2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

DroneEARS: Robust Acoustic Source Localization with Aerial Drones

Prasant Misra, A. Anil Kumar, Pragyan Mohapatra, and Balamuralidhar P.

***Abstract*— Micro aerial vehicles (MAVs), an *emerging* class of aerial drones, are fast turning into high value mobile sensing assets. While MAVs have a large sensory gamut at their disposal; vision continues to dominate the external sensing scene, with *limited* usability in scenarios that offer acoustic clues. Therefore, we endeavor to provision a MAV auditory system (i.e., *ears*); and as part of this goal, our preliminary aim is to develop a *robust acoustic localization* system for detecting sound sources in the physical space-of-interest. However, devising this capability is *extremely* challenging due to strong *ego-noise* from the MAV propeller units, which is both wideband and non-stationary. It is well known that beamformers with large sensor arrays can overcome high noise levels; but in an attempt to cater to the platform (i.e., space, payload and computation) constraints of a MAV, we propose DroneEARS: a binaural sensing system for geo-locating sound sources. It combines the *benefits* of sparse (two elements) sensor array design (for meeting the platform constraints), and our proposed mobility-aided beamforming (for overcoming the severe ego-noise and its other complex characteristics) to significantly enhance the received signal-to-noise ratio (SNR). We demonstrate the efficacy of DroneEARS by empirical evaluations, and show that it provides a *SNR improvement* of $15 - 18$ dB compared to many conventional and widely used techniques. This SNR gain translates to a source *localization accuracy* of approximately 40 cm within a scan region of $6\,\text{m} \times 3\,\text{m}$, that is, *one* order of magnitude better than competing methodologies.**

I. Introduction

Aerial drones are *fast* emerging as a powerful new class of mobile sensing systems, with a *rich sensory* gamut neatly combined with on-board communication and computing elements. In addition, the micro-multirotor-hovering class of aerial drones provide remarkable *mobile agility* [1] to maneuver across any type of terrain and space (indoor/outdoor). This new class is referred to as micro aerial vehicles (MAVs), and it is increasingly finding usage across a *wide* range of applications ranging from industrial to humanitarian.

Vision is the *most* promising mode of sensing in MAVs [2]. Compared to other spatially rich sensors such as sonar and laser range finders (that are commonly used in terrestrial vehicles), vision sensors require comparatively lower energy to interrogate the environment; and for a comparable mass, they can gather richer information and span wider fields of view [3]. While the advantages of drone vision are clearly obvious; on the flip side, it becomes *unusable* in camera obstructed, occluded conditions, or in scenarios that offer profound acoustic clues (such as in emergency response and search operations). Fundamentally, scene analysis (such as source localization) using acoustic clues should be possible even if visual detection is hindered. Acoustic sensing is, therefore, *beneficial* in a variety of MAV supported applications, and will *enhance* its operational capability by offering a multi-modal sensing experience.

Problem context and challenges. MAV-based acoustic sensing offers *significant* hurdles in its effective realization, and is primarily due to very strong *ego-noise* that masks the sound sources. The ego-noise is generated mainly by the propeller units, and is *extremely strong* due to its close vicinity to the acoustic recorders than the sound source itself. It is also *non-stationary* due to rotation speed change of individual motors during flight, and has a *wide* dynamic range and frequency span. All of these factors lead to an *extremely low signal-to-noise ratio* (SNR) condition.

There are numerous citations in existing literature that have tackled similar problem statements. It is well understood that noise levels of high severity can be efficiently compensated through an acoustic array with a large number of sensing elements along with sophisticated beamforming techniques [4], and this combination can finely focus the received power - in the form of a narrow beam - to a specific point-of-interest in the physical sensing space. However, catering to such exhaustive system and computational requirements within the space, size, payload, and energy constraints of a MAV is not only tedious, but also *extremely* challenging. Motivated by the need to overcome these challenges, and to provision an efficient acoustic sensing system for MAVs, we propose DroneEARS.

Contributions. DroneEARS is a *binaural* sensing system designed for MAVs to *locate* sound sources, an important auditory pre-requisite and cue for spatial awareness. This *new* computing framework combines the benefits of our *proposed* sparse sensor array model (for meeting the platform constraints), and mobility-aided beamforming (for overcoming the severe ego-noise); and enhances the SNR of the received signal by $15 - 18$ dB compared to conventional techniques using an equivalent set of sensor elements. Our empirical evaluations show that its localization accuracy of estimating the sound source is approximately within 40 cm for a scan region of $6\,\text{m} \times 3\,\text{m}$, which is *one* order of magnitude better than other competing techniques.

The remainder of this paper is organized as follows. We elaborate on the design of DroneEARS in Section II, present its evaluation in Section III, and summarize our work with concluding remarks in Section IV.

All authors are with Embedded Systems and Robotics, TCS Research & Innovation, Bangalore, India. Email: {prasant.misra, achannil.kumar, mohapatra.pragyan, balamurali.p}@tcs.com

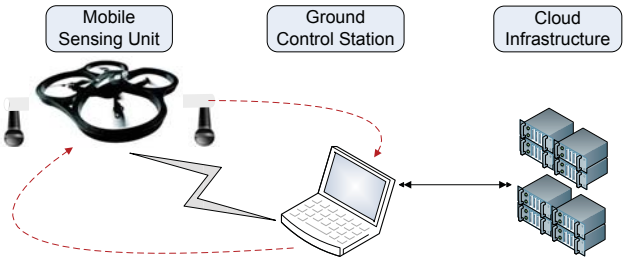


Fig. 1: **Architecture of DroneEARS.**

II. The Design of DroneEARS

To ground our discussion, we explain the architecture and MAV platform details to better motivate the design choices made regarding the sparse acoustic array and beamforming methodology, introduced in the latter half of this section.

System architecture. The system is composed of *three* units, as represented in Fig. 1. The *mobile sensing unit* consists of a MAV with a wide range of onboard sensors. The primary responsibility of this unit is to sense the physical space, gather and transfer back the raw data to the ground control station (GCS). The *GCS unit* is a standard computer that communicates with the mobile sensing unit using a long-range radio. GCS allows the pilot to configure mission parameters. Besides that, it acts as an interface to transfer data sent by the mobile sensing unit to the *cloud infrastructure unit*. This backend forms the final system unit and command/control centre, where raw data and low-level inferences are permanently stored, curated, processed, and fused with archival data for more fine-grained insights.

Platform details. We use the popular AR.Drone 2.0 quadcopter, developed by Parrot S.A., as the mobile sensing unit. As the name *quadcopter* implies, this aircraft consists of *four* propellers unit situated in the same horizontal plane located around its body. They provide lift, stability and directional control, and lateral movement to allow the aircraft to propel itself forward in any direction. Each propeller unit consists of a 10 cm long blade mounted on a 15 W/28500 revolutions/minute (RPM) brushless motor. We refer our readers to [5] for additional details regarding this platform.

Analysis of ego-noise. Our aim is to provision an *acoustic sensing system* on this quadcopter unit to localize sound sources. However, the combination of the mechanical noise (generated by the rotation of the motors) and the wind noise (generated by the immense flow of air into the recording system during propeller rotation and aircraft motion) - referred to as *ego-noise* collectively - makes the platform very loud to gather any usable acoustic information.

There have been few studies in the recent past that have analyzed the spatial and spectral characteristics of ego-noise [6]. The spatial characterization studies have concluded that wind noise can significantly deteriorate the recording quality; and that the positions below the MAV receive more noise than positions that are above or beside it due to the downward flow of wind from the propellers. The spectral characteristics exhibit the traits of : strong power (especially, below 3 kHz), wide bandwidth (100 Hz to 18 kHz), and non-stationarity (induced due to the movement of the MAV, and the change in rotation speed of the four motors).

Multi-channel techniques, which make use of sensor arrays to sample the signal in a spatio-temporal manner and perform beamforming by exploiting the phase and correlation properties, are effective in source localization. Fixed beamformers are robust to low SNR levels, but usually need a large sensor array. Multi-channel, adaptive beamformers overcome the native limitations of fixed beamformers to a great extent; but may still display limited gain under mobile conditions, where the acoustic mixing network changes dynamically. In the next section, we explore the design basics of sensor arrays and beamformers to arrive at an optimal make for MAVs.

A. Overview of DroneEARS

Design rationale: With regards to beamforming, the configuration and built (i.e., geometry) of the sensor array greatly impacts the beam characteristics. While the array geometry can range from linear forms to complicated spirals, we resort to an uniform linear array (ULA) due to its simpler design and working principles, which complement the MAV platform constraints.

An ULA arranges identical sensor elements along a line with uniform spacing. It is *desirable* to have : (i) a larger aperture length A (i.e., distance between the first and the last array element, where distance is measured in terms of wavelength λ) to obtain a narrower beam that can better resolve sound sources, and (ii) the distance B between successive elements of the sensor array to be less than $\lambda/2$ to avoid the undesirable grating lobes that results due to spatial aliasing. Although A can be enlarged by embedding more array elements, this approach may not be feasible for our target platform due to system and computational limitations, as well as restrictions in deployment and storage. Alternately, in view of the platform constraints, the dimensions of A can be increased by using fewer array elements, but with a larger B. While such a *sparse array* design is more durable for MAVs, it invariably introduces *beamforming ambiguities* due to spatial aliasing. Mobility plays a *vital* role in alleviating the combined challenge of lowering the computation load to a desirable extent, and provisioning a larger A with reduced grating lobes.

Motivating insights: The DroneEARS system performs an exhaustive search for the sound source in the the target region. For performing an efficient search operation, it first divides the search space into a grid of predefined cell sizes. It uses the mobility of the MAV to beam the power to each cell at different measurement locations and at finer frequency resolutions; and subsequently, consolidates all the beamed power to estimate the source location. The illustrative example shown in Fig. 2 depicts this high-level operation.

We apply *two* key observations in this regard. *First*, each independent measurement made at a particular spatial point and frequency sub-band will produce a beam pattern, where the main lobe can be made to point to the location of the sound source with multiple grating lobes pointing to

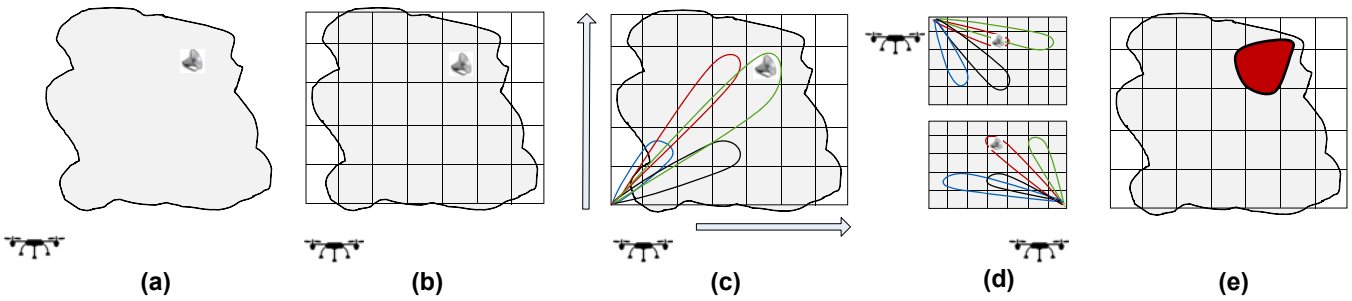


Fig. 2: **The high-level operation view of DroneEARS.** *Note: the speaker symbol in* (a)-(d) *denotes the original location of the sound source, and the red blob in* (e) *represents its estimated location.*

other noisy locations; but an efficient fusion across all such spatially separated measurement points and frequency sub-bands will ultimately subdue the unwanted grating lobes. *Second*, even if B is chosen in a manner such that the sensor array is sensitive to a specific narrowband frequency range, the wide spectrum of the received signal can be robustly handled through sub-banding their individual frequency responses and fusing them efficiently; which further helps in reducing the power of the grating lobes. While the wideband nature of the signal is used for aperture extension, *mobility combined with the signal's wide bandwidth is used for removing the grating lobes*. This new manner of consolidating intra-band and inter-measurement fusion significantly enhances the overall SNR of the received signal, and hence, improves the source detection and localization accuracy. It also increases the degrees-of-freedom (DOF) beyond a single source to reliably *localize multiple sound sources in the search space, if there is good spatial separation among them*.

B. Details of DroneEARS

1) Signal Model: We consider an ULA comprising of M channels indexed by $m = \{1, 2, ..., M\}$, and equally spaced on a line with an inter-element distance of B. It is mounted on a MAV, and receives signals radiated by Q wideband stationary sound sources (where, $Q < M$). The sources are assumed to be located in the far field. We denote the $D \times 1$ position of the $\mathbf{q}^{th}$ transmitter as $\mathbf{p}_q$. Since an ULA can only resolve sources in 2D, we assume D = 2. The MAV is supposed to take measurements at the l^{th} position for a total of L different locations, where $l = \{1, 2, ..., L\}$.

The complex signal envelope observed by the m^{th} sensor element at the l^{th} position is expressed as:

$$r_{m,l}(t) = \sum_{q=1}^{Q} e^{j2\pi\tau_{m,l}(\mathbf{p}_q)} s_q(t - \tau_{m,l}(\mathbf{p}_q)) + \eta_{m,l}(t) \quad (1)$$

$\eta_{m,l}(t)$ denotes the additive white noise in Eq. (1), while $\tau_{m,l}(\mathbf{p}_q)$ represents the delay of the m^{th} sensor with respect to a reference sensor for the transmitter located at $\mathbf{p}_q$. The delay $\tau_{m,l}(\mathbf{p}_q)$ depends on the array geometry. For an ULA $\tau_{m,l}(\mathbf{p}_q) = \mathbf{x}_l^T \vec{\mathbf{u}}_{m,q}$, where $\mathbf{x}_l$ denotes the position of the l^{th} sensor with respect to the reference sensor, and $\vec{\mathbf{u}}_{m,q}$ denotes the unit vector pointing along the direction of the q^{th} transmitter from the m^{th} measurement location [7].

We make the following assumptions in this signal model.
• *Assumption 1* : The number of transmitters are less than the number of elements in the sensor array (i.e., $Q < M$)[1].
• *Assumption 2* : The Q transmitters are stationary, and they continuously transmit for the entire measurement duration.
• *Assumption 3* : The precise location of the MAV is known at every measurement point.

2) Problem Formulation: Given the above signal model and assumptions, the goal is to : *estimate the location of* $\boldsymbol{p}_q$ *transmitters for* $q = \{1, 2, ..., Q\}$, subject to the constraints of wideband acoustic signal and noise sources, low SNR of the received signal, and low processing complexity.

3) Proposed Approach: Since the signal is wideband in nature, the received signal is decomposed into narrow sub-bands using a non-overlapping filter bank. Let us denote the i^{th} sub-band of signal $s_q(t)$ by $s_q^{(i)}(t)$, where $i = \{1, 2, ..., N_s\}$ and N_s represents the number of sub-bands (and therefore, the number of filters). By using the standard narrowband assumption, $s_q^{(i)}(t - \tau_{m,l}(\mathbf{p}_q))$ can be approximated as $s_q^{(i)}(t)$. Therefore, the signal observed at the l^{th} measurement position corresponding to the i^{th} sub-band can now be expressed as :

$$\mathbf{r}_l^{(i)}(t) = \sum_{q=1}^{Q} \mathbf{a}_l^{(i)}(\mathbf{p}_q) s_q^{(i)}(t) + \eta_l^{(i)}(t) \quad (2)$$

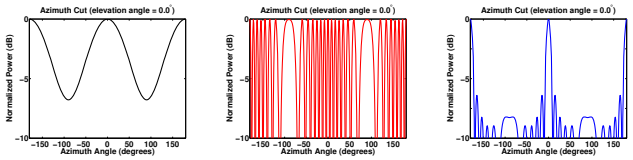
where $s_q^{(i)}(t)$ represents the i^{th} sub-band of the q^{th} source, $\mathbf{a}_l^{(i)}(\mathbf{p}_q)$ denotes the array response at the l^{th} measurement location for the i^{th} sub-band for the signal source located at $\mathbf{p}_q$, and $\mathbf{r}_l^{(i}(t)$ is a $M \times 1$ vector. The above expression can be expressed in the matrix form as :

$$\mathbf{r}_l^{(i)}(t) = \mathbf{A}_l^{(i)} \mathbf{s}^{(i)}(t) + \eta_l^{(i)}(t) \quad (3)$$

where $\mathbf{A}_l^{(i)} = [\mathbf{a}_l^{(i)}(\mathbf{p}_1), ..., \mathbf{a}_l^{(i)}(\mathbf{p}_Q)]$ and $\mathbf{s}^{(i)} = [s_1^{(i)}, ..., s_Q^{(i)}]^T$. It is important to note that the matrix $\mathbf{A}_l^{(i)}$ shall not be known in advance. In fact, Eq. (3) can be viewed as a parameter estimation problem, where the goal is to estimate the parameters $\mathbf{P} = \{\mathbf{p}_q\}_{q=1}^{Q}$.

Using all the above N_s equations is a *must* to obtain a good localization accuracy. However, it is important to

[1]Note that when the transmitters are spatially well separated, as shall be demonstrated in Section III, the proposed method is capable of even localizing $Q > M$ sources.



(a) Lower frequency beam 500 Hz (b) Higher frequency beam 5 kHz (c) Intra-band beam fusion

Fig. 3: **Rationale of intra-band beam fusion.** (a) & (b) *show the beams corresponding to* 500 *Hz (single beam at* 0^o*, but of wider width) and* 5 *kHz (narrow beam, but with grating lobes at* 180^o *that spreads into* -90^o *and* $+90^o$*) respectively.* (c) *The intra-band fusion of Eq. (6) leads to a relatively narrow beam, but with significant reduction in grating lobes. In combination with the inter-measurement fusion of Eq. (7), the robustness of the estimation process improves significantly.*

note that for a given inter-sensor spacing B; using those sub-bands that satisfy Nyquist criterion $B < \lambda/2$ yield good estimates, while erroneous results are obtained for those sub-bands that do not satisfy this condition. It can be overcome by the method stated by Tang et al. [8], where such a parameter estimation problem is casted into a sparse representation framework, and is solved using a convex optimizer. However, this approach is computationally demanding, and will enormously intensify when combined for all L measurement locations. We make use of the sparse representation framework, but propose a simple yet efficient estimation method.

Sparse representation framework. The locations of the smaller cells in the scan region are denoted by $\tilde{\mathbf{P}} = \{\tilde{\mathbf{p}}_1, \tilde{\mathbf{p}}_2, ..., \tilde{\mathbf{p}}_{N_p}\}$, where N_p denotes the number of smaller cells and $N_p \gg Q$. Assuming $\mathbf{P} \subset \tilde{\mathbf{P}}$, Eq. (3) can be casted into the sparse representation framework as :

$$\mathbf{r}_l^i(t) = \tilde{\mathbf{A}}_l^{(i)} \tilde{\mathbf{s}}^{(i)}(t) + \eta_l^{(i)}(t) \tag{4}$$

where $\tilde{\mathbf{A}}_l^{(i)} = [\mathbf{a}_l^{(i)}(\tilde{\mathbf{p}}_1), \mathbf{a}_l^{(i)}(\tilde{\mathbf{p}}_2), ..., \mathbf{a}_l^{(i)}(\tilde{\mathbf{p}}_{N_p})]$, and is of size $M \times N_p$. $\tilde{\mathbf{A}}_l^{(i)}$ bears similarity to $\mathbf{A}_l^{(i)}$. In this representation, the k^{th} row of the sparse vector $\tilde{\mathbf{s}}^{(i)}(t)$ will be non-zero and will be equal to $s_k^{(i)}(t)$, if a source signal is present at location $\mathbf{p}_k$. It is important to note that Eq. (4) is for the measurement at one single location l and one sub-band i. In the rest of this section, we propose a computationally efficient method using this framework.

Intra-band and Inter-measurement fusion. The entire region-of-interest is divided into smaller cells, and the power of each cell corresponding to each sub-band and from each measurement location is first estimated. This is performed by beamforming [4] and by appropriately fusing the individual power measurements, such that the power in the respective cell where the target signal is present gets a boost compared to other empty cells. For the l^{th} measurement location and corresponding to the i^{th} sub-band, the power corresponding to the $\tilde{\mathbf{p}}_k$ cell location is estimated as :

$$\Gamma_{l,\tilde{\mathbf{p}}_k}^{(i)} = \frac{1}{(\mathbf{a}_l^{(i)}(\tilde{\mathbf{p}}_k))^H (R_l^{(i)})^{-1} \mathbf{a}_l^{(i)}(\tilde{\mathbf{p}}_k)}. \tag{5}$$

where the covariance matrix $R_l^(i) = \sum_{t=1}^{T} \mathbf{r}_l^i(t)(\mathbf{r}_l^i)^H(t)$, and T denotes the length of the snapshots.

In the first fusion step, referred to as *intra-band fusion*; for a given measurement location l, the power corresponding to each cell location across N_s sub-bands is fused as :

$$\Delta_{l,\tilde{\mathbf{p}}_k} = \prod_{i=1}^{N_s} \Gamma_{l,\tilde{\mathbf{p}}_k}^{(i)} \tag{6}$$

Recall from the previous discussion that the sub-band corresponding to $B \leq \lambda/2$ yields a single but a wider beam due to smaller aperture length (Fig. 3(a)); while the sub-band corresponding to the $B > \lambda/2$ yields a comparatively narrower beam due to larger aperture length, but produces many undesirable grating lobes due to spatial aliasing (Fig. 3(b)). By using the above intra-band fusion across sub-bands; not only do the undesired grating lobes significantly reduce, but it also results in a comparatively narrower beam. Fig. 3(c) shows the beam pattern obtained after fusion; where we notice the reduction of grating lobes, and a narrower beamwidth compared to Fig. 3(a) and Fig. 3(b) respectively.

In the second fusion step, referred to as *inter-measurement fusion*, the power across the L measurement locations are fused; and the final aggregated power corresponding to the $\tilde{\mathbf{p}}_k$ location is estimated as :

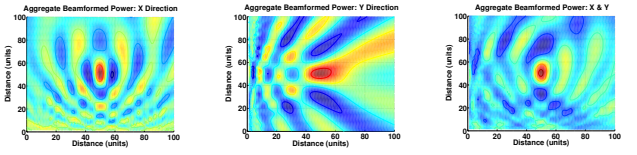
$$W_{\tilde{\mathbf{p}}_k} = \sum_{l=1}^{L} \Delta_{l,\tilde{\mathbf{p}}_k} \tag{7}$$

As a final outcome, the power corresponding to the cell where the sound source is present increases compared to the beamforming power corresponding to individual measurement location; with significant reduction in the power levels of the unwanted grating lobes and noisy peaks (Fig. 4). Thus, it not only overcomes spatial aliasing, but also improves the overall computation efficiency by estimating the location of the sound source(s) (Fig. 5) in a single processing step.

III. Evaluation

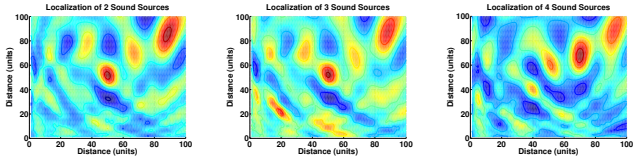
In this section, we evaluate the performance quality and limitations of the DroneEARS framework. Empirical data was collected by recording the noise and the sound source signal separately, and summing the two signals at different input SNRs that vary from -25 dB to 25 dB.

Array configuration. The audible acoustic frequency range is from 20 Hz-20 kHz. However, most common sound sources in the physical world have a frequency of less than 5 kHz (i.e., $\lambda = 6.6$ cm). We choose a binaural (i.e., two element) sensor array with a large aperture in order to meet the MAV platform constraints with reasonable degree of array efficiency. The aperture A of the array can be widened by considering a large B between the two sensing elements. In fact, for an ULA of two sensing elements, the aperture length



(a) Location ground truth: (50,50) (b) Location ground truth: (50,50) (c) Location ground truth: (50,50)

Fig. 4: **Location parameter estimation of a single sound source at -25 dB.** *Grating lobes of significant power levels are observed when measurements are taken either across the x-axis and y-axis, but fusing the measurements along both directions suppresses the unwanted lobes and noisy peaks to a reasonable extent.*



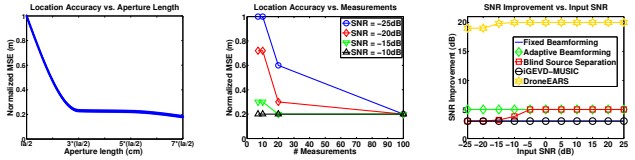
(a) Location ground truth: (50,50), (90,90) (b) Location ground truth: (20,20), (50,50), (90,90) (c) Location ground truth: (40,40), (50,50), (70,70), (90,90)

Fig. 5: **Location parameter estimation of multiple sound sources at -25 dB.** *Such a mechanism works only if there is good spatial separation between the different sound sources.*

A is the same as the inter-element distance B. We study the impact of B, ranging from $\lambda/2$ to $7\lambda/2$, on the localization performance of the sensor array with DroneEARS computing framework in Fig. 6(a). The result shows that the localization accuracy, measured in terms of the mean square error (MSE), improves by a significant margin of 70 cm as B takes on values from $\lambda/2$ to λ; but thereafter shows a gain of approximately 10 cm, as B goes all the way upto $7\lambda/2$. Since our aim is to robustly localize sound source with a maximum frequency of 5 kHz, we choose B to be $7\lambda/2$ (or 23 cm).

Localization quality. Mobility, and taking measurements at multiple locations is an integral component of the DroneEARS workflow. We study the impact of measurement count on the localization performance of DroneEARS at different SNR levels in Fig. 6(b). From the figure, we note that the received signals of very low SNR need more number of mobile measurements to achieve an average measure of localization accuracy compared to ones with higher SNR levels. This effect is very profound at the received SNR level of -25 dB, where at least 20 such spatial measurements are required to reduce the MSE by 50%. However, the number of measurements needed for signals that are received with higher SNR (-20 dB and above) levels to obtain a good location estimate is quite small.

The computation rationale of DroneEARS is to boost up the aggregate SNR of the received signal in the presence of strong, wideband and non-stationary ego-noise. Fig. 6(c) quantifies this measure and shows the SNR improvement (i.e., the different between the input and the output SNRs) obtained by DroneEARS. It is compared against fixed beamforming (using delay-and-sum [9]), adaptive beamforming (using MaxNSR [10]), blind source separation (using independent component analysis [11]), and generalized Eigen value decomposition for multiple signal classification (iGEVD-MUSIC [12]). The result shows that DroneEARS outperforms all of these algorithms by a significant margin of $15-18$ dB. A major limitation of most competing techniques is that they assume to have a good knowledge of the noise correlation matrix. While it can be reliably estimated in static operating conditions, the additional mobility factor makes it very challenging to dynamically derive it.

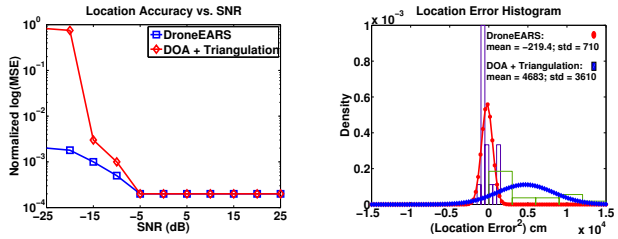


(a) λ is shorted to la (b) (c)

Fig. 6: **Performance analysis of DroneEARS.** (a) *The location accuracy improves with a larger aperture length, but this improvement tapers down significantly beyond $3\lambda/2$ and attains a constant level beyond $7\lambda/2$.* (b) *A minimum of 20 mobile measurements are desirable to obtain an average localization accuracy with low SNR signals.* (c) *DroneEARS outperforms all compared techniques by a significant margin of $15-18$ dB. It is important to note that DroneEARS is a mobile measurement scheme, while others are based on a static measurement setup.*

The configuration of the array (in terms of the number of sensing elements M, the inter-element separation distance B, and the aperture length A) has a major impact on the performance of the localization algorithms compared in Fig. 6(c). In fact, the performance of all of these algorithms would improve significantly if the count of the sensing elements are increased for the same aperture length. An alternate localization technique is to estimate the DOA of the signal source at multiple measurement points, and translate it to an equivalent location co-ordinate by triangulation. In this regard, we use the cross-correlation based DOA estimation technique proposed by Kottege [13] as it is well suited for two element ULA configurations[2]. Fig. 7(a) compares the location accuracy of DroneEARS with the (DOA [13] + triangulation) technique. The result shows that DroneEARS outperforms this competing methodology by more than *two* orders of magnitude for very low SNR conditions (< -20 dB), but this advantage starts to taper down aggressively for SNRs greater than -15 dB. We adopt the learning from these performance studies, and conduct an experiment to localize a sound source in an indoor setting with a scan region limited to a 6 m$\times$3 m area in the centre of a room. Fig. 7(b) shows the result of this exercise, where we

[2]The cross-correlation based DOA estimation technique proposed by Kottege [13] is guided by the principle that if the signal source is positioned sufficiently far away compared to B (which is equal to A for a two element ULA), then DOA can be expressed using the polar coordinate equation of the asymptotes of the hyperbola as: $\theta = \pm \arctan\left(\frac{\sqrt{B^2-\Delta B_1^2}}{\Delta B_1}\right)$; where: $\tan\theta$ is the gradient of the asymptotes, and ΔB_1 is the path difference of the traversing signal to the two array elements.



(a) (b)

Fig. 7: **Localization quality of DroneEARS.** (a): *Both the schemes are based on mobile measurements with a two element ULA.* (d) *DroneEARS recorded a mean error of 17cm and standard deviation (std.) of 26cm, which was 10 times better than the other method. Note: histogram shows the (location error)*^{2}.

notice that DroneEARS estimates the location of the sound source with a mean error of 17 cm, and standard deviation of 26 cm. Thus, its overall location error is less than 40 cm, which is significantly (i.e., 10 times) better than the other competing technique.

IV. Discussion and Conclusion

Discussion. Beamforming with microphone arrays has a rich history, and using it for location finding is an active topic of investigation. In the last four decades, many beamformers have been designed to efficiently sense the target location under both normal and high stress conditions [14], [15]. Existing techniques such as classical adaptive beamformers [9], time delay estimators, linear predictive [16] and minimum-norm methods [17], Root-MUSIC [18], ESPIRIT [19], cumulant-based techniques [20], etc., do offer noise resilience and are known to work robustly, but for static operating conditions with sensor arrays comprising of large number of receiving elements. The performance of many of these techniques severely deteriorate with extremely sparse array geometries (such as with just two sensing elements), when compensating for noise that is jointly strong, wideband and non-stationary. New variants of MUSIC such as GEVD-MUSIC [21] and iGEVD-MUSIC [12] have been proposed in the recent past for MAV-based auditory scene analysis. However, the results reported have been shown to work till -5 dB (with a 8-element sensor array); which is much lower than typical ego-noise strength levels.

Conclusion. We presented a binural sensing technique for MAVs - DroneEARS - to localize sound sources in the physical search space in a robust manner. By exploiting the inherent mobility of the flying vehicles, we showed that it is possible to overcome the limitations of sparse sensor arrays (with as low as two sensor elements) and strong ego-noise that is both wideband and non-stationary, by using a combination of intra-band and inter-measurement beam fusion. We designed the theoretical framework of DroneEARS; and validated its working through empirical analysis and evaluations, where we showed *one* order of magnitude improvement in acoustic source localization accuracy than other competing techniques.

DroneEARS would further benefit by mechanisms to enhance the received SNR; which was, by far, the most daunting task in realising this solution. The initial success of DroneEARS has motivated us to explore efficient hardware designs for performing acoustic-related computations onboard the MAV, and executing it in a principled manner remains an important future direction.

References

[1] D. Floreano and R. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, no. 7, pp. 460–466, May 2015.

[2] R. Carloni, V. Lippiello, M. D. Auria, M. Fumagalli, A. Y. Mersha, S. Stramigioli, and B. Siciliano, "Robot vision: Obstacle-avoidance techniques for unmanned aerial vehicles," *IEEE Robotics Automation Magazine*, vol. 20, no. 4, pp. 22–31, Dec 2013.

[3] A. J. Davison and D. W. Murray, "Simultaneous localization and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865–880, Jul 2002.

[4] L. C. Godara, "Application of antenna arrays to mobile communications. ii. beam-forming and direction-of-arrival considerations," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[5] Parrot, "Parrot AR Drone 2.0," https://www.parrot.com/fr/drones/parrot-ardrone-20-elite-edition, 2012.

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

[7] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[8] Z. Tang, G. Blacquière, and G. Leus, "Aliasing-free wideband beamforming using sparse signal representation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3464–3469, 2011.

[9] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[10] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using maxnsr blocking matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1493–1508, Sept 2015.

[11] J. F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct 1998.

[12] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 3288–3293.

[13] N. Kottege, "Underwater acoustic localisation in the context of autonomous submersibles," Ph.D. dissertation, The Australian National University, 2009.

[14] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin Heidelberg: Springer-Verlag, 2008.

[15] S. Gannot and I. Cohen, *Adaptive Beamforming and Postfiltering*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 945–978.

[16] J. Burg, *Maximum entropy spectral analysis*, ser. Stanford Exploration project. Stanford University, 1975.

[17] S. S. Reddi, "Multiple source location-a digital approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-15, no. 1, pp. 95–105, Jan 1979.

[18] B. D. Rao and K. V. S. Hari, "Performance analysis of root-music," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1939–1949, Dec 1989.

[19] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[20] A. Swami and J. M. Mendel, "Cumulant-based approach to harmonic retrieval and related problems," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1099–1109, May 1991.

[21] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 3943–3948.