Drone audition for search and rescue: Datasets and challenges

Antoine Deleforge

INCEEUROPE

CidB
Centre d'information sur le Bruit

QUIET DRONES
International Symposium
on
UAV/UAS Noise
Paris – 25th to 27th May 2020

Drone audition for search and rescue: Datasets and challenges

Antoine Deleforge, Inria Nancy – Grand Est: antoine.deleforge@inria.fr

Summary

Processing audio signals recorded from a microphone array embedded in an unmanned aerial vehicle (UAV) has received increasing research interest in the recent years and has been referred to as *drone audition*. An important field of application is search and rescue, where humans in disaster areas need to be quickly found. UAVs equipped with high-resolution cameras have already been used in humanitarian responses, while audio-based UAV-embedded sound localisation remains an open research challenge. Microphones could provide a critical complementary modality to vision in situations where visual feedbacks are limited due to bad lighting conditions (night, fog) or obstacles limiting the field of view. This paper provides an overview of the technical and methodological challenges faced by drone audition in the context of search and rescue and presents two publicly available datasets that aim at fostering research in this area. Some localisation and noise-reduction results obtained using baseline methods are also presented. While static localisation of speech sources from a distance of four meters can be efficiently achieved, in-flight localisation from larger distances remains a challenge.

1. 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. Drones have already been used by humanitarian organizations in places like Haiti and the Philippines to map areas after a natural disaster, using high-resolution embedded cameras, as documented in a recent United Nation report [2]. While a number of UAV-embedded tools to address such situations have been developed using video cameras [3], audio-based source localization from UAVs has received relatively less research attention. UAVs equipped with a microphone array could 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 [4]. Sound signals primarily capture two types of information: (i) Geometrical information via the sound propagation path, e.g., the position of emitting sources with respect to the receiving microphone array and (ii) the semantic content of emitted sounds, e.g., speech or an emergency whistle. Methods for retrieving information of the first type include sound source localization or acoustic echo retrieval. Methods for retrieving information of the second type include speech recognition or sound event detection and classification.

Figure 1: Microphones embedded in a drone may help localizing people for search and rescue in disaster area.

In the specific context of drone audition, these traditional audio signal processing (ASP) tasks are made challenging by a variety of effects. One major issue is the noise produced by the UAV itself, generically referred to as *egonoise* in robotics [5]. Due to the quickly changing speed of motors to stabilize the vehicle in the air or to change 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 signal-to-noise ratio (SNR) can easily reach -15 dB or less [6]. Another factor affecting ASP performance is wind noise. The wind is produced by the rotating propellers, the UAV movement in the air and may occur naturally in outdoor scenarios. This wind noise has high power and is of low-frequency. Hence, it easily overlaps with speech signals that typically occur in a similar frequency range [7]. Last, UAV often encounter very dynamical situations. This means that drone-audition methods must be designed to adapt to quickly changing spatial configurations (direction, distance, etc.) environment (presence of reflecting surfaces, wind, etc.) and noise (level, type, etc.). All these challenges need to be tackled at the same time and in near real-time for real world applications 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 (camera, lasers, gyroscope, motor controllers, inertial measurement unit, compass, etc.) may be available. Multimodal approaches fusing information from multiple sensors in order to enhance drone perception and navigation present a promising research avenue, as investigated in [8, 9, 10, 11].

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the recent literature in drone audition, in particular on ego-noise reduction, sound source localization, echolocation and datasets. Section 3 reviews two recently created datasets, namely, the collaborative drone egonoise datasets from the IEEE Signal processing Cup 2019 (SPCup19, [26]) (Section 3.1), and the DREGON dataset [6] dedicated to UAV-based sound localization (Section 3.2). Finally, concluding remarks and outlooks are provided in Section 4.

2. Related Work

2.1 Egonoise Reduction

Egonoise reduction is a topic of interest in robot audition for some years [5]. A first category of methods performs egonoise reduction solely based on audio signals. This includes the work in [12], which compares *unsupervised* methods such as beamforming, blind source separation (BSS) and time-frequency filtering algorithms on UAV-embedded recordings. More recently, the same authors proposed a framework combining time-frequency filtering and BSS for egonoise reduction [13]. It shows very promising performance in a realistic outdoor flight scenario involving a human speaker at a 4 meters distance. Alternatively, the *supervised* methods proposed in [14, 15, 16] rely on pre-recorded noise-only signals to learn a dictionary that represents the egonoise via techniques such as nonnegative matrix factorization (NMF) or K-SVD. This dictionary is then used to model the noise characteristics in noisy recordings and improve its reduction. These methods have been successfully applied to ground robots but not yet to UAVs to the best of our knowledge.

A second category of methods makes use of additional sensors than audio to improve egonoise reduction performance, and in particular, *motor data* such as the rotors' speed or inertial measurements. These include dictionary-based methods [17, 18, 19] as well as Gaussian process [8] or neural-network based approaches [9]. The idea is to predict the spectral characteristics of the egonoise at runtime by learning its relationship to the current motor and/or inertial status of the robot.

2.2 Sound Source Localization

Sound source localization (SSL) is a long standing and extensively studied topics in robotics [20], but is still relatively new in the specific context of drone audition. Many robotic SSL approaches developed in recent years are different variations of the MUltiple SIgnal Classification (MUSIC) algorithm, e.g., [8, 21, 22, 23, 28]. Generalized Cross Correlation (GCC) methods [24] and their variants were also successfully used in robot SSL [4, 25, 6] and are generally less computationally expansive than their MUSIC counterparts. They were notably applied to UAV-embedded SSL in [4, 6], and used as a baseline in the recent SPCup19 drone SSL challenge [26] (see Section 4). An open-source python implementation of this baseline is available at [27]. Reducing ego-noise using the multichannel Wiener filter as a pre-processing step to GCC-based SSL was shown to significantly improve performance in [6].

In [8], 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 to improve robustness to noise of a MUSIC-like method. In the same spirit, [9] 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. In [28], two different UAV microphone array designs are proposed and MUSIC variants are compared on an outdoor SSL scenario. Good SSL success rates are obtained in relatively mild SNR conditions (around 0 dB). The authors propose to adapt the algorithms depending on the considered scenario and emphasize their high computational costs as a drawback for real-time applicability.

2.3 Echolocation for Navigation

Another more recent and emerging research topic in drone audition is the use of acoustic echoes for navigation in closed environments, and in particular for auditory *simultaneous localization and mapping* (SLAM), which is relevant for search and rescue. When sound propagates from a source in an environment, early reflections of the sound on nearby surfaces result in delayed and attenuated copies of the emitted signals at receivers, commonly referred to as *echoes*. The

timings of these echoes contain rich information on the geometry on the environment and can be used to detect obstacles, a principle used in nature by echoing bats to orientate themselves. This principle is referred to as *active echolocation* when a perfectly known and controlled source signal is used and as *passive echolocation* using a partially known or unknown source signal. In the context of UAVs, using a sound emitter placed on the drone (active) or even the drone egonoise itself (passive) is an attractive avenue for navigation in closed-environments where vision cannot be fully relied on. This idea has recently received some research attention. In [29, 30] a perfectly known source placed at the center of a circular array mounted on a drone is used to detect nearby surfaces. The method is validated on simulated data only, but using real drone egonoise. In [31], the same approach is successfully used for a real-world ground-robot navigation demo, where the noise level is much lower than in typical UAV scenarios. Conversely, in [10], cameras, depth sensors and laser sensors are used to identify reflectors in an environment and build a corresponding acoustic model that can localize non-line-of-sight sound sources in a ground-robot navigation scenario.

2.4 Datasets

While the following sections focus on the SPCup-Egonoise [26] and DREGON [6] datasets, at least two other recently published datasets for drone audition are worth mentioning. The AIRA-UAS dataset was captured indoor with an 8-channel circular microphone array mounted on three types of drones [32]. The recording drone is flying either alone or in the presence of other drones. It aims at evaluating UAV-embedded sound source localization and separation methods, which could be applied to search-and-rescue but also to unauthorized drone operation detection. Also of interest is the AVQ dataset, consisting of outdoor, synchronized audio-visual recordings from a flying drone equipped with an 8-microphone-array as well as a camera [11]. The dataset includes scenarios 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.

3. Drone Audition Datasets

3.1 The SPCup-Egonoise dataset

The IEEE Signal Processing Cup is a yearly international competition involving teams of undergraduate students. The 6th edition (SPCup19) took place from November 2018 to May 2019 and was focused on UAV-embedded sound localization for search and rescue [26]. On top of the main required SSL tasks, a bonus task asked participants to gather their own sound recordings using microphones mounted on a UAV. Eleven teams participated to this task, which resulted in the *SPCup-Egonoise dataset*, now publically available online for research purpose at [33]. The dataset includes recordings using 1- to 16-channel microphone arrays and features a variety of drone model and array geometries. The diversity of this dataset is illustrated in Figure 2. As can be seen, noise levels variy widely depending on the setup. Team *Idea! SSU* placed their sensors the farthest away from the UAV propellers, resulting in relatively mild low frequency noise, while recordings from teams *ChuMS* and *Maverick* feature clipping (microphone saturation) due to extreme noise loudness. It can also be observed that different drone models feature different harmonic comb patterns, from very distinct and spectrally spread patterns (teams *KU Leuven* and *NSS Chellamma*) to less distinct ones (team *Shout COOE*), through sparser spectra (team *AGH*) or lower frequency ones (team *Diagonal Unloading*).

Common to all these recordings is the presence of recognizable patterns that characterize drone egonoise. These correspond to wind (low frequency, sporadic bursts), propeller rotations (harmonic combs) and other electronic and mechanical sources (random stationary background). These patterns are better seen in Figure 4, showing the spectrogram of a speech source recorded from a flying UAV [6]. As can be seen, wind noise is particularly detrimental to speech, as it lies within the same frequency range and is very loud. However, it rarely occurs in all

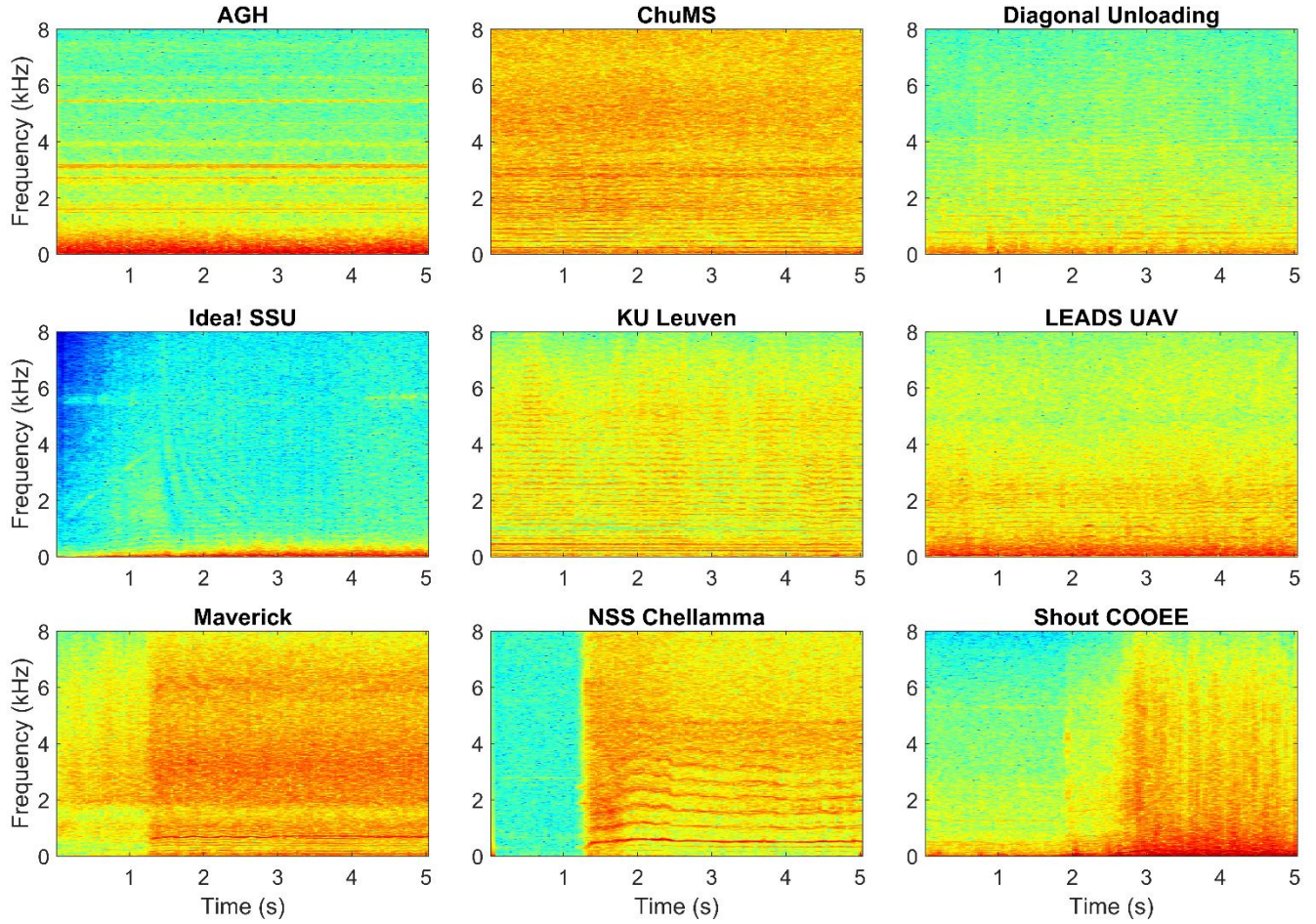

Figure 2: Example spectrograms of 5-second excerpts of egonoise recordings made by different teams that participated to the IEEE Signal Processing Cup 2019 [26]. The same color scale from blue to red is used to represent log-magnitudes (dB) in each spectrogram.

channels at the same time, suggesting the use of wind noise detection for adaptive channel selection. Wind noise reduction using NMF has also been explored in [7]. The random stationary background components are the easiest to remove using statistical methods such as the multichannel Wiener filter (MWF), as showed, e.g., in [6]. MWF can also be used to efficiently remove harmonic components, assuming that the harmonic pattern is known and constant over time. While this assumption is reasonable for stationary or constant-velocity flights, it is no longer valid in scenarios that are more dynamic, with numerous changes of velocities and directions. For such situations, a promising avenue is to use motor-speed data to predict harmonic patterns, as illustrated in Figure 5 and as explored in [42] on a ground robot.

3.2 The DREGON Sound Localization dataset

Figure 3: DREGON setup.

The DREGON dataset was published in 2018 [6] with the aim of fostering research in UAV-embedded SSL for search-and-rescue applications. It is publicly available online at [34]. The setup used is a MikroKopter quadrotor UAV equipped with a cube-shaped 8-microphone array mounted under the drone via a 3D-printed structure, as depicted in Figure 3. The dataset contains both noisy (motors on) and clean (motors off, hand-held) in-flight and static audio recordings. These are continuously annotated with the 6 degrees-of-freedom placements of both the target source and microphone array using a precise VICON motion capture system.

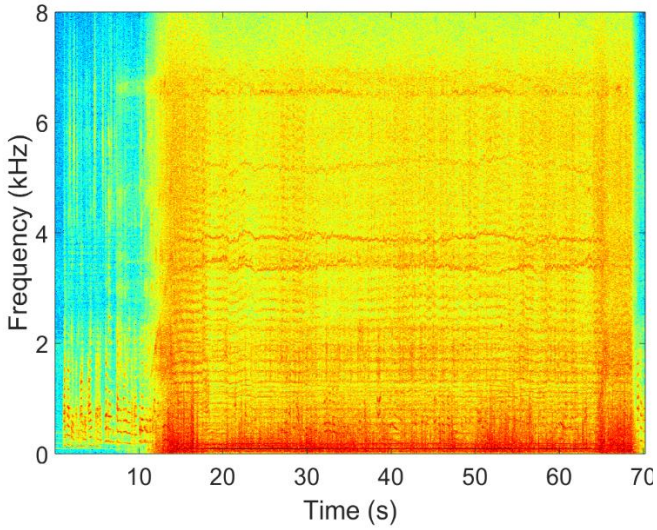

Figure 4: Spectrogram of the recording of a speech source by a microphone embedded in a drone [6]. First the drone is idle, then the motors are powered on and the drone takes off.

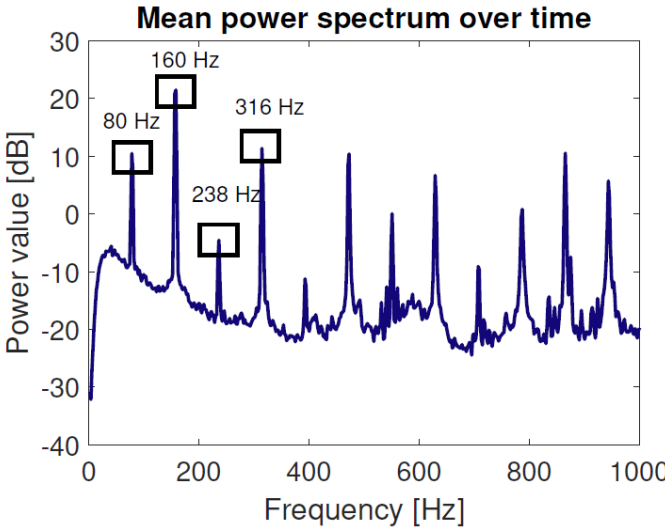

Figure 5: Mean power spectrum of the sound generated by an individual drone rotor spinning at 80 rotations per second [6]. Peaks at harmonics proportional to 80 Hz can be observed.

Different flight patterns featuring varying dynamics are available. The target source is a static loudspeaker on the ground, emits either speech or white (broadband) noise, and is placed 2 to 4 meters from the UAV. In addition to audio and position signals, the rotational speeds of each of the four individual rotors as well as inertial measurements data are available at all time, and all signals are time-stamped for synchronization. The dataset features in-flight recording sessions in two different rooms with respective volumes 10m x 10m x 2.5m and 12m x 12m x 3.5m and mild reverberation times (RT60 < 200ms).

Figure 6 shows some of the sound source localization results obtained on the dataset using GCC-PHAT [24] after pre-processing the signals with multichannel Wiener filtering (MWF). The noise statistics used for MWF were pre-computed from recordings of each individual rotors spinning at 80 rotations per second, and were fixed over time. As can be seen, very satisfactory SSL results can be obtained with this approach when the emitting sound source is broadband, even in harsh noise conditions (SNR < -10 dB). However, when the emitted signal is speech, much poorer results are obtained under similar SNR, in particular when speed and position change. This can be explained by the strong spectral overlap between speech, wind, and propeller noises, and suggest that a fixed MWF does not sufficiently reduces these noises for SSL in these conditions (See Section 3.1).

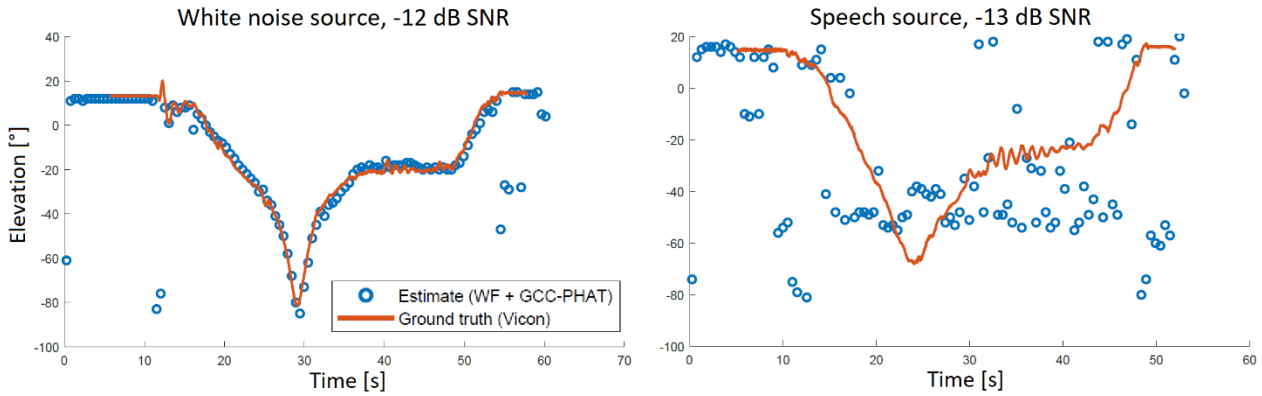

Figure 6: Estimated elevation angles of a static, ground sound source recorded from a flying drone using multichannel Wiener filtering (WF) followed by GCC-PHAT, versus ground truth [6]. Left: the source emits white noise. Right: the source emits speech.

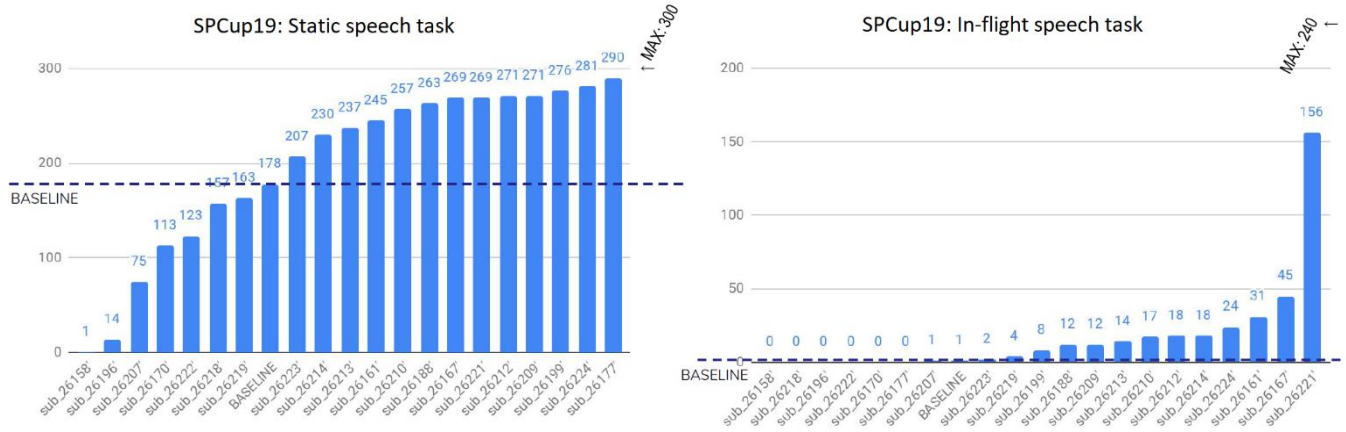

Figure 7: Results obtained by the 20 student teams that participated to the SPCup19 [26]. Left: number of correctly localized static speech sources from a stationary drone (angular error <10°). Right: number of correctly localized speech source from a flying drone.

Finally, some of the results obtained on the SSL tasks of the SPCup19, which were based on a subset of DREGON, are shown in Figure 7. Note that the baseline used for this challenge was also GCC-PHAT, but without Wiener pre-filtering. Here, it can be observed that while near-perfect results are obtained by several teams on the static speech localization task, the in-flight speech localization task proves much more challenging, despite similar SNR conditions. The best performing team nevertheless achieves an encouraging 65% localization accuracy on that realistic task. Crucially to this success, the team used Kalman filtering to smooth estimated trajectories, pre-filtered signals using an adaptive MWF, fused localization estimates of both MUSIC and GCC-PHAT methods via k-mean clustering, and processed the angular spectra obtained from these methods using handcrafted heuristics.

Amongst the top 12 performing teams, the most popular pre-filtering method was MWF, and the most popular SSL method was GCC-PHAT. About half of the team used motor speeds provided with audio signals to estimate noise characteristics and about half used some form of angular trajectory smoothing and post-processing. Other notable ideas included the use of a speech activity detector or channel selection to reduce the effect of wind noise. Interestingly, only two teams made used of machine learning techniques such as deep neural networks, probably owing to the limited amount of data available for training. The newly released SPCup19-egonoise dataset could constitute a first step towards employing such approaches in the future.

4. Conclusion

In this paper, we have explored some of the numerous challenges and research opportunities brought by the recent and emerging topic of drone audition. An important application for this new field is audio-based search and rescue, which involves a large number of audio signal processing tasks, from sound source localization to echolocation, from noise reduction to sound event and speech detection and recognition. While many of these tasks have long been studied, the specific setting of drone audition, involving compact arrays near loud noise sources and very dynamic soundscapes make them particularly challenging. Most of these problems remain open as of today, but emerging research efforts in noise reduction and source localization along the past few years have shown promising results, notably fostered by the recent development of publicly available datasets.

References

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

[2] U.N.O. for the Coordination of Humanitarian Affairs (2014), "Unmanned aerial vehicles in humanitarian response," [Online].

[3] L. Lopez-Fuentes, J. van de Weijer, M. Gonzalez-Hidalgo, H. Skinnemoen and A. D. Bagdanov (2017), “Review on computer vision techniques in emergency situations,” Multimedia Tools and Applications.

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

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

[6] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge (2018), “Dregon: Dataset and methods for uav-embedded sound source localization,” in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[7] M. Schmidt, J. Larsen, and F.-T. Hsiao (2007), "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, pp. 431 – 436.

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

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

[10] I. An, M. Son, D. Manocha, and S.-e. Yoon (2018), “Reflection-Aware Sound Source Localization,” in IEEE International Conference on Robotics and Automation (ICRA), pp. 66–73, IEEE

[11] L. Wang, R. Sanchez-Matilla and A. Cavallaro (2019), "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5320-5325.

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

[13] L. Wang and A. Cavallaro (2020), "A blind source separation framework for ego-noise reduction on multi-rotor drones," in IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[14] A. Deleforge and W. Kellermann (2015), "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 355-359).

[15] N. Mae, M. Ishimura, S. Makino, D. Kitamura, N. Ono, T. Yamada and H. Saruwatari (2017), "Ego noise reduction for hose-shaped rescue robot combining independent low-rank matrix analysis and multichannel noise cancellation," in International Conference on Latent Variable Analysis and Signal Separation (pp. 141-151). Springer, Cham.

[16] T. Haubner, A. Schmidt and W. Kellermann (2018), "Multichannel Nonnegative Matrix Factorization for Ego-Noise Suppression," in Speech Communication; 13th ITG-Symposium, pp. 1-5.

[17] A. Deleforge, A. Schmidt and W. Kellermann (2019), "Audio-motor integration for robot audition," Multimodal Behavior Analysis in the Wild (pp. 27-51). Academic Press.

[18] A. Schmidt, A. Deleforge and W. Kellermann (2016), "Ego-noise reduction using a motor data-guided multichannel dictionary," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[19] A. Schmidt, H. W. Löllmann and W. Kellermann (2018), "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[20] C. Rascon and I. Meza (2017), “Localization of sound sources in robotics: A review,” Robotics and Autonomous Systems, vol. 96, pp. 184–210.

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

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

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

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

[25] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud (2013), “The ManyEars open framework,” Autonomous Robots, vol. 34, no. 3, pp. 217–232.

[26] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel and L. Marcenaro (2019), "Audio-Based Search and Rescue With a Drone: Highlights From the IEEE Signal Processing Cup 2019 Student Competition [SP Competitions]," in IEEE Signal Processing Magazine, vol. 36, no. 5, pp. 138-144.

[27] https://github.com/Chutlhu/SPCUP19

[28] K. Hoshiba et al. (2017), “Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments,” Sensors, 17, 2535, pp. 1–16.

[29] U. Saqib, J. R. Jensen (2019), "Sound-based Distance Estimation for Indoor Navigation in the Presence of Ego Noise", 27th European Signal Processing Conference (EUSIPCO), 1-5.

[30] U. Saqib, S. Gannot, J. R. Jensen (2020), "Estimation of acoustic echoes using expectation-maximization methods", EURASIP Journal on Audio, Speech, and Music Processing (1), 1-15.

[31] U. Saqib and J. R. Jensen (2020) "A Model-based Approach to Acoustic Reflector Localization with a Robotic Platform," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) [Accepted].

[32] O. Ruiz-Espitia, J. Martinez-Carranza and C. Rascon (2018), "AIRA-UAS: an Evaluation Corpus for Audio Processing in Unmanned Aerial System," in International Conference on Unmanned Aircraft Systems (ICUAS), pp. 836-845.

[33] http://dregon.inria.fr/datasets/the-spcup19-egonoise-dataset/

[34] http://dregon.inria.fr/datasets/dregon/