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

CIS centre for intelligent sensing

Queen Mary University of London

Ricardo Sanchez-Matilla, Lin Wang, Andrea Cavallaro

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

Introduction

Objective

- To enable a micro aerial vehicle (MAV) to hear multiple sound sources (e.g. talking people)

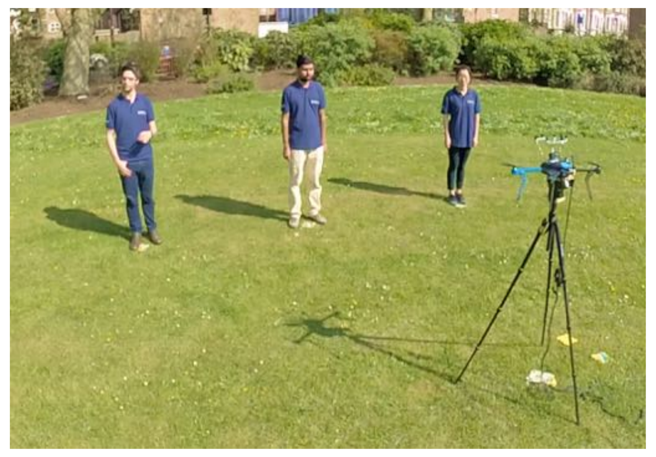
Challenges

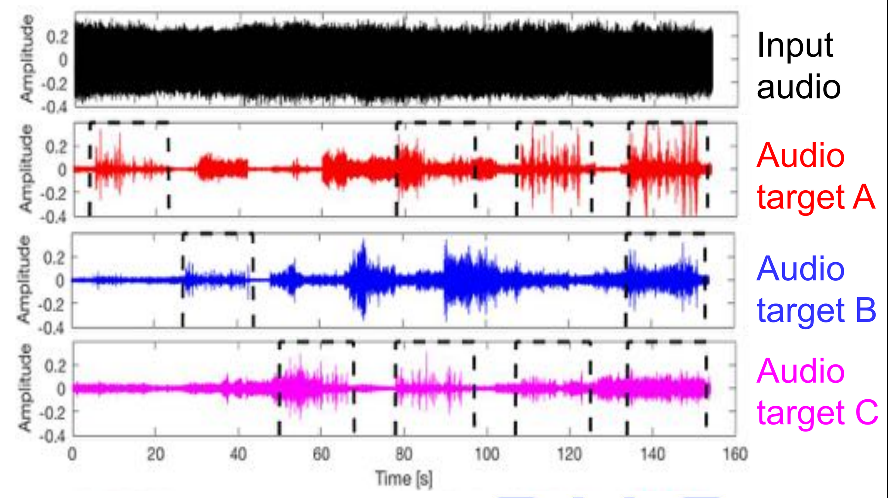
- Extremely low signal-to-noise ratio [1-3]
- Multiple sound sources at unknown locations

Proposed audio-visual solution

- Video: to locate candidate sound sources
- Audio: to enhance each sound source

Demo





Up to 3 people moving and speaking simultaneously

http://cis.eecs.qmul.ac.uk/projects/multimodalmav/

Proposed method

Framework

- Audio-visual calibration
- Visual target detection and tracking
- Spatially informed time-frequency audio filtering

Given

- $\boldsymbol{I}_k$: video signal
- $\{x_1(n)\dots x_M(n)\}$: microphone-array signals
- $\boldsymbol{O}_C, \boldsymbol{O}_M$: camera and microphone-array origins
- $\boldsymbol{R}$: location of microphones

To estimate

- N : number of speakers
- $\{\theta_v^1 \dots \theta_v^i \dots \theta_v^N\}$: direction of arrival (DOA) of the potential sound sources on video coordinates
- $\{\theta_a^1 \dots \theta_a^i \dots \theta_a^N\}$: direction of arrival (DOA) of the potential sound sources on audio coordinates
- $\{y_1(n) \dots y_N(n)\}$: enhanced speeches

Assumption

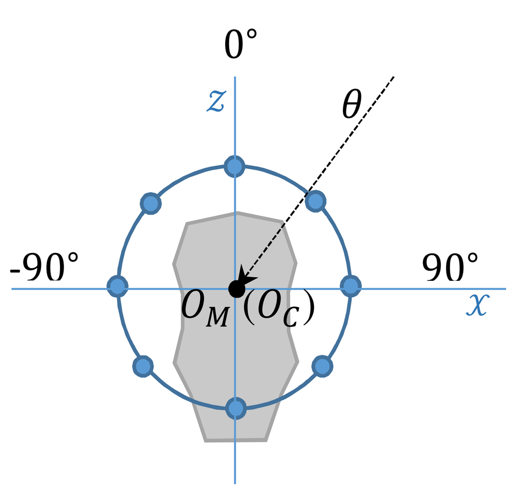
- Hovering MAV and static active speakers

Hardware prototype

- 3DR IRIS Quadcopter: 0.55m×0.55m
- GoPro HERO3+
- 8-microphone array: diameter 0.2m

Geometrical alignment

- Calculate geometrical relation between audio and video reference systems using a calibration sequence (i.e. a sequence with a clean sound easy to localize via audio and video)
- To estimate parameters (a_1, a_2) that linearly relate both signals

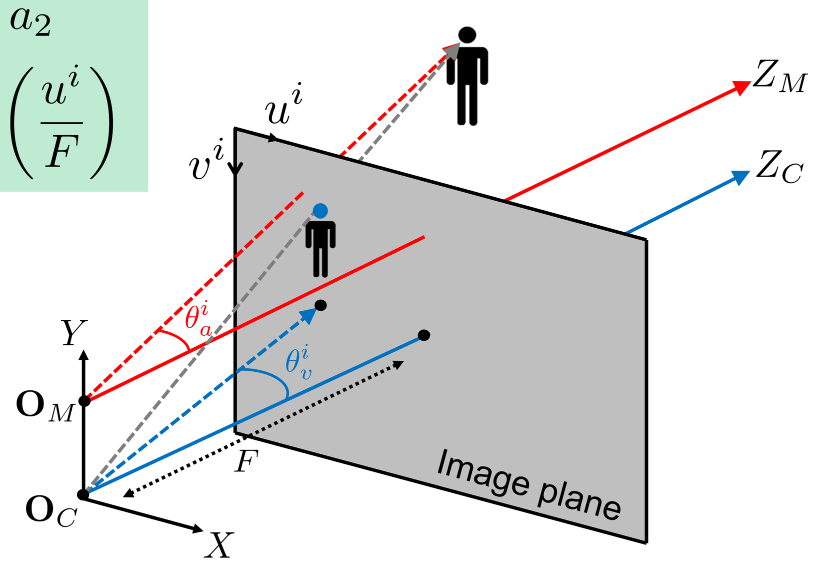


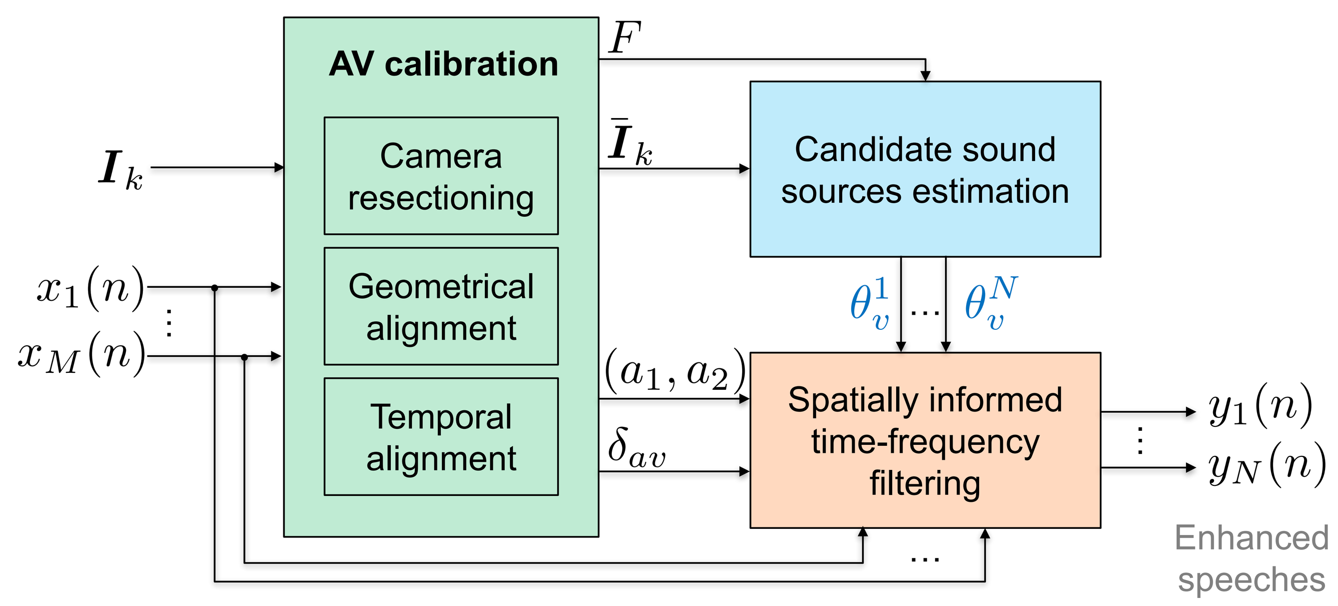
Top-down 2D geometrical representation

$$\theta_a^i = a_1 \theta_v^i + a_2$$

$$\theta_v^i = \arctan\left(\frac{u^i}{F}\right)$$

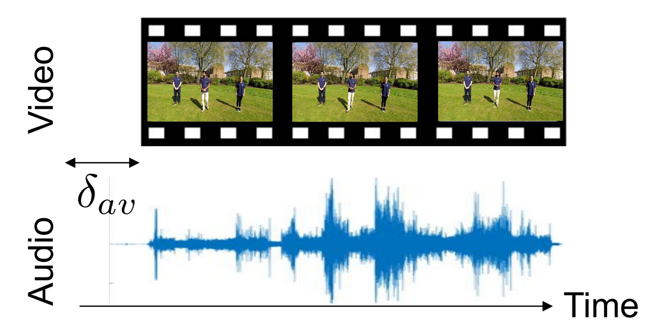
F : focal length





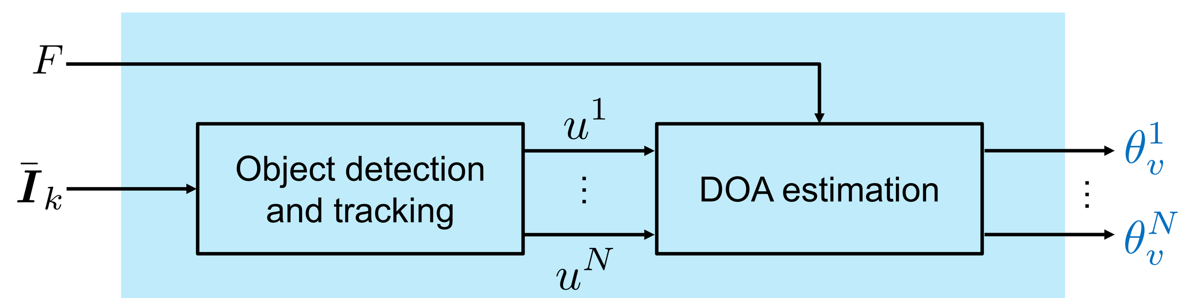
Temporal alignment

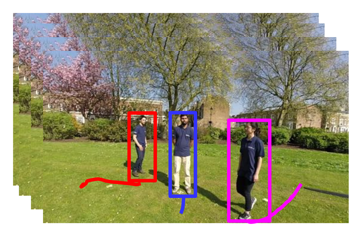
- to detect the time offset between the microphone array and the GoPro audio signals (δ_{av})



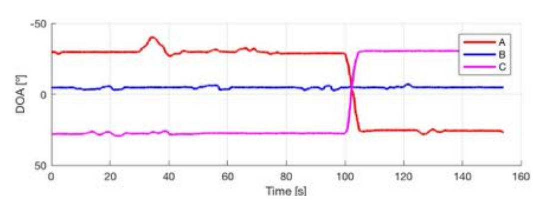
Candidate sound sources estimation

- to track the DOA (θ_v^i) of each target using EA-PHD-PF framework [4]





Trajectory for target A, B and C



DOA for target A, B and C

Time-frequency filters [2]

- to separate and enhance each target speech (i.e. for each DOA, θ_a^i)

Adaptive filter — Microphone correlation matrix — Target sound correlation matrix

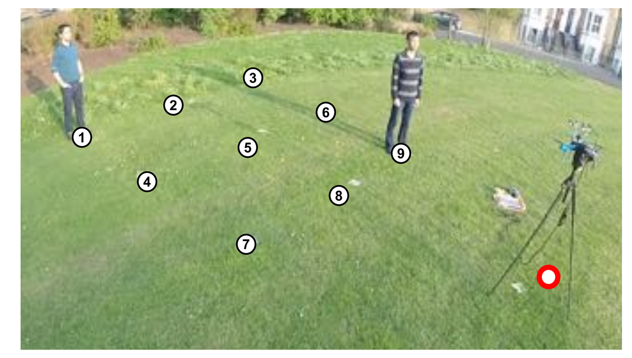
$$\boldsymbol{w}_{\text{TF}}(\omega, l, \theta_a^i) = \boldsymbol{\Phi}_{xx}^{-1}(\omega, l)\,\boldsymbol{\phi}_{ss1}(\omega, l, \theta_a^i)$$

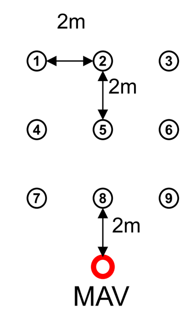
$$y_{\text{TF}}(\omega, l, \theta_a^i) = \boldsymbol{w}_{\text{TF}}^{\text{H}}(\omega, l, \theta_a^i)\,\boldsymbol{x}(\omega, l)$$

Experimental results

Time-frequency spatial filter for comparison

- Audio-only [3]
- Audio-visual (**proposed**)





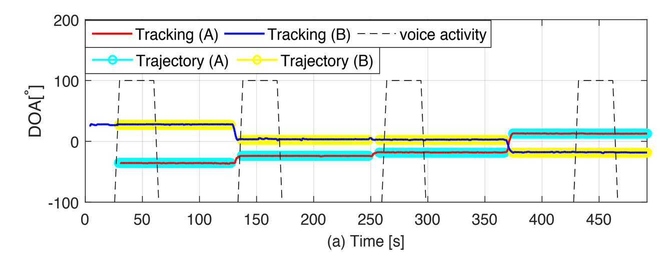
Setup

- Up to 2 people moving and speaking simultaneously
- Ego-noise and speeches recorded separately and added at different SNR for evaluation purposes

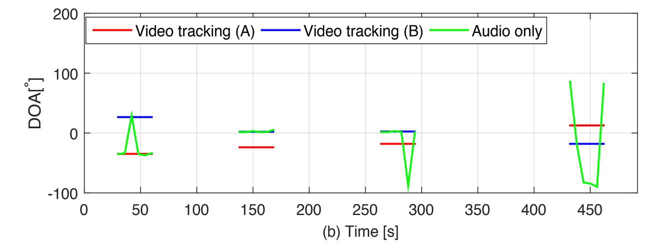
Evaluation measures

- DOA error
- Ratio between target and interfering speaker signal (SIR)
- Ratio between target and ego-noise signals (SNR)
- Perceptual Evaluation of Speech Quality (PESQ) [5]

Target sound localization results (two speakers)

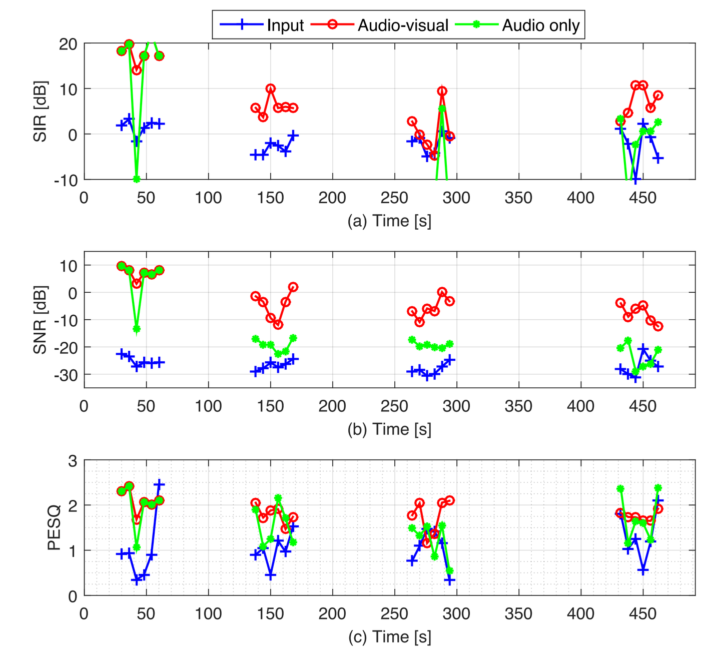


- Continuous video tracking results
- Accurate estimation for multiple speakers



- Audio-only and video-only tracking results at voice-active periods
- Audio-only cannot estimate the DOA for multiple speakers

Sound enhancement results (two speakers)



Results for one speaker

- Audio-only algorithm
  - can estimate only one DOA at a time
  - performs well only when the DOA is correct
- Audio-visual algorithm
  - can separate signals
  - outperforms state of the art

Conclusions

New audio-visual method that

- localizes multiple simultaneously talking speakers whose number is unknown
- robustly works with strong ego-noise (SNR < -15dB)
- separates and enhances each sound

References

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

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

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

[4] R. Sanchez-Matilla, F. Poiesi and A. Cavallaro. “Online multi-target tracking with strong and weak detections”. ECCV. 2016

[5] A. Rix, et al. “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs”. Proc. of IEEE ICASSP. 2001