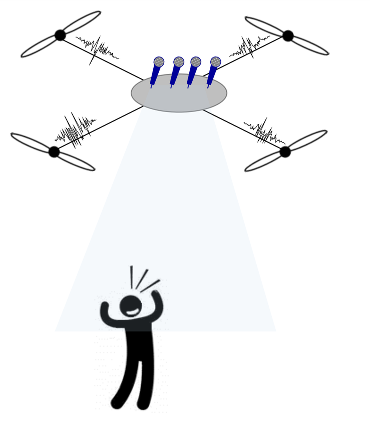
Tracking a moving sound source from a multi-rotor drone

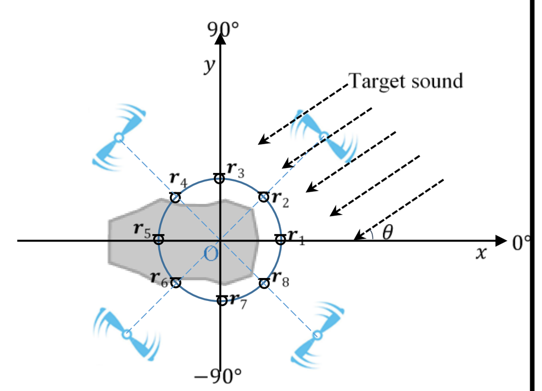
Lin Wang, Ricardo Sanchez Matilla, Andrea Cavallaro

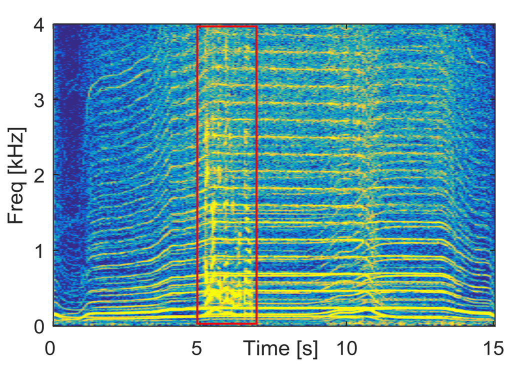
{lin.wang, ricardo.sanchezmatilla, a.cavallaro}@qmul.ac.uk

1. Introduction

- Sound source tracking with a microphone array mounted on a mini-drone
  - human robot interaction
  - surveillance, serach & rescue
- Signal model
  - M microphones: $\boldsymbol{x}(n) = \boldsymbol{s}(n) + \boldsymbol{v}(n)$
  - microphone location: $\boldsymbol{R}$
  - direction of arrival (DOA) of the target sound: $\theta_d(n)$
    - θ_d is time-varying and unknown
  - objective: to estimate and track $\theta_d(n)$ given $\boldsymbol{x}$ and $\boldsymbol{R}$
- Main challenge [3]
  - ego-noise generated by rotating rotors and propellers
  - extremely low signal-to-noise ratio (e.g. SNR < -15 dB)
  - dynamics due to time-varying DOA
- Solution
  - block-wise processing
  - time-frequency spatial filtering for DOA estimation
  - particle filtering to smooth the estimation

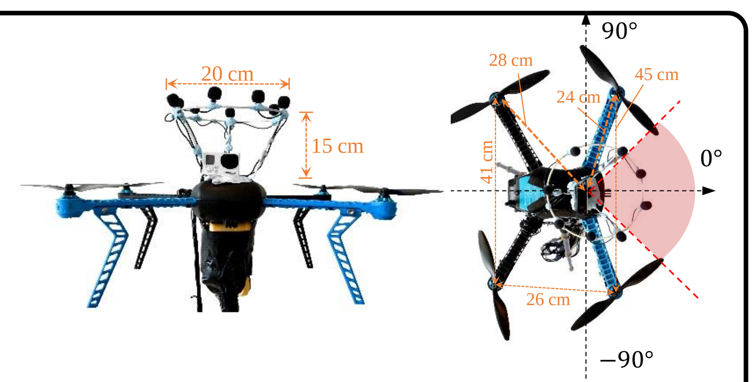


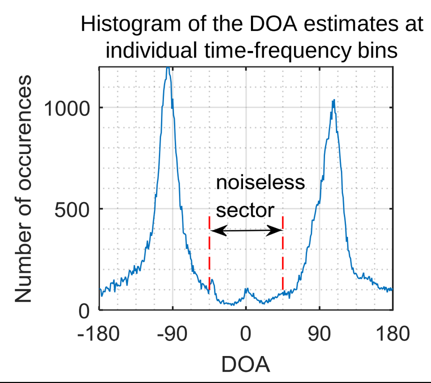




2. Hardware prototype

- Auditory drone [1]
  - 3DR IRIS Quadcopter
  - Circular microphone array
    - $M = 8$ elements with diameter 0.2m
- Spatial characteristics of ego-noise [4]
  - the location of the rotors and propellers are fixed
  - ego-noise tends to arrive from the back side
  - noisy sector vs noiseless sector ($\boldsymbol{\theta_L} = [-45°, 45°]$)
- This noiseless sector will be exploited in the proposed method (peak detection and tracking)





3. Time-frequency spatial filtering for source localization

- Based on the time-frequency sparsity of ego-noise and target sound [2]
- Workflow
  - to estimate the DOA at individual time-frequency bins
  - to formulate a set of spatially informed filters pointing at candidate directions
  - the output signal tends to show a high Kurtosis value when the spatial filter points at the target direction

$\theta \in [\theta_1, \cdots, \theta_D]$: candidate directions
$\boldsymbol{x}^{\boldsymbol{b}}(k,l) = [x_1(k,l), \cdots, x_M(k,l)]^{\mathrm{T}}$: the STFT signal in the b-th block
k, l: time and frequency indices; b: block index (neglected for clarity)

$x_1(k,l)$ ⋯ $x_M(k,l)$ $\boldsymbol{R}$

Local DOA estimation
γ_{TF}: SRP-PHAT function at (k,l)
$\theta_{\mathrm{TF}}(k,l) = \underset{\theta}{\operatorname{argmax}}\, \gamma_{\mathrm{TF}}(k,l,\theta)$

$\{\theta_{\mathrm{TF}}(k,l)\}$

$\{\theta_1, \cdots, \theta_D\}$

Spatially informed filtering
$c_d(k,l,\theta) = exp\left(-\frac{(\theta_{TF}(k,l)-\theta)^2}{2\sigma^2}\right)$
$\boldsymbol{\Phi}_{ss}(k,l,\theta) = \frac{1}{L}\sum_l c_d^2(k,l,\theta)\boldsymbol{x}^H(k,l)\boldsymbol{x}(k,l)$
$\boldsymbol{\Phi}_{xx}(k,l) = \frac{1}{L}\sum_l \boldsymbol{x}^{\mathrm{H}}(k,l)\boldsymbol{x}(k,l)$
$\boldsymbol{w}(k,l,\theta) = \boldsymbol{\Phi}_{xx}^{-1}(k,l)\boldsymbol{\phi}_{ss}^{1}(k,l,\theta)$
$y(k,l,\theta) = \boldsymbol{w}^{\mathrm{H}}(k,l,\theta)\boldsymbol{x}(k,l)$

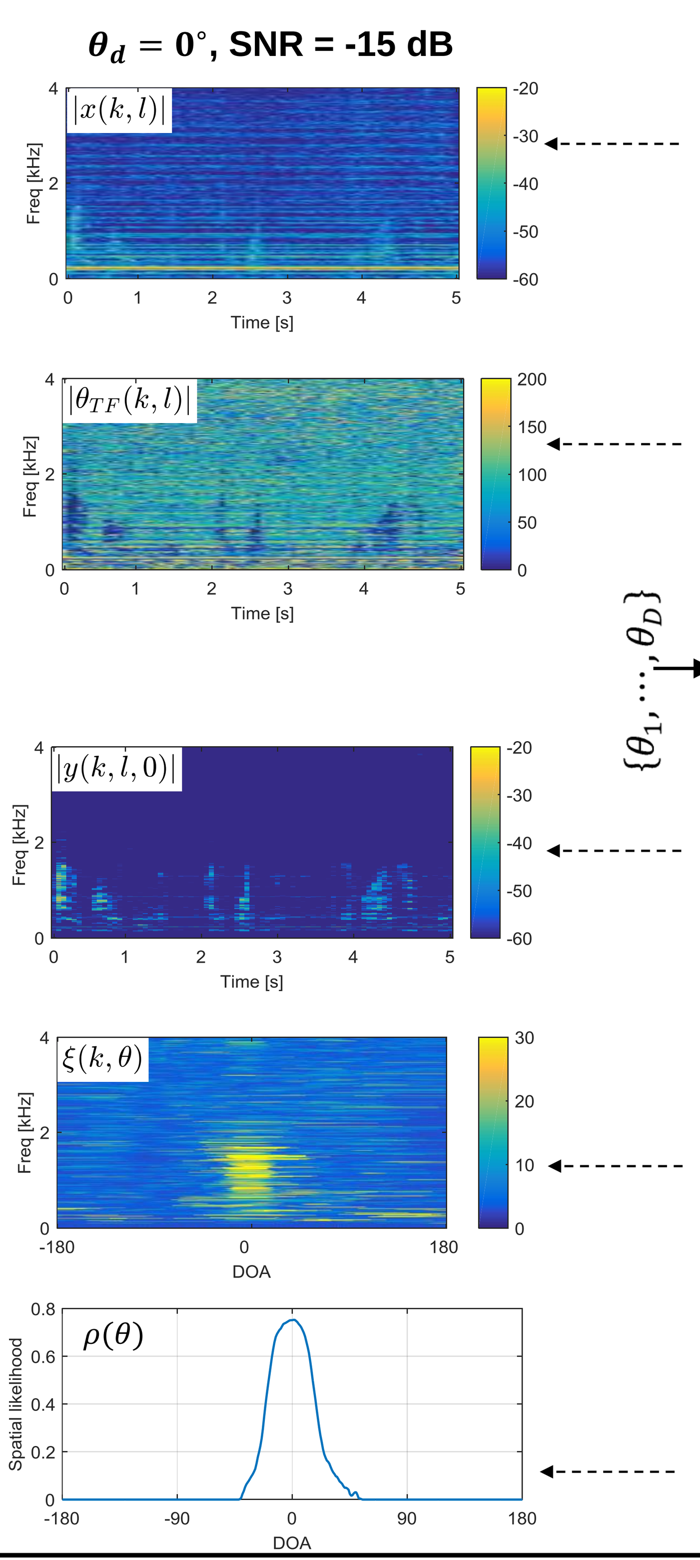
$y(k,l,\theta_1)$ ⋯ $y(k,l,\theta_D)$

Kurtosis
$\xi(k,\theta) = \mathrm{Kurtosis}(y(k,:,\theta))$

$\xi(k,\theta_1)$ ⋯ $\xi(k,\theta_D)$

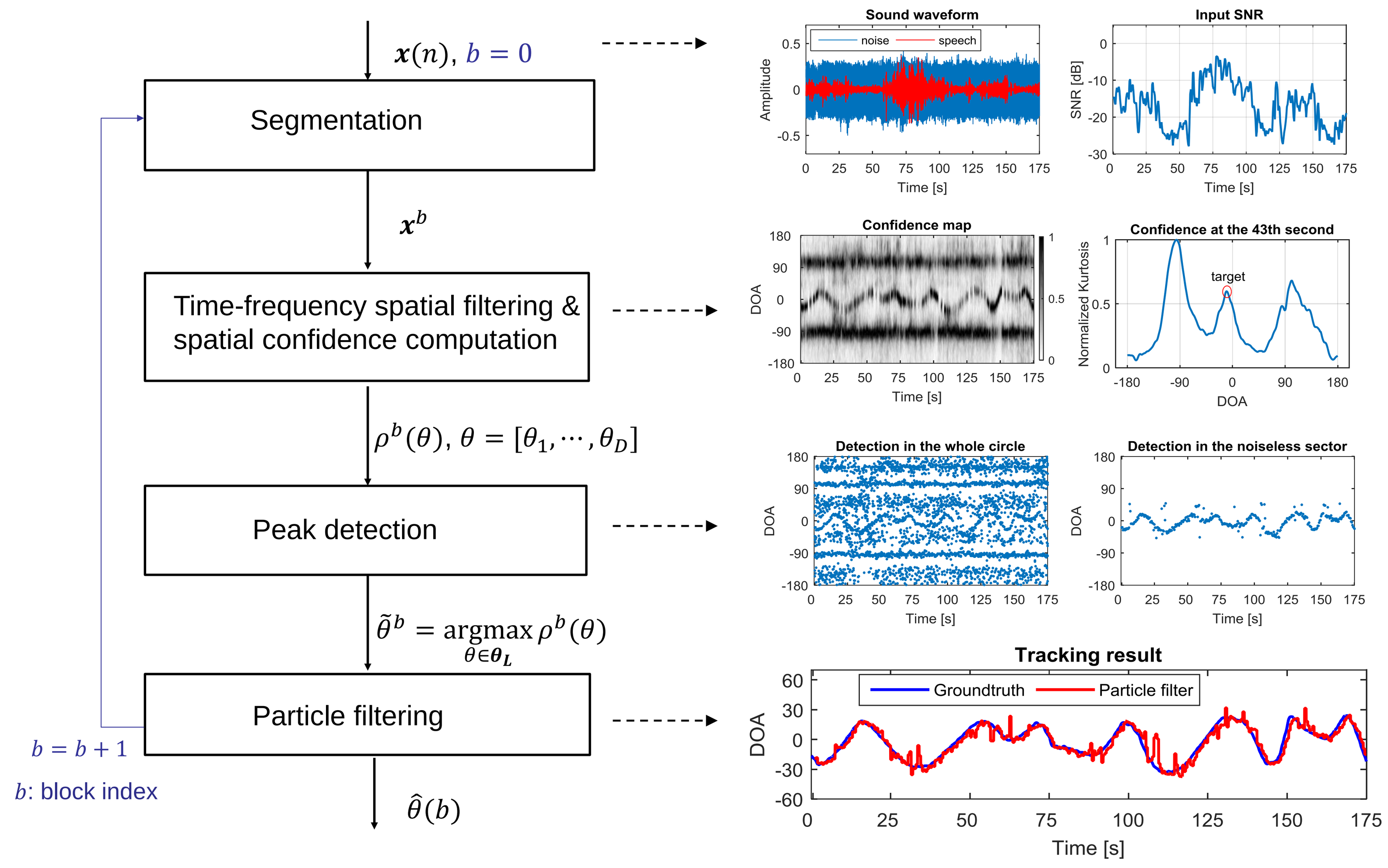
Spatial likelihood map
$\tilde{\rho}(\theta) = \sum_k \xi(k,\theta)$
$\rho(\theta) = \mathrm{Normalize}(\tilde{\rho}(\theta))$

$\rho(\theta)$



4. Tracking framework

- Peak detection in the noiseless sector
  - assuming speakers always in front of the drone

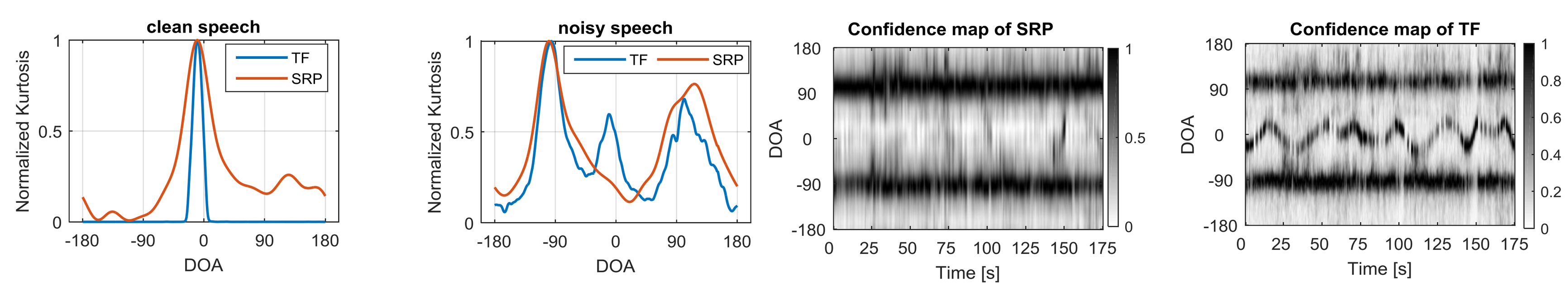


5. Experimental results

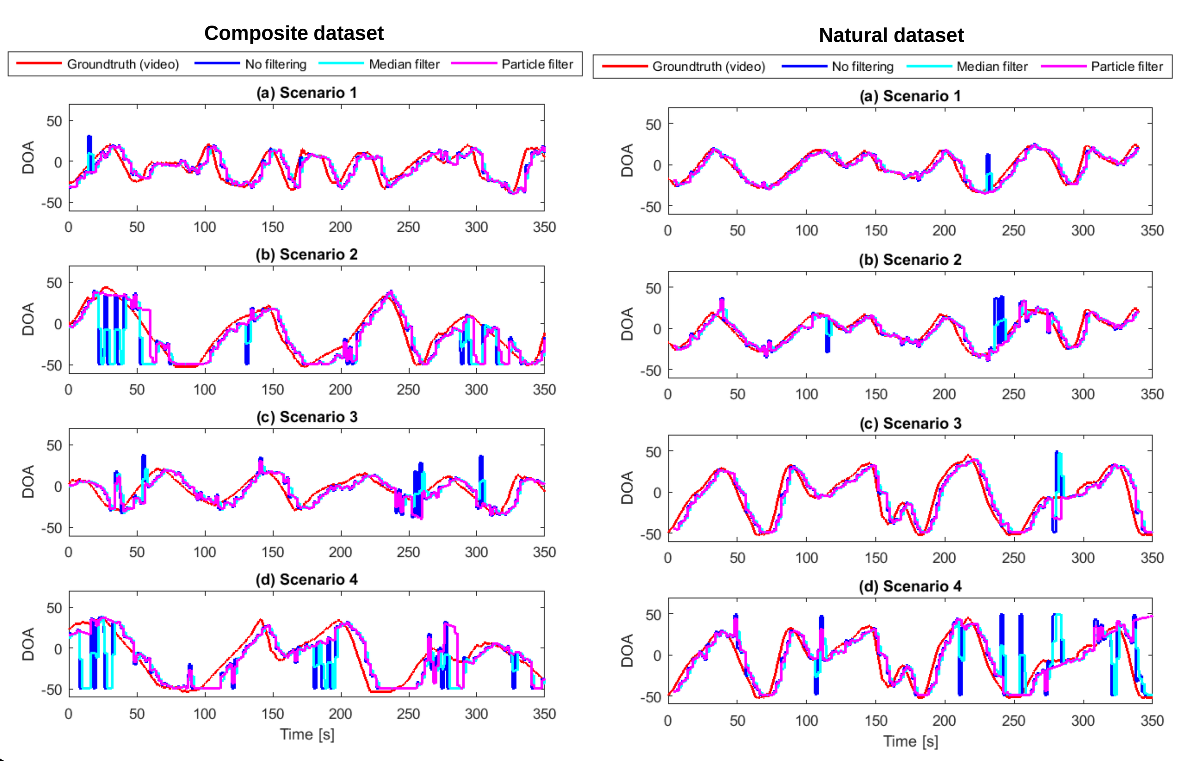
- Setup
  - open space: auditory drone on tripod
    - operating with stationary and dynamic power
  - loudspeaker moving: constrained (noiseless sector) or unconstrained area
  - camera on the drone to capture the moving loudspeaker: video groundtruth [5]
- Dataset
  - composite (ego-noise and speech recorded separately)
  - natural (ego-noise and speech recorded simultaneously)
  - four scenarios:

Loudspeaker \ Drone	Stationary	Dynamic
Constrained	S1	S2
Unconstrained	S3	S4

- Evaluation measure
  - localization error in comparison with video ground-truth
- Source localization result comparison
  - SRP-PHAT and time-frequency spatial filtering (TF) [2]



- Tracking result comparison
  - no filtering, median filtering, particle filtering
  - block size: 2 seconds with half overlap



Localization error [degree]: mean (standard deviation)

	Scenario	No filtering	Median filtering	Particle filtering
Composite	S_1	3.5 (4.7)	4.6 (5.5)	3.8 (3.9)
	S_2	4.3 (7.8)	4.9 (5.7)	4.4 (4.5)
	S_3	7.4 (9.0)	9.3 (9.0)	8.2 (7.8)
	S_4	8.0 (17.6)	10.9 (13.8)	8.4 (21.1)
Natural	S_1	8.7 (7.5)	9.9 (7.8)	9.1 (7.4)
	S_2	8.8 (8.4)	8.7 (6.5)	8.3 (6.5)
	S_3	14.7 (18.9)	15.4 (16.0)	10.3 (8.8)
	S_4	16.4 (19.5)	16.4 (16.7)	11.5 (9.3)

6. Conclusions

- Moving sound source tracking by combining:
  - time-frequency spatial filtering
  - peak detection in the noiseless sector
  - particle filtering
- Verified with real data and video groundtruth

References

[1] L. Wang and A. Cavallaro, “Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles,” *Proc. AVSS*, 2016.

[2] L. Wang and A. Cavallaro, “Time-frequency processing for sound source localization from a micro aerial vehicle,” *Proc. ICASSP*, 2017.

[3] L. Wang and A. Cavallaro, “Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles,” *IEEE Sensors. J.*, 2017.

[4] L. Wang and A. Cavallaro, “Acoustic sensing from a multi-rotor drone”, *IEEE Sensors. J.*, 2018.

[5] 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*, 2017.

Audio Demos & Datasets

http://www.eecs.qmul.ac.uk/~andrea/sst.html
http://www.eecs.qmul.ac.uk/~andrea/ear-in-the-sky.html
http://cis.eecs.qmul.ac.uk/projects/multimodalmav/
http://www.eecs.qmul.ac.uk/~andrea/auditory-mav.html

Acknowledgment

This work is supported by the ARTEMIS-JU and the UK Technology Strategy Board (Innovative UK) through the COPCAMS Project, under grant 332913; and also supported by the UK EPSRC project NCNR (EP/R02572X/1).