Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement

Lin Wang, Ricardo Sanchez-Matilla, Andrea Cavallaro

***Abstract*— We present an audio-visual dataset recorded outdoors from a quadcopter and discuss baseline results for multiple applications. The dataset includes a scenario for source localization and sound enhancement with up to two static sources, and a scenario for source localization and tracking with a moving sound source. These sensing tasks are made challenging by the strong and time-varying ego-noise generated by the rotating motors and propellers. The dataset was collected using a small circular array with 8 microphones and a camera mounted on the quadcopter. The camera view was used to facilitate the annotation of the sound-source positions and can also be used for multi-modal sensing tasks. We discuss the audio-visual calibration procedure that is needed to generate the annotation for the dataset, which we make available to the research community[1].**

I. Introduction

Audio-visual sensing from a quadcopter is of interest for applications such as search and rescue, human-drone interaction and multimedia broadcasting [1]–[5]. However, the quality of sounds recorded from a quadcopter is poor due to the strong and time-varying ego-noise generated by the rotating motors and propellers, which cause extremely low signal-to-noise ratios, e.g. smaller than -15 dB [6], [7]. Moreover, the movement of the quadcopter itself and natural wind further complicate the analysis of sounds emitted by sources in the environment.

A number of microphone-array algorithms have been proposed to address these challenges for sound source localization [3], [7]–[15] and enhancement to extract target sounds masked by the strong ego-noise [1], [7], [15]–[24]. Based on the role of the onboard microphones and sensors used, these algorithms can be categorized as unsupervised or supervised approaches. Unsupervised approaches use microphone signals for the acoustic sensing task with beamforming [15], [17], [23], blind source separation [16], [17], time-frequency spatial filtering [7], [8], [17] or post-filtering [23]. Supervised approaches use additional sensors to monitor the quadcopter and to predict the ego-noise in order to assist the sound source localization process [13], [14] or the adaptive ego-noise cancellation [1], [19]. As quadcopters are generally equipped with an onboard camera, audio-visual processing methods can use the visual information to facilitate the localization of the sound source and the enhancement of the sound of target sources [9], [18].

Indoor datasets with multi-channel sound recordings captured from a drone platform are becoming available for sound source localization [25] and sound enhancement [26]. DREGON was captured with an 8-channel cube-shape microphone array mounted on a Mikrokopter drone [25]. AIRA-UAS was captured with an 8-channel circular microphone array mounted on three types of drones, a DJI Matrice 100, a 3DR Solo and a Parrot Bebop 2 [26]. These two datasets were collected indoors to facilitate the annotation using external positioning systems.

In this paper we present AVQ, the first annotated outdoor *Audio-Visual* dataset from a *Quadcopter* drone. The dataset can be used for audio-visual and audio-only tasks such as sound enhancement, sound source localization and tracking. We use an 8-element microphone array mounted on a quadcopter to record sounds in the environment as well as a camera to allow multi-modal tasks. The dataset consists of two subsets that capture up to two *static* sound sources emitting sound in front of the drone; and a *moving* sound source (see Fig. 1). We also describe the audio-visual calibration framework that we use to align temporally and geometrically the audio and visual signals.

The paper is organized as follows. Sec. II introduces the hardware and the calibration process we use between audio and video devices. Sec. III introduces the recording scenarios and the annotation of the dataset. Sec. IV analyzes the dataset, discusses potential applications and presents baseline results. Finally, in Sec. V we draw conclusions.

II. Audio-visual calibration

The sensing platform is composed of a 3DR IRIS quadcopter, an 8-microphone circular array located 15 cm above the body of the drone (diameter $d = 20$ cm, Boya BY-M1 omnidirectional Lavelier microphones) and a GoPro camera at the center of the microphone array. The positioning of the microphone array aims to avoid the noise caused by the wind blowing downwards from the propellers. The microphone signals are sampled synchronously at the rate of 44.1 kHz with a multichannel audio recorder (Zoom R24). The audio is recorded into the SD card of Zoom R24, while the video (sound and image) is recorded into the SD card of the GoPro. Since the microphone array and the GoPro camera work independently, a calibration procedure is needed to align temporally and geometrically the audio and video streams. We position the platform in a park at the height of 1.8 m on a tripod to record the audio-visual dataset.

As the GoPro camera has its own built-in microphone, we estimate the unknown time offset, δ_{av}, between the streams from the microphone array and the stream from the camera by matching the audio sequences from the

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.

[1]http://cis.eecs.qmul.ac.uk/projects/avq/

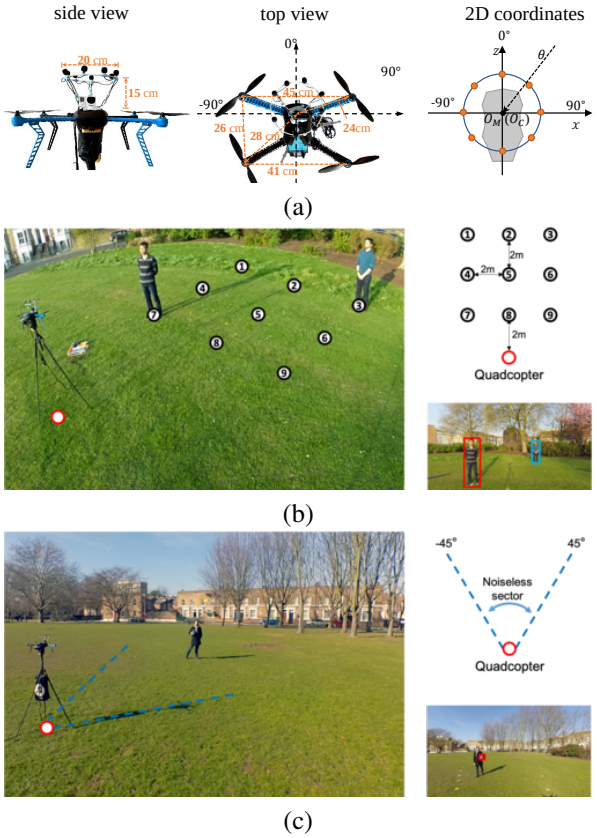

Fig. 1. The AVQ recording setup. (a) Side and top view of the audio-visual sensing platform; and 2D-coordinate system. O_M and O_C denote the centers of the microphone array and of the camera in the 2D plane, respectively. (b) Recording environment for Subset $\mathcal{S}1$: two people talk from nine locations. (c) Recording environment for Subset $\mathcal{S}2$: a loudspeaker is carried by a person walking in front of the drone. The left and right panels of (b) and (c) show the overall scene and the view from the onboard camera, respectively.

(a) (b)

Fig. 2. Image captured by the camera mounted on the drone (a) before and (b) after image undistortion.

microphone array and from the microphone of the camera using a calibration sound (e.g. clapping). All the audio and video sequences provided in the AVQ dataset are already temporally synchronized using this method.

Next, we use camera resectioning and geometrical alignment to represent audio and visual observations in a unified coordinate system.

A. Resectioning

The resectioning procedure aims to undistort the image and to estimate the camera parameters for geometrical alignment. After recording a calibration video of a checkerboard captured at different locations, we estimate the camera parameters with the Matlab Camera Calibration

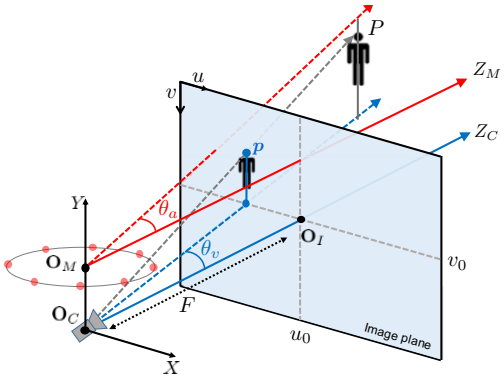

Fig. 3. The 3D coordinate systems for the microphone array, (X, Y, Z_M), and the camera, (X, Y, Z_C). The centers of the microphone array and camera are $\boldsymbol{O}_M$ and $\boldsymbol{O}_C$, respectively; $\boldsymbol{O}_I = (u_0, v_0)$ is the principal point (center) of the image; and F is the focal length of the camera. The sound source P is projected onto the image plane as $\mathbf{p}$ with visual angle θ_v. The audio angle from the sound source P to the array is θ_a.

Toolbox [30], obtaining the radial and tangential lens distortion parameter $\boldsymbol{\xi}$, and the intrinsic parameter $\boldsymbol{K}$:

$$\boldsymbol{K} = \begin{bmatrix} F_u & 0 & 0 \\ c_s & F_v & 0 \\ u_0 & v_0 & 1 \end{bmatrix}, \tag{1}$$

where F_u and F_v represent the horizontal and vertical components of the camera focal length, respectively, and the camera focal length is $F = \frac{F_u + F_v}{2}$ (measured in pixels); (u_0, v_0) indicate the location of the principal point (optical center) in the image; and c_s is the skew axis coefficient.

The parameter $\boldsymbol{\xi}$ is used to undistort the image frame as

$$\bar{\boldsymbol{I}} = \mathcal{D}(\boldsymbol{I}, \boldsymbol{\xi}), \tag{2}$$

where $\mathcal{D}(\cdot)$ represents the undistortion procedure, $\boldsymbol{I}$ and $\bar{\boldsymbol{I}}$ denote an image frame before and after undistortion, respectively [30]. An example is given in Fig. 2 illustrating an undistorted image processed with the estimated parameter $\boldsymbol{\xi}$. All the video sequences provided in the AVQ dataset are already undistorted.

The parameter $\boldsymbol{K}$ will be used in the geometric alignment when estimating from an image the visual angle of a sound source (visual object) with respect to the camera center.

B. Geometrical alignment

The geometrical alignment associates audio and video events in a unified coordinate system (Fig. 3). The 3D position P of a real-world object is projected on the image plane, where it is denoted as $\boldsymbol{p}$. Let θ_a and θ_v be the angles, on a 2D horizontal plane, of the object with respect to the microphone array and the camera. When an object emits a sound, its direction of arrival (DOA) can be estimated either from the microphone-array signals, θ_a, or from the visual signal, θ_v. Since the microphone array and the video camera have their own coordinate systems, to infer the DOA of the sound from the corresponding object in the image we need to know the relationship between θ_a and θ_v. In practice the centers of the microphone array O_M and the camera O_C

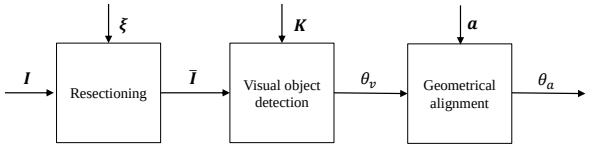

Fig. 4. Audio-visual calibration of the AVQ dataset. We undistort the original image $\boldsymbol{I}$ using the lens distortion parameter $\boldsymbol{\xi}$ (Eq. 2) and detect visual objects in the undistorted image $\bar{\boldsymbol{I}}$. From the location of the visual object we compute the visual angle θ_v using the camera intrinsic parameter $\boldsymbol{K}$ (Eq. 4) and convert to the audio angle θ_a using the calibration parameter $\boldsymbol{a}$ (Eq. 3).

are not perfectly aligned in Fig. 3. We thus represent the relationship between θ_a and θ_v as

$$\theta_a = a_1\theta_v + a_2, \tag{3}$$

where $\boldsymbol{a} = [a_1, a_2]^{\mathrm{T}}$ are unknown constants. To estimate a_1 and a_2, we record the sound from a speaker at L different locations with both the microphone array and the camera while the drone is muted. As an example, let us use the sound from the location Q. For the audio, the DOA of the sound, θ_a^Q, can be estimated from the microphone signal with the SRP-PHAT algorithm [27]. For the video, we manually label the sound emitting point (the mouth of the speaker) in the image, $\boldsymbol{p}_Q = (u_Q, v_Q)$, and then estimate its DOA as

$$\theta_v^Q = \arctan\frac{u_Q}{F}. \tag{4}$$

We thus estimate a set of DOAs of the speaker from the audio as $\boldsymbol{\theta}_a = [\theta_a^1, \ldots, \theta_a^L]^{\mathrm{T}}$ and from the video as $\boldsymbol{\theta}_v = [\theta_v^1, \ldots, \theta_v^L]^{\mathrm{T}}$. The vector of parameters $\boldsymbol{a}$ is then estimated from $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_v$ using least-square fitting.

Given the parameters $\boldsymbol{\xi}$, $\boldsymbol{K}$ and $\boldsymbol{a}$, we can calibrate the audio and video sequences provided in the dataset. Fig. 4 illustrates the calibration steps.

III. Dataset

The dataset consists of the $\mathcal{S}1$ and $\mathcal{S}2$ subsets, with natural and composite scenarios. $\mathcal{S}1$ includes up to two sound sources at fixed locations, whereas $\mathcal{S}2$ includes a moving sound source. In the *natural* scenario, the target sound and the ego-noise are recorded simultaneously. In the *composite* scenario, the target sound and the ego-noise are recorded separately, thus allowing one to evaluate the performance at different input signal-to-noise ratios (SNRs) and to compute the output SNR after processing [28].

In $\mathcal{S}1$, two people (the sound sources) talk at nine predefined locations in front of the drone (Fig. 1(b)). The distance between these locations and the drone varies between 2 m and 6 m. We record only composite scenarios, i.e. the clean speech and the ego-noise are recorded separately. When recording the ego-noise, the quadcopter operates at 50%, 100% or 150% of the power level of the hovering state. When recording speech, the two people talk in turns for about 40 s each and then move to the next location.

In $\mathcal{S}2$, a loudspeaker (the sound source) playing speech is carried by a person (Fig. 1(c)). As the relative location of the microphone array and the motors and propellers is fixed, the ego-noise tends to arrive from the side closer to the motors (back side of the array, with respect to the field of view of the camera), thus creating a sector with lower ego-noise (the front of the array). This allows us to identify a *noiseless sector* $[-45°, 45°]$ where a target sound can be more easily detected [7]. We record natural and composite scenarios. The drone operates either with a constant hovering power or with a time-varying power between 50% and 150% of the hovering state. The loudspeaker moves either in a *constrained* area (inside the noiseless sector) or in an *unconstrained* area (in front of the drone). The distance between the loudspeaker and the drone varies between 2 m and 6 m. Each trajectory lasts for about 3 minutes.

TABLE I
AVQ dataset: specifications.

Sub	Seq	Mod	Dur	VG	VAD	Type	Drone	Source
$\mathcal{S}1$	seq1	A	120s			EO	constant (50%)	/
	seq2	A	120s			EO	constant (100%)	/
	seq3	A	40 s			EO	constant (150%)	/
	seq4	AV	797s	✓	✓	SO	muted	2 sources 9 locations
	misc	microphone location, AV calibration parameters						
$\mathcal{S}2$	seq1	A	210s			EO	constant (100%)	/
	seq2	A	214s			EO	dynamic	/
	seq3	AV	215s	✓	✓	SO	muted	cons.
	seq4	AV	217s	✓	✓	SO	muted	uncons.
	seq5	AV	303s	✓	✓	MIX	constant (100%)	cons.
	seq6	AV	271s	✓	✓	MIX	constant (100%)	uncons.
	seq7	AV	258s	✓	✓	MIX	dynamic	cons.
	seq8	AV	249s	✓	✓	MIX	dynamic	uncons.
	misc	microphone location, AV calibration parameters						

KEY - Sub: Subset; Seq: Sequence; Mod. Modality; Dur: Duration; VG: Video ground-truth; VAD: voice activity detection; A: Audio-only; AV: Audio-visual; EO: ego-noise only; SO: speech only; MIX - mixture; cons: constrained area; uncons: unconstrained area; misc: miscellaneous.

Table I summarizes the specifications of the AVQ dataset. The audio is in WAV format with sampling rate 44.1 kHz. The video is in MP4 format with frame rate 30 fps, resolution 1920×1080 for $\mathcal{S}1$ and 1280×720 for $\mathcal{S}2$, and wide field of view (i.e. 70 vertical degrees and 120 horizontal degrees before undistortion). The total duration of the recordings is about 50 minutes.

To obtain the ground-truth locations of the sound source, for $\mathcal{S}1$ we use a person detector [29] and for $\mathcal{S}2$ we use a visual marker to assist the loudspeaker detection (see the example in the right panels of Fig. 1(b) and (c)). Fig. 5(a) and (b) depict the video ground-truth locations, θ_v, and the voice activity detector (VAD) information of the sound source for $\mathcal{S}1$ and $\mathcal{S}2$, respectively.

IV. Analysis and Results

A. Analysis

Fig. 6 depicts the time-domain, spectral, and spatial characteristics of the ego-noise of a drone operating at a

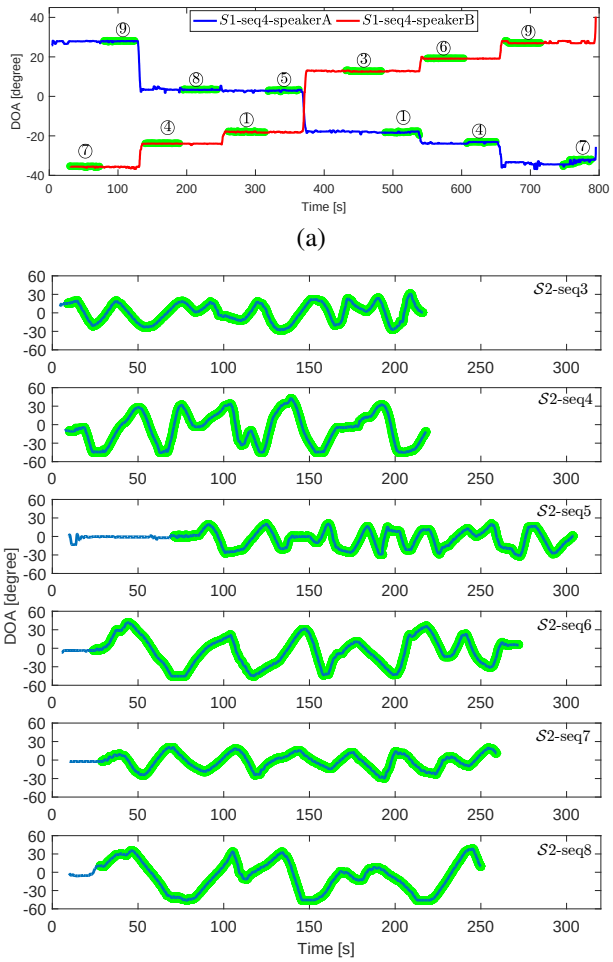

Fig. 5. AVQ dataset: video ground-truth trajectory of the sound sources. Green thick lines: voice activity periods; circled numbers: sound source locations shown in Fig. 1(b). (a) Subset $\mathcal{S}1$. (b) Subset $\mathcal{S}2$.

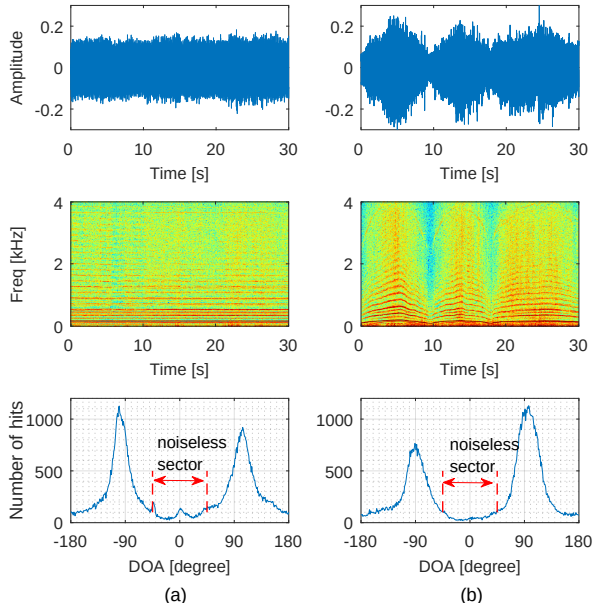

Fig. 6. Time-domain waveform, time-frequency spectrogram, and spatial analysis of the ego-noise. (a) The ego-noise of a drone operating with constant power (at hovering power). (b) The ego-noise of a drone operating with time-varying power (between 50% and 150% of the hovering power).

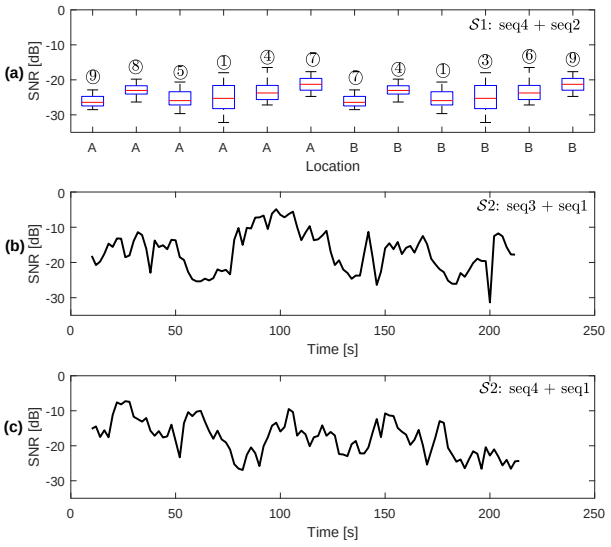

Fig. 7. The SNR of composite sequences generated by mixing a clean speech with the ego-noise (drone operating at hovering state). (a) $\mathcal{S}1$: seq4 + seq2. (b) $\mathcal{S}2$: seq3 + seq1. (c) $\mathcal{S}2$: seq4 + seq1.

constant power and a time-varying power. In both cases, the duration of the ego-noise sample is 30 seconds. The ego-noise mainly consists of narrow-band harmonic noise, which is caused by the mechanical sound of the rotating motors, and full-band noise, which is caused by the rotating propellers cutting the air. As can be observed from the spectrogram, the fundamental of the harmonics typically varies with time, corresponding to the operating power of the drone. The spatial characteristics are illustrated by the histogram of the local DOA estimation at individual time-frequency bins in the 30-second duration [8]. In both Fig. 6(a) and (b), the histogram plot presents several high peaks, which correspond to the direction of arrival of the sound of the motors. The histogram always presents low histogram values in the sector $[-45°, 45°]$, which we refer to as the noiseless sector. This is because the microphone array is placed at the front of the body of the drone (see Fig. 1(a)) and the ego-noise tends to arrive from the back side of the array. The time-frequency sparsity and the spatial characteristics of the ego-noise can be exploited for acoustic sensing algorithms [7].

Fig. 7 illustrates the input SNR of some composite sequences generated by mixing a clean speech with the ego-noise (when the drone is operating at hovering state). The SNR over a segment $\mathbb{B}$ is defined as the power ratio between the clean speech $s(n)$ and the ego-noise $v(n)$ [28]:

$$\text{SNR} = 10\log_{10}\frac{\sum_{n\in\mathbb{B}} s^2(n)}{\sum_{n\in\mathbb{B}} v^2(n)}. \tag{5}$$

We compute the SNR on active VAD periods and over non-overlapping segments of 2 seconds long. In Fig. 7(a), we generate the composite data using the sequences in $\mathcal{S}1$

TABLE II

AVQ DATASET: TESTING SCENARIOS.

Evaluation	Scenario	Modality	Sequence involved	
			$\mathcal{S}1$	$\mathcal{S}2$
Sound enhancement (up to 2 sources)	composite	A/AV	seq1-seq4	
Sound localization (up to 2 sources)	composite	A/AV	seq1-seq4	
Source tracking	composite	A/AV		seq1-seq4
Source tracking	natural	A/AV		seq5-seq8

KEY - A: audio-only scenario; AV: audio-visual scenario.

and boxplot the SNR per speaker location. In all locations, the median SNR is lower than -20 dB. In Fig. 7(b) and (c), we generate the composite data using the sequences in $\mathcal{S}2$ and plot the variation of SNR over time. Since the distance between the speaker and the drone changes with time, the SNR also varies dynamically.

B. Baseline results

AVQ enables the evaluation of sensing performance in different scenarios (see Table II). $\mathcal{S}1$ can be used to generate composite scenarios to evaluate sound enhancement and source localization at different input SNRs [28]. $\mathcal{S}2$ can be used to generate composite or natural scenarios to evaluate the tracking performance of the moving sound source. Based on the modality used, the dataset is appropriate for evaluating audio-only and audio-visual joint processing algorithms.

We use AVQ to evaluate the performance of state-of-the-art (baseline) acoustic sensing algorithms based on sparsity-based time-frequency spatial filtering [7], [17]. Considering that the energy of speech and the ego-noise are usually concentrated at isolated time-frequency bins, the algorithms compute local DOAs of the acoustic signal at individual time-frequency bins. These local DOAs are used to construct a spatial filter steering at the desired direction for sound enhancement [17], or to estimate the location of a target sound by steering the spatial filter at a set of candidate directions [7], [8]. When the sound source is moving, the time-frequency spatial filtering can be used to estimate the source location in a block-wise manner and the accuracy can be improved with a tracker [9]. When a camera is available, the time-frequency filtering can be steered at the target direction estimated from the video [18].

Next, we show two examples of using the AVQ dataset for evaluating the sound enhancement, source localization and tracking performance of the baseline algorithms.

The first example is generated with the sequences $\mathcal{S}1$-seq2 and $\mathcal{S}1$-seq4 and is used to evaluate the performance of sound source localization and sound enhancement for a single sound source (i.e. speaker B in Fig. 5(a)) embedded in the ego-noise, assuming that the VAD information of the sound source is known. Fig. 8(a) presents the sound source localization result achieved by time-frequency spatial filtering [8], and compares it with the video ground-truth. Fig. 8(b) presents the sound enhancement performance achieved by two algorithms: the audio-only [7], which

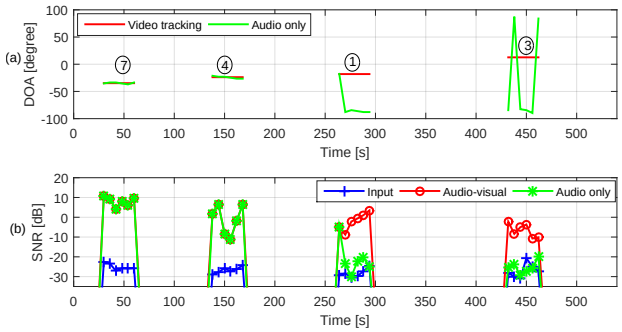

Fig. 8. Baseline results for (a) source localization and (b) sound enhancement using the dataset $\mathcal{S}1$.

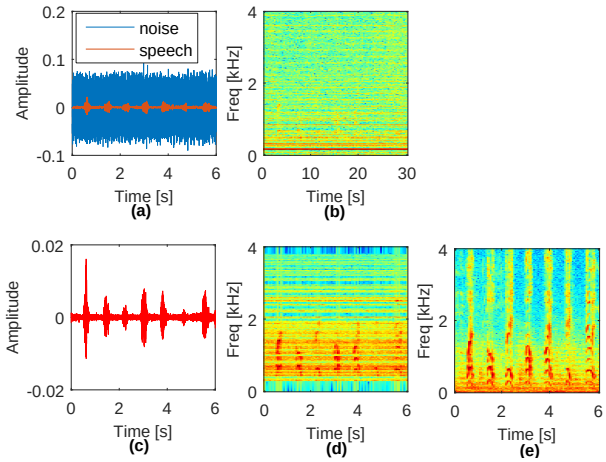

Fig. 9. Example of processing results using the data recorded at location ⑦. (a) Time-domain waveforms of speech and ego-noise. (b) Spectrogram of the mixture signal. (c,d) Time-domain waveform and spectrogram of the enhanced output. (e) Spectrogram of the clean speech signal for reference. The input and output SNRs are -22 dB and 10 dB.

enhances a target direction estimated from the audio signal; and the audio-visual algorithm [18], which enhances a target direction estimated from the video signal. Fig. 9 shows an example of the processed results using the data recorded at location ⑦. The input SNR is extremely low (-22 dB) and the clean speech is masked by the strong ego-noise thus making sound source localization and sound enhancement very challenging. However, the baseline algorithm manages to extract the target speech from the noisy signal, with an output SNR of 10 dB.

The second example is generated with all the eight sequences from $\mathcal{S}2$ and is used to evaluate the tracking performance of the moving sound source. Table III presents the mean and standard deviation of the localization error by comparing the video ground-truth with the results achieved by only using a time-frequency spatial filtering (TF) and combining TF and particle filter tracking (TFT) [9]. The results of these time-frequency spatial filtering methods on AVQ are available with the dataset and the ground-truth annotation at http://cis.eecs.qmul.ac.uk/projects/avq/. More baseline results using the AVQ dataset are also presented in [9], [18].

TABLE III

BASELINE SOURCE TRACKING RESULTS USING DATASET $\mathcal{S}2$, IN TERMS OF MEAN (STANDARD DEVIATION) LOCALIZATION ERROR IN DEGREES.

Composite ($S2$)			**Natural** ($S2$)		
Sequence	TF	TFT	Sequence	TF	TFT
seq1 + seq3	3.5 (4.7)	3.8 (3.9)	seq5	8.7 (7.5)	9.1 (7.4)
seq2 + seq3	4.3 (7.8)	4.4 (4.5)	seq6	8.8 (8.4)	8.3 (6.5)
seq1 + seq4	7.4 (9.0)	8.2 (7.8)	seq7	14.7 (18.9)	10.3 (8.8)
seq2 + seq4	8.0 (17.6)	8.4 (21.2)	seq8	16.4 (19.5)	11.5 (9.3)

KEY - TF: time-frequency spatial filtering; TFT: TF + tracking.

V. CONCLUSION

We presented AVQ, an audio-visual dataset recorded outdoors with an 8-microphone circular array and a camera mounted on a quadcopter. We also presented results of state-of-the-art algorithms on this dataset and the audio-visual calibration procedure that is needed to generate the ground-truth annotation of the position of the sound sources from the video captured by the onboard camera. We hope that AVQ will inspire the work of researchers in the audio-visual sensing problems in the presence of strong ego-noise.

Acknowledgement: L. Wang acknowledges the support by the Institute of Coding, which is supported by the Office for Students (OfS) and the Higher Education Funding Council for Wales (HEFCW).

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[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] J. Cacace, R. Caccavale, A. Finzi, and V. Lippiello, "Attentional multimodal interface for multidrone search in the Alps" in *Proc. IEEE Int. Conf. Syst. Man, Cybernetics*, Budapest, Hungary, 2016, pp. 1178-1183.

[6] 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.

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

[8] 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.

[9] L. Wang, R. Sanchez-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. 1-8.

[10] 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.

[11] 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.

[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] 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.

[14] 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.

[15] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Beamforming-based acoustic source localization and enhancement for multirotor UAVs," in *Proc. European Signal Processing Conf.*, Rome, Italy, 2018, pp. 987-991.

[16] 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.

[17] 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.

[18] 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*, Mountain View, USA, 2017, pp. 1591-1599.

[19] 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.

[20] 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.

[21] R. P. Fernandes, E. C. Santos, A. L. L. Ramos, and J. A. Apolinario Jr., "A first approach to signal enhancement for quadcopters using piezoelectric sensors," in *Proc. Int. Conf. Transformative Sci. Eng. Business Social Innovation*, Fort Worth, USA, 2015, pp. 536-541.

[22] S. Yoon, S. Park, and S. Yoo, "Two-stage adaptive noise reduction system for broadcasting multicopters," in *Proc. IEEE Int. Conf. Consum. Electron.*, Las Vegas, USA, 2016, pp. 219-222.

[23] 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.

[24] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, 2018, pp. 1-5.

[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. 1-8.

[26] O. Ruiz-Espitia, J. Martinez-Carranza, and C. Rascon, "AIRA-UAS: an evaluation corpus for audio processing in unmanned aerial system," in *Proc. Int. Conf. Unmanned Aircraft Systems*, Dallas, USA, 2018, pp. 836-845.

[27] 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. 1573-1588, 2016.

[28] 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.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Las Vegas, USA, 2016, pp. 779-788.

[30] J. Heikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 1997, pp. 1106-1112.