DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization

Martin Strauss, Pol Mordel, Victor Miguet, Antoine Deleforge

DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization

Martin Strauss[1], Pol Mordel[2], Victor Miguet[3] and Antoine Deleforge[4]

***Abstract*— This paper introduces DREGON, a novel publicly-available dataset that aims at pushing research in sound source localization using a microphone array embedded in an unmanned aerial vehicle (UAV). The dataset contains both clean and noisy in-flight audio recordings continuously annotated with the 3D position of the target sound source using an accurate motion capture system. In addition, various signals of interests are available such as the rotational speed of individual rotors and inertial measurements at all time. Besides introducing the dataset, this paper sheds light on the specific properties, challenges and opportunities brought by the emerging task of UAV-embedded sound source localization. Several baseline methods are evaluated and compared on the dataset, with real-time applicability in mind. Very promising results are obtained for the localization of a broad-band source in loud noise conditions, while speech localization remains a challenge under extreme noise levels.**

I. Introduction

Unmanned aerial vehicles (UAV), commonly referred to as *drones*, have been of increasing influence in recent years. Applications such as autonomous human transport machines or delivery devices for postal services are being envisioned [1]. Search and rescue scenarios where humans in emergency situations need to be quickly found in areas difficult to access also constitute a potentially large field of application. While a number of UAV-embedded tools to address such situations have been developed using video cameras [2], audio-based source localization from UAVs has received much less research attention [3], [4], [5]. Though, UAVs equipped with a microphone array may present several advantages in emergency situations, especially whenever there is a lack of visual feedback due to bad lighting conditions (night, fog, etc.) or obstacles limiting the field of view [3].

With that in mind, to push forward investigations of audio properties during a UAV flight and the development of new embedded sound source localization method for search and rescue, this paper presents a novel state-of-the-art dataset called *DREGON* for DRone EGonoise and localizatiON. A quadrotor UAV equipped with an 8-channel cube-shaped microphone array (Fig. 1) was used to record in-flight audio in different scenarios with or without a target source simulated by a loudspeaker emitting various sounds. The two large rooms used for the flights were low-reverberant and equipped with a motion capture system. The system was used in all flights to obtain precise ground truth 6 degrees-of-freedom (DoF) coordinates of the UAV and target source in the room at all time. Additional synchronized signals of interest obtained from sensors embedded in the UAV and referred to as *log signals* are also included. These include the commanded and measured speed of each of the four propellers and inertial sensor measurements at all time. External fixed viewpoint video recordings of all flights are included for reference. In addition to in-flight recordings, reference noiseless recordings of static sound sources from all-sphere directions as well as recordings of the individual propellers at various speed are included. This is the first time a dataset of this kind is publicly released, to the best of the authors' knowledge.

Fig. 1: The MikroKopter quadrotor UAV used for the DREGON dataset, with our 3-D printed 8-channel microphone array mounted on the bottom. Green circles highlight two of the passive markers used for motion capture and yellow circles highlight two of the microphones (best seen in colors).

As an illustration, the performance of different popular *sound source localization* (SSL) methods, namely, variations of the GCC-PHAT [6], [7] and MUSIC [8], [9] algorithms, are compared on the proposed dataset with a focus on real time capabilities. Encouragingly, the best performing method enables in-flight estimation of the azimuth and elevation of a broadband source with 2° accuracy, while in-flight speech localization remains a challenge. Finally, different properties

[1]Martin Strauss is with Friedrich-Alexander University, Erlangen, Germany. martin.strauss@fau.de

[2]Pol Mordel is with CNRS/IRISA Rennes - Bretagne Atlantique, Rennes, France. pol.mordel@irisa.fr

[3]Victor Miguet is with ENS Rennes, France.

[4]Antoine Deleforge is with Inria Rennes - Bretagne Atlantique, Rennes, France. antoine.deleforge@inria.fr

of the recorded flight noise are analyzed and some leads for future research are outlined. The dataset as well as MATLAB code for the baseline methods are publicly available on the project website: `dregon.inria.fr`.

A. Challenges and Opportunities of the Dataset

The SSL capability of a UAV is influenced by a variety of effects. One major issue is the noise produced by the UAV itself, generically referred to as *ego-noise* in robotics [10]. Due to the quickly changing speed of motors to stabilize the vehicle in the air or to change in its position, the noise profile is highly non-stationary. Additionally, since the microphones are mounted on the drone itself, they are very close to the noise sources leading to high noise levels. Because of this, the SNR can easily reach -15 dB or less [11] making SSL very difficult. Another factor impacting localization performance is wind noise. The wind is produced by the rotating propellers, the UAV movement in the air and may also occur naturally in outdoor scenarios. This wind noise has high power and is of low-frequency. Hence, it easily overlaps with speech signals which typically occur in a similar frequency range [12]. Last, SSL must be performed using relatively short time windows, due to the fast movements of the UAV relative to potential sound sources. All these challenges need to be tackled at the same time and in near real-time for real-life SSL scenario such as search and rescue.

On the bright side however, using microphones embedded in a UAV comes with interesting opportunities. Additionally to audio signals, other signals recorded by various embedded sensors (gyroscope, motor controllers, inertial measurement unit, compass, ...) may be available. We believe that multi-modal approaches fusing information from multiple sensors in order to enhance SSL and source tracking present a promising direction for future research, as investigated in [13], [14].

B. Related work

Ego-noise reduction is a topic of increasing interest in robotics for several years now [10]. Different techniques have been reported to perform well, *e.g.*, non-negative matrix factorization [15], deep neural networks [16] or dictionary learning [17]. On the other hand, SSL is a long standing and extensively studied topics in robotics [18]. However, both ego-noise reduction and SSL for the specific setting of UAV-embedded microphones are still rather new topics [3], [19], [4], [14], [20], [5], [11], [11]. Many SSL approaches developed in recent years for robotics are different variations of the *MUltiple SIgnal Classification* (MUSIC) algorithm. For instance, [19] presents an incremental version of the *generalized eigenvalue decomposition* (GEVD)-MUSIC algorithm [21] referred to as iGEVD-MUSIC. The method is showed to perform better than the original GEVD version at a lower computational cost in an outdoor scenario. In [4], it is claimed that using *generalized singular value decomposition* (GSVD) instead of GEVD in MUSIC is more efficient. The proposed incremental GSVD-MUSIC (iGSVD) method shows improved performance compared to iGEVD. Additionally, a *correlation matrix scaling* (CMS) is applied to further improve results. Alternatively to MUSIC, Generalized Cross Correlation (GCC) methods are used for robot SSL in [3] and in the general framework ManyEars [22].

In [13], a UAV-embedded SSL method using both pre-recorded and on-flight propeller speed data is proposed. These data are used to estimate an adaptive noise correlation matrix in a Gaussian process regression model. This matrix is then used in GEVD-MUSIC to improve robustness to noise. As a result, the proposed method outperforms all comparison methods especially in high SNR conditions. In the same spirit, [14] presents a Deep Neural Network (DNN) approach to UAV-embedded SSL. To overcome the large training data requirements of DNNs, a partially shared network learning multiple tasks at the same time is implemented.

The authors of [11] compare beamforming, blind source separation (BSS) and time-frequency filtering algorithms on UAV-embedded recordings. The recordings are made with a circular microphone array placed over a UAV fixed on a tripod. The method requires the position of the sound source to be known beforehand, as opposed to the SSL scenario addressed here. BSS for UAV ego-noise reduciton is also investigated in [20]. In [5], two different UAV microphone array designs are proposed and the SEVD-MUSIC and iGSVD-MUSIC algorithms are compared on an outdoor SSL scenario. SSL success rates of almost 100 % are obtained even in low SNR conditions (< 0 dB). The authors propose to adapt the algorithms depending on the considered scenario and emphasize their high computational costs as a major drawback for real-time applicability.

II. The Dregon Dataset

This section reports on the hardware and protocols used to gather the data and on the detailed content of the dataset. A summary of the dataset content can be found in Table I.

A. Hardware and Recording Environments

A quadrotor UAV MK-Quadro from MikroKopter (HiSystems GmbH, Moormerland, Germany) (see Fig. 1) was employed. This customizable quadrotor is equipped with four MK2832-35 motors[1]. The usual MK-Quadro setup was extended with an ODROID-XU4 Linux Computer (Hardkernel co., Ltd., GyeongGi, South Korea) running ROS and the TeleKyb-genom3 framework [23] for implementing the low-level flight control receiving the body-frame velocity commands (25), and for exchanging data via Wi-Fi with the ground station. For the sound recordings, a specifically designed 3-D printed cube-shaped microphone array was mounted under the UAV. This placing was motivated by the assumption that a potential target source would be located under the UAV, enabling spatial filtering of ego-noise.

[1] `http://wiki.mikrokopter.de/MK2832-35`

TABLE I: The DREGON dataset: content summary.

<table>
<tr><th colspan="4">DREGON dataset</th></tr>
<tr><td colspan="3">Noise-Free Source signals</td><td>In Flight Noise-Only Recordings</td></tr>
<tr><td rowspan="4">Type
Distance

Azimuth
Elevation</td><td colspan="2">Chirp, white noise, speech (semi-anechoic room)</td><td rowspan="2">Hovering, up and down, rectangle,
spinning cw and ccw (smallest room),
free flight (both rooms)</td></tr>
<tr><td>1.2 m</td><td>2.4 m</td></tr>
<tr><td>-30° -15° 0°
-45° -30° -15° 0°</td><td>-45° -30° -15° 0°
-45° -30° -15° 0°</td><td>Log signals</td></tr>
<tr><td colspan="2"></td><td>Vicon motion capture data</td></tr>
<tr><td colspan="2"></td><td>Video recordings</td></tr>
<tr><td colspan="3">Individual Motor Recordings</td><td>In-Flight Source Recordings</td></tr>
<tr><td>Commanded speed [turns/s]</td><td colspan="2">50 60 70 80 90</td><td>White noise (loud), speech (loud),
white noise (less loud), speech (less loud),
”silent flight” + white noise.</td></tr>
<tr><td rowspan="3">Motors</td><td colspan="2" rowspan="3">Individual motors, all four at 70 turns/s</td><td>Log signals</td></tr>
<tr><td>Vicon motion capture data</td></tr>
<tr><td>Video recordings</td></tr>
</table>

Further, this location partially protects the microphone array from possible shocks and from the wind generated by the rotors. The array contains eight microphones (one on each vertex, see Fig. 1) and is connected to a sound card[2] mounted on the drone via a 3D-printed case. All audio recordings have a sampling frequency of 44.1 kHz. The total weight of the equipped quadrotor is 1.68 kg. The propellers' speed ranges from 15 turns/s when they starts, up to around 95 turns/s at maximum. The UAV was flown in two different carpeted indoor rooms with respective volumes 10 m×10 m×2.5 m and 12 m×12 m×3.5 m and reverberation time under 150 ms. A low reverberation time was chosen for this dataset because typical search and rescue scenarios occur in outdoor environments, where little reverberation is present. Both rooms were equipped with a twelve-camera Vicon motion capture system[3] (see Fig. 2). On top of audio and Vicon data, log signals obtained from the embedded inertial measurement unit (3D anguar velocities and accelerations) and the four propellers (commanded and feedback speed of each of the 4 propellers) are provided in the dataset. The log and Vicon signals are synchronized together via timestamps provided by the robotic middleware ROS. Synchronized timestamps were then manually created for each audio sample based on motors onsets and outsets. Maximal synchronization errors are estimated to be below 100 ms. All flights were video-recorded by a wide-angle camera mounted on a tripod for reference.

B. In-flight Source Recordings

Four in-flight recordings of a single static source were performed in the largest room. The UAV was tele-operated by a human and was performing free-flights combining sequences of take-off, landing, stabilization, hovering, straight line, circles and spinning in a realistic way. The source was simulated by a Genelec loudspeaker placed on a table in the center and emitting either random speech utterances from the TIMIT dataset [24] or white noise at two different levels, achieving recorded audio signal-to-noise ratios (SNRs) ranging from -16 to -10 dB. The Vicon motion capture system was used to obtain the 6 DoF coordinates of both the UAV and the loudspeaker, thanks to trackers placed on each of them. These coordinates were notably used to calculate the ground truth 3D positions of the source in the UAV's frame, as showed in Fig.5. On top of the four recordings obtained from real UAV flights, one recording of a ”silent flight” with a white noise source was made. The UAV motors were off and a human operator was carrying the UAV to simulate a flight. This recording also includes 10s of ”room silence”, for reference.

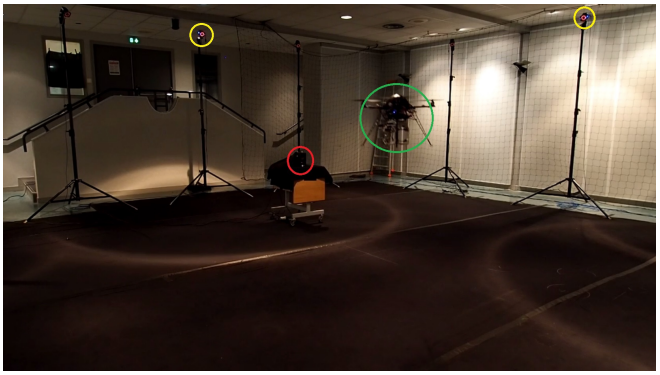


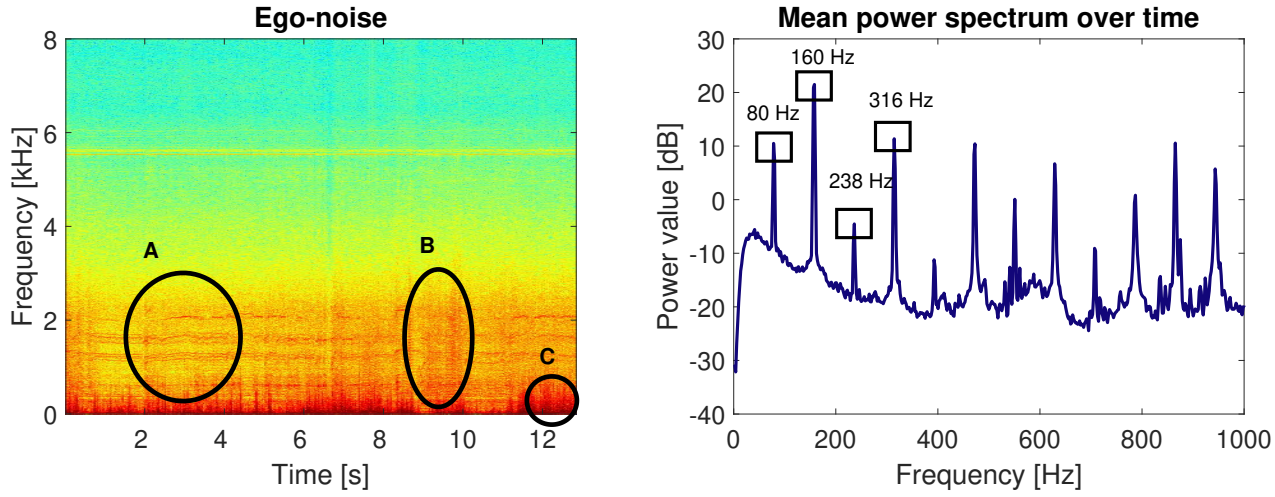
Fig. 2: Setup of the room were in-flight source recordings (Sec. II-B) were performed. The flying UAV is highlighted by a green circle, the loudspeaker simulating the source by a red circle, and two of the twelve Vicon cameras by yellow circles (best seen in colors).

C. In-flight Noise-Only Recordings

In order to isolate the noise characteristics of UAV-embedded audio signals, a number of recordings were made with the UAV flying in the absence of a sound source. Different types of flights were performed: *hovering*, *up-and-down*, *spinning clockwise (cw) and counterclockwise (ccw)*, *rectangle* and *free flight*. While free flights were performed in both rooms, the other types were performed in

[2] https://sourceforge.net/p/eightsoundsusb/wiki/Main_Page/

[3] Vicon motion capture systems Ltd., Oxford, UK.



(a) (b)

Fig. 3: (a) Log-magnitude spectrogram recorded by one embedded microphone during free flight. In **A** harmonic components, in **B** structural noise and in **C** wind noise can be observed. (b) Mean power spectrum of the sound generated by an individual motor rotating at 80 turns/s. Peaks at harmonics proportional to 80 Hz can clearly be identified.

the smallest room only. *Hovering* simply describes taking off and holding position, while *up-and-down* includes multiple vertical movements. For *spinning cw and ccw* the drone was taking off, holding position and spinning around the yaw angle in both directions. The fourth type of flight performed was a *rectangle*, where the drone takes off and flies in a horizontal rectangular shape around the room.

In addition to flight noise, other recordings were made while the UAV was fixed on a nylon string hanging freely in the room and held still by a human operator. This was done in order to record pure motor noise at a chosen speed without having the influence of balancing and stabilizing movements. Each individual motor and the four of them jointly were hence recorded for 10s at speeds of 50, 60, 70, 80 and 90 turns/s. An interesting feature of these recordings is that they contain very few wind noise.

D. Noise-Free Recordings

While testing SSL methods on realistic flight scenarios is necessary, it is also desirable to test them on more controlled settings. To this aim, the DREGON dataset includes a number of noise-free recordings of a target sound source using the microphone array of previous experiments. These recordings can be conveniently summed together and with the noise-only recordings of previous section to simulate noisy audio excerpts of any length involving any number of static sources from any fixed directions and at any SNR. They were performed in a third semi-anechoic and acoustically isolated room in order to have as little interference as possible. The microphone array was fixed on a tripod while a loudspeaker was moved to different elevation and azimuth angles and two different distances to the array. From a 1.2 m distance we used all combinations of azimuth angles $\{-45°, -30°, -15°, 0°\}$ and elevation angles $\{-30°, -15°, 0°\}$ yielding 12 different directions. From a 2.4 m distance an additional elevation angle of $-45°$ was added yielding 16 different positions. Note that thanks to the rotational symmetries of the designed microphone array, over 100 source directions around the whole sphere can be conveniently simulated using these directions, by appropriately permuting the 8 channels. For each direction, a white noise sound, a chirp and random speech utterances from the TIMIT dataset [24] were recorded. This resulted in a total of 84 annotated clean source signals. The signal durations were 10 s for chirps, 2 s for white noise sounds and 14s in average for speech signals.

III. Acoustic Analysis of UAV Noise

The acoustic noise at microphones during UAV flights consists of several parts. As pointed out in [20], three main components can be identified: mechanical noise (ego-noise), air flow noise created by the propellers and wind noise. Fig. 3 shows a noise spectrogram obtained from one of the microphones during a free flight, where these three components are highlighted. As can be seen, most of the noise's energy is located in frequencies below 3 kHz. This is particularly problematic when the target source signal is human voice, which usually covers a similar frequency range. Among the different components, wind noise, located below 1 kHz, is by far the loudest one when present. However, we observed that wind noise was rarely present in more than 4 out of 8 channels at once. This deserves further investigation and could be exploited by removing corrupted channels using a wind detection method [12].

In general it can be observed that the noise contains many non-stationary components, making its modelling non-straightforward. However there seems to exist an overall background component which is roughly stationary during

most of the flight. Our hypothesis is that this component is generated mainly by the motor rotations and vibrations of the structure. Additionally, Fig. 3 shows the existence of several spectral harmonic components. To investigate this further, we used individual motor recordings of the DREGON dataset (Section II-C). In Fig. 3 it can be seen how a motor speed of 80 turns/s leads to peaks in the power spectral density along the harmonics proportional to 80 Hz. This observation paves the way for methods exploiting instantaneous motor speed logs to cancel these peaks. While these harmonic components are highly non-stationary whenever the drone undergoes rapid changes of directions, it still seems rather reasonable to assume that the noise is approximately stationary during the most steady parts of the flight (with exception to wind components). Based on this approximation, a stationary filtering technique described in Section IV-B is used to reduce noise in the SSL experiments of Section V.

IV. Baseline Sound Localization Methods

In this section some baseline sound localization methods are briefly outlined. These methods are tested on the dataset and results are reported in Section V.

A. Sound Source Localization

The most widely used feature to localize a sound source is the pairwise time difference of arrival (TDOA) of the source signal at two microphones. A number of methods exist to estimate TDOAs and directions of arrival from recorded signals. In this study, we focus on two such methods which are widely used in SSL, namely the generalized cross-correlation phase-transform (GCC-PHAT) method [6], [7] and the multiple signal classification (MUSIC) method [8], [9].

GCC-PHAT as described in *e.g.*, [7], [25], [22], takes as input a pair of time-domain audio signals, and a set of directions of arrival (DOA) to be tested. A DOA is represented by a single angle from which the source signal impinges at the microphone pair, assuming the source is in the far field. Assuming further that sound propagates in free field from the source to each microphone, a one-to-one mapping exists between DOAs and TDOAs. Based on these assumptions, the methods output a *score* for each tested DOA, referred to as the *angular spectrum* for this microphone pair [25]. The Multi-Channel BSS Locate Matlab toolbox[4] (mBSSL) implements a general framework that efficiently combines estimated angular spectra from multiple microphone pairs to estimate the 2D direction (azimuth, elevation) of one or several sound sources, given the array geometry. It can be viewed as a computationally efficient variation of SRP-PHAT [26]. For convenience, we refer to mBSSL used with GCC-PHAT angular spectra as simply GCC-PHAT in the following.

[4] http://bass-db.gforge.inria.fr/bss_locate/

The method MUSIC [8], [9] is a so-called subspace method known to be more robust to noise than GCC-PHAT. It also outputs an angular spectrum given a set of test DOAs. However, MUSIC is most effective when using an array containing more than two microphones. Because the array used in DREGON is non-linear, angular spectra estimated for this whole array must cover a two-dimensional DOA grid search (azimuth, elevation). For this reason, our implementation of MUSIC is much more computationally demanding than our implementation of GCC-PHAT, which benefits from mBSSL to efficiently aggregate one-dimensional angular estimates. Several implementations of MUSIC exist (see, *e.g.*, [18]). The one used in this study relies on the eigenvalue decomposition (EVD) of the frequency-domain covariance matrix of observed 8-channel signals, and is referred to as EVD-MUSIC.

B. Multichannel Wiener Filtering

The multichannel Wiener filter (MWF) is a classical signal processing technique which can be used to reduce stationary noise from a multichannel signal, see *e.g.* [10]. Based on the observation that UAV-embedded noise signals are roughly stationary during steady parts of the flights (see Section III), we investigated the use of MWF as a pre-propressing step to the source localization methods described in previous section. Let

$$X(f,t) = S(f,t) + N(f,t) \tag{1}$$

where $X(f,t)$, $S(f,t)$ and $N(f,t)$ in $\mathbb{C}^M$ respectively denote the M-channel observed, image-source and noise signals in the short-time Fourier domain and (f,t) is the frequency-time index. The MWF estimate of the source signal is given by $\hat{S}(f,t) = \mathbf{W}_{MWF}(f)X(f,t)$ where

$$\mathbf{W}_{MWF} = \mathbf{R}_{XX}(f)^{-1}(\mathbf{R}_{XX}(f) - \mathbf{R}_{NN}(f)) \tag{2}$$

denotes the optimal Wiener filter and $\mathbf{R}_{XX}(f), \mathbf{R}_{NN}(f) \in \mathbb{C}^{M\times M}$ respectively denote the observed and the noise covariance matrices at frequency f. In practice, $\mathbf{R}_{XX}(f)$ and $\mathbf{R}_{NN}(f)$ cannot be computed exactly and are replaced by their sample estimates. $\hat{\mathbf{R}}_{NN}(f)$ is pre-computed using a noise-only signal of sufficient length, and $\hat{\mathbf{R}}_{XX}(f)$ is directly estimated from multichannel input data. This latter estimation can be done recursively in an efficient way for real-time applications. See [10] for more details.

An 8-channel Wiener filter was applied to the recordings of the DREGON dataset as a pre-prossing step to GCC-PHAT for robust SSL, yielding the WF+GCC-PHAT method. Pre-processing EVD-MUSIC via Wiener filtering is in fact strictly equivalent to the so-called GEVD-MUSIC method [9], which was hence also used in our experiments.

V. Experiments and Results

In this section, the sound localization methods described in previous section are applied to the DREGON dataset and their results are compared and discussed. In all the following

TABLE II: Static sound source localization of 500ms signals at 0 dB SNR. The results are displayed in the format *avg* ± *std* (*fail%*) where *avg* and *std* denote the mean and standard deviations of the mean azimuth and elevation absolute angular errors in degree for successful localizations only ($< 10^\circ$ error), while *fail* denotes the percentage of failures.

Methods / Scenarios	free flight	hovering	rectangle	spin cw and ccw	up and down
GCC-PHAT	$2.43^\circ \pm 1.7^\circ$ (27.8%)	$2.41^\circ \pm 1.5^\circ$ (12.2%)	$2.45^\circ \pm 1.5^\circ$ (7.81%)	$2.36^\circ \pm 1.6^\circ$ (14.2%)	$2.52^\circ \pm 1.6^\circ$ (12.5%)
WF + GCC-PHAT	$2.45^\circ \pm 1.8^\circ$ (1.09%)	$2.31^\circ \pm 1.5^\circ$ (0.00%)	$2.36^\circ \pm 1.6^\circ$ (0.16%)	$2.35^\circ \pm 1.6^\circ$ (0.16%)	$2.41^\circ \pm 1.6^\circ$ (0.47%)
SEVD-MUSIC	$3.01^\circ \pm 2.4^\circ$ (13.6%)	$3.00^\circ \pm 2.4^\circ$ (13.4%)	$3.04^\circ \pm 2.3^\circ$ (23.6%)	$3.03^\circ \pm 2.4^\circ$ (49.8%)	$2.77^\circ \pm 2.4^\circ$ (31.6%)
GEVD-MUSIC	$1.98^\circ \pm 1.9^\circ$ (0.16%)	$2.05^\circ \pm 1.9^\circ$ (3.59%)	$1.93^\circ \pm 1.9^\circ$ (2.19%)	$1.93^\circ \pm 1.9^\circ$ (2.50%)	$1.84^\circ \pm 1.8^\circ$ (2.50%)

experiments, audio signals are downsampled to 16 kHz and the short time Fourier window is fixed to 64 ms with 50 % overlap. The global angular search space has a resolution of 5° in both azimuth and elevation.

A. Static-source localization

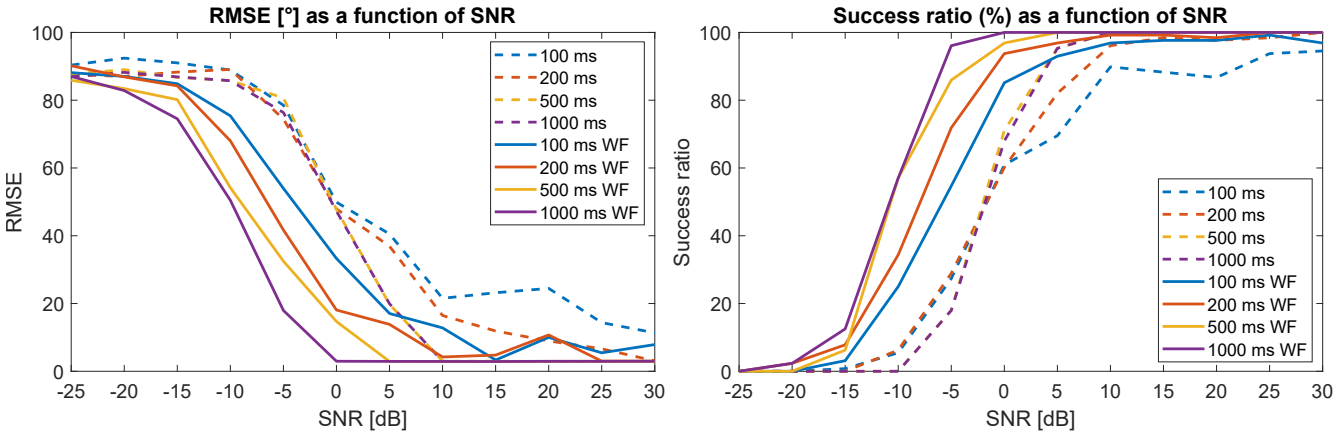
The first experiments use combinations of noise-free static source recordings (Sec. II-D) and noise-only in-flight recordings (Sec. II-C). We limited our source signals to speech emitted from a 2.4 m distance. For each test, a speech segment from a given direction and a noise segments are randomly selected and are weighted and summed to form a test mixture of desired length and SNR. Note that the speech signals used in our experiments are such that they do not contain significant pauses. The signal used to pre-compute the fixed noise covariance matrix of MWF and GEVD-MUSIC (See Sec. IV-B) is the sum of all four individual motor noise signals running at 80 turns/s. Indeed, flight logs revealed that this was the most commonly reached speed in practice.

First, we compare the 4 baseline methods GCC-PHAT, WF+GCC-PHAT, SEVD-MUSIC and GEVD-MUSIC on different flight scenarios. The analysis window, *i.e.*, the duration of every test signal, is fixed to 500 ms. The SNR is fixed to 0 dB in all tests. For every source direction, 40 mixtures are generated from randomly chosen segments, resulting in 640 tests per scenario. For both azimuth and elevation angles, three evaluation metrics are considered: the percentage of *failures*, defined as localization error larger than 10°, and the average and standard deviation of the absolute angular errors of successful estimates. The failure cases are removed from average and standard deviations in order to avoid a too strong influence of outliers. The results are showed in Table II. It can be seen that the mean absolute error of successful localizations is roughly between 2° and 3° for all methods and do not show a high variance, suggesting that all methods share a similar maximal angular resolution. A major difference can be found in the amount of failures. The methods with no noise pre-processing reach failure rates up to 28 % for GCC-PHAT in free-flight and 50 % in SEVD-MUSIC when the UAV spins. On the other hand, GEVD-MUSIC have less then 4 % and WF + GCC-PHAT less than 1 % outliers in all scenarios. These results are very promising. While WF+GCC-PHAT and GEVD-MUSIC show comparable performance, the former is to be preferred. Indeed, it achieved computational times that would easily allow real-time implementations with propper optimization (about 1.5s to process 1s of signal on average), while our implementations of MUSIC-based methods were about twenty times slower using the same hardware, for the reasons detailed in Section IV-A. Because of this computational bottleneck, the following experiments will focus on GCC-PHAT.

In a second experiment, the influence of analysis-window sizes and SNR levels on the GCC-PHAT and WF+GCC-PHAT methods is analyzed. The analysis window sizes varies from 100, 200, 500 and 1000 ms, and the tested SNR values vary from -25 dB to +30 dB in 5 dB steps. Hence, 28 different test conditions are created. To evaluate these, we calculate the *root-mean-squared-error* (RMSE) of estimated directions. Additionally, an SSL result is counted as a success only if both azimuth and elevation errors are lower than 5°, and the success rates are displayed. As can be seen in Fig 4, Wiener pre-processing significantly boosts GCC-PHAT performance in all test conditions. The figure also reveals that the analysis window size has a strong influence on localization accuracy, larger windows systematically implying better performance. At -5 dB, a 96 % success rate is achieved using 1000 ms windows, while this falls to 84 % for 500 ms windows. Nevertheless, this experiment suggests that 500 ms is a good compromise between real-time applicability and accuracy, with a near 100 % success rate for all positive SNRs, and a relatively small decrease in performance compared to 1000 ms. In general, the experiments of this section show that for moderate SNRs (> 0 dB), stationary pre-filtering is sufficient to obtain satisfying localization performance.

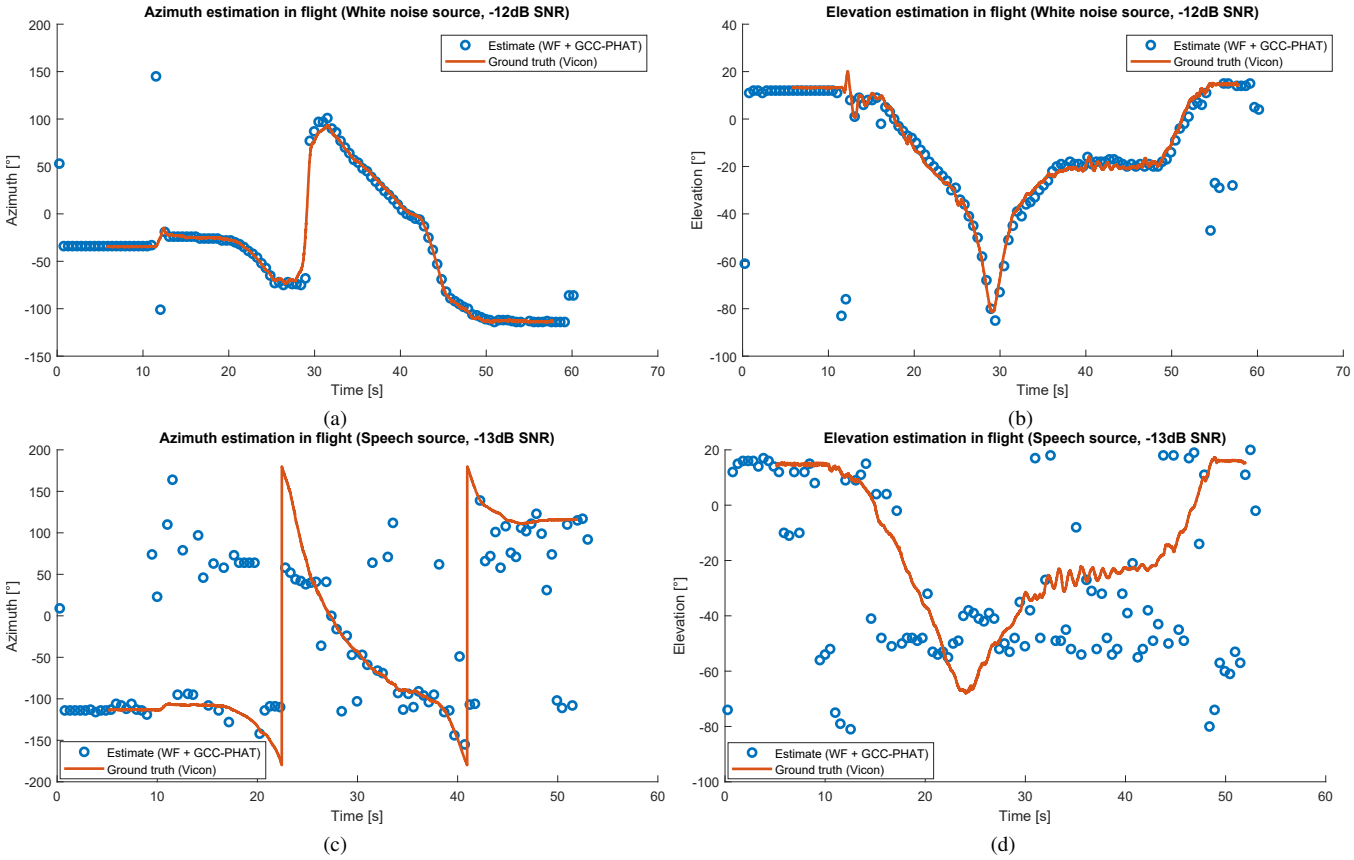
B. In-flight source localization

In this section, the best performing method WF+GCC-PHAT is used on the more realistic scenarios where the UAV flies freely in the presence of a target white noise or speech source, whose range and direction are hence constantly varying in the UAV's frame. A 500 ms sliding analysis window is used over the entire recordings. The Wiener filter noise covariance matrix is estimated from a 7s noise-only recording in free-flight. For the speech scenario, the 3 channels containing heaviest wind-noise are manually removed and the search grid does not includes elevations above 20° to avoid spurious localizations. Fig. 5 displays estimated trajectories against ground truth trajectories of azimuth and elevation angles. As can be seen, the method achieves outstanding performance in white-noise source localization, despite an average SNR of -12dB over this run.



(a) (b)

Fig. 4: (a) Sound localization RMSE in various SNR conditions for different analysis window sizes. Dashed lines correspond to vanilla GCC-PHAT while solid lines shows results with an additional Wiener-filter pre-processing. (b) Same as (a) but for the success ratio, were a localization is considered successful if both azimuth and elevation errors are less than 10°.



(a) (b) (c) (d)

Fig. 5: Target sound source trajectories estimated with WF+GCC-PHAT against ground truth trajectories obtained with the Vicon system during a free flight.

With exception of a few outliers at take off and landing, near-perfect localization is achieved with errors under 2° in both azimuth and elevation for all test frames. However, we observe that performance drastically drops when a speech source is used instead, despite a similar SNR of -13dB. This is consistent with the experiments of previous section, and can be explained by the fact that the energy of speech and UAV noise signals is mainly located in a similar frequency range. These results confirm that more accurate noise models should be developed to tackled speech localization under heavy noise for search and rescue with UAVs. Additional modalities such as motor speed and inertial measurement data could be leveraged for this aim.

VI. Conclusion

This paper presented DREGON, a novel state-of-the-art dataset that aims at pushing forward research in embedded sound source localization with a UAV for search and rescue. The dataset consists of noise-free, noise-only and in-flight audio recordings of a target source emitting speech or white noise signals. The position of the target source in the UAV's frame is precisely annotated in all recordings using a motion capture system. Additional modalities such as inertial measurements and motor speeds acquired by sensors embedded in the UAV as well as video recordings of all flight are included. The dataset as well as a number of baseline sound localization methods applicable to the data are made freely available online for the research community. Our preliminary investigation on the dataset revealed a number of insights on the problem at hand. Notably, the acoustic characteristics of UAV noise make speech localization particularly challenging in very low SNRs. This calls for the development of refined source and noise models. Promising directions are the use of wind noise reduction methods [12] and the use of structured models [15] or extra modalities to predict ego-noise [13], [17].

VII. Acknowledgement

The authors are thankful to François Bodin, Guillermo Andrade-Barroso and Vincent Drevelle of IRISA (Rennes, France) for their help with the hardware used to build this dataset. The authors are also thankful to Lupinenweg. This study has been partially funded by the Fondation Rennes 1.

References

[1] R. D. Andrea, "Guest Editorial Can Drones Deliver?" *IEEE Transactions on Automation Science and Engineering*, vol. 11, pp. 647–648, 2014.

[2] L. Lopez-Fuentes, J. van de Weijer, M. González-Hidalgo, H. Skinnemoen, and A. D. Bagdanov, "Review on computer vision techniques in emergency situations," *Multimedia Tools and Applications*, 2017.

[3] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 4737–4742.

[4] T. Ohata *et al.*, "Improvement in Outdoor Sound Source Detection Using a Quadrotor-Embedded Microphone Array," in *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 1902–1907.

[5] K. Hoshiba *et al.*, "Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments," *Sensors, 17, 2535*, pp. 1–16, 2017.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 32, no. 4, pp. 474–484, 2002.

[8] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, pp. 276–280, 1986.

[9] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 2009–2014.

[10] H. W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Speech Communication; 11. ITG Symposium; Proceedings of*. VDE, 2014, pp. 1–4.

[11] L. Wang and A. Cavallaro, "Microphone-Array Ego-Noise Reduction Algorithms for Auditory Micro Aerial Vehicles," *IEEE SENSORS*, vol. 17, no. 8, pp. 2447–2455, 2017.

[12] M. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind Noise Reduction using Non-Negative Sparse Coding," in *Machine Learning for Signal Processing 17 - Proceedings of the 2007 IEEE Signal Processing Society Workshop, MLSP*, 09 2007, pp. 431 – 436.

[13] K. Furukawa *et al.*, "Noise Correlation Matrix Estimation for Improving Sound Source Localization by Multirotor UAV," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 3943–3948.

[14] T. Morito *et al.*, "Partially Shared Deep Neural Network in Sound Source Separation and Identification Using UAV-Embedded Microphone Array," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 1299–1304.

[15] T. Tezuka, T. Yoshida, and K. Nakadai, "Ego-motion noise suppression for robots based on Semi-Blind Infinite Non-negative Matrix Factorization," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6293–6298.

[16] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal noise suppression for speech recognition by small robots." in *European Conference on Speech Communication and Technology (INTERSPEECH), 2013 IEEE/RSJ International Conference on*, 2005, pp. 2685–2688.

[17] A. Schmidt, A. Deleforge, and W. Kellermann, "Ego-Noise Reduction Using a Motor Data-Guided Multichannel Dictionary," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 1281–1286.

[18] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[19] K. Okutani *et al.*, "Outdoor Auditory Scene Analysis Using a Moving Microphone Array Embedded in a Quadcopter," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 3288–3293.

[20] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, pp. 152–158.

[21] K. Nakamura *et al.*, "Intelligent Sound Source Localization for Dynamic Environments," in *Intelligent Robots and Systems (IROS), 2009 IEEE/RSJ International Conference on*. IEEE, 2009, pp. 664–669.

[22] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Autonomous Robots*, vol. 34, no. 3, pp. 217–232, 2013.

[23] V. Grabe, M. Riedel, H. H. Bülthoff, P. R. Giordano, and A. Franchi, "The telekyb framework for a modular and extendible ros-based quadrotor control," in *European Conference on Mobile Robots, ECMR 2013*, 2013, p. pp. 1925.

[24] J. S. Garofolo *et al.* (1993) Timit Acoustic-Phonetic Continuous Speech Corpus LDC93S1. [Online]. Available: https://catalog.ldc.upenn.edu/ldc93s1.WebDownload. Philadelphia:LinguisticDataConsortium

[25] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[26] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001, pp. 157–180.