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.

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, \( \delta_{\text{avr}} \), between the streams from the microphone array and the stream from the camera by matching the audio sequences from the
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 Toolbox [30], obtaining the radial and tangential lens distortion parameter $\xi$, and the intrinsic parameter $K$:

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

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 $e_s$ is the skew axis coefficient.

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

$$\bar{I} = D(I, \xi),$$

where $D(\cdot)$ represents the undistortion procedure, $I$ and $\bar{I}$ 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 $\xi$. All the video sequences provided in the AVQ dataset are already undistorted.

The parameter $K$ 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 $p$. Let $\theta_a$ and $\theta_v$ 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, $\theta_a$, or from the visual signal, $\theta_v$. 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 $\theta_a$ and $\theta_v$. In practice the centers of the microphone array $O_M$ and the camera $O_C$...
are not perfectly aligned in Fig. 3. We thus represent the relationship between $\theta_a$ and $\theta_v$ as

$$\theta_a = a_1 \theta_v + a_2,$$

(3)

where $a = [a_1, a_2]^T$ 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, $\theta^a_Q$, 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, $p_Q = (u_Q, v_Q)$, and then estimate its DOA as

$$\theta^v_Q = \arctan \frac{u_Q}{v_Q}.$$  

(4)

We thus estimate a set of DOAs of the speaker from the audio as $\theta_a = [\theta^a_1, \ldots, \theta^a_L]^T$ and from the video as $\theta_v = [\theta^v_1, \ldots, \theta^v_L]^T$. The vector of parameters $a$ is then estimated from $\theta_a$ and $\theta_v$ using least-square fitting.

Given the parameters $\xi$, $K$ and $a$, 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 $S1$ and $S2$ subsets, with natural and composite scenarios. $S1$ includes up to two sound sources at fixed locations, whereas $S2$ 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 $S1$, 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 $S2$, 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 $[-45^\circ, 45^\circ]$ 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 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 $1920 \times 1080$ for $S1$ and $1280 \times 720$ for $S2$, 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 $S1$ we use a person detector [29] and for $S2$ 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, $\theta_v$, and the voice activity detector (VAD) information of the sound source for $S1$ and $S2$, 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
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 \([-45^\circ, 45^\circ]\), 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 \(B\) is defined as the power ratio between the clean speech \(s(n)\) and the ego-noise \(v(n)\) [28]:

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

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 \(S_1\) and \(S_2\).
TABLE II
AVQ DATASET: TESTING SCENARIOS.

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Scenario</th>
<th>Modality</th>
<th>Sequence involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound enhancement</td>
<td>composite</td>
<td>A/V</td>
<td>seq1-seq4</td>
</tr>
<tr>
<td>(up to 2 sources)</td>
<td></td>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>Sound localization</td>
<td>composite</td>
<td>A/V</td>
<td>seq1-seq4</td>
</tr>
<tr>
<td>(up to 2 sources)</td>
<td></td>
<td></td>
<td>S2</td>
</tr>
<tr>
<td>Source tracking</td>
<td>composite</td>
<td>A/V</td>
<td>seq1-seq4</td>
</tr>
<tr>
<td>Source tracking</td>
<td>natural</td>
<td>A/V</td>
<td>seq5-seq8</td>
</tr>
</tbody>
</table>

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 S2 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). S1 can be used to generate composite scenarios to evaluate sound enhancement and source localization at different input SNRs [28]. S2 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 S1-seq2 and S1-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 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 S2 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/](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 S2, IN TERMS OF MEAN (STANDARD DEVIATION) LOCALIZATION ERROR IN DEGREES.

<table>
<thead>
<tr>
<th>Composite (S2)</th>
<th>Natural (S2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>TF</td>
</tr>
<tr>
<td>seq1 + seq3</td>
<td>3.5 (4.7)</td>
</tr>
<tr>
<td>seq1 + seq3</td>
<td>4.3 (7.8)</td>
</tr>
<tr>
<td>seq1 + seq4</td>
<td>7.4 (9.0)</td>
</tr>
<tr>
<td>seq2 + seq4</td>
<td>8.0 (17.6)</td>
</tr>
</tbody>
</table>

KEY: TF: time-frequency spatial filtering; TFF: 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 (OIS) and the Higher Education Funding Council for Wales (HEFCW).

REFERENCES