Multi-Sound-Source Localization for Small Autonomous Unmanned Vehicles with a Self-Rotating Bi-Microphone Array

Deepak Gala, Nathan Lindsay, and Liang Sun

Abstract—While vision-based localization techniques have been widely studied for small autonomous unmanned vehicles (SAUVs), sound-source localization capability has not been fully enabled for SAUVs. This paper presents two novel approaches for SAUVs to perform multi-sound-sources localization (MSSL) using only the interaural time difference (ITD) signal generated by a self-rotating bi-microphone array. The proposed two approaches are based on the DBSCAN and RANSAC algorithms, respectively, whose performances are tested and compared in both simulations and experiments. The results show that both approaches are capable of correctly identifying the number of sound sources along with their three-dimensional orientations in a reverberant environment.

I. INTRODUCTION

Small autonomous unmanned vehicles (e.g., quadcopters and ground robots) have revolutionized civilian and military missions by creating a platform for observation and permitting access to locations that are too dangerous, too difficult or too costly to send humans. These small vehicles have shown themselves to be remarkably capable in a lot of applications, such as surveying and mapping, precision agriculture, search and rescue, traffic surveillance, and infrastructure monitoring, to name just a few.

The sensing capability of unmanned vehicles has been enabled by various sensors, such as RGB cameras, infrared cameras, LiDARs, RADARs, and ultrasound sensors. However, these mainstream sensors are subject to either lighting conditions or line-of-sight requirements. On the end of the spectrum, sound sensors have the superiority to conquer line-of-sight constraints and provide a more efficient approach for unmanned vehicles to acquire situational awareness thanks to their omnidirectional nature.

Among the sensing tasks for unmanned vehicles, localization is of utmost significance [1]. While vision-based localization techniques have been developed based on cameras, sound source localization (SSL) has been achieved using microphone arrays with different numbers (e.g., 2, 4, 8, 16) of microphones. Although it has been reported that the accuracy of the localization is enhanced as the number of microphones increases [2], [3], this comes with the price of algorithm complexity and hardware cost, especially due to the expense of the Analog-to-Digital converters (ADC), which is proportional to the number of speaker channels. Moreover, arrays with a particular structure (e.g., linear, cubical, circular, etc.) will be difficult to control, mount and maneuver, which makes them unsuitable to be used on small vehicles.

Humans and many other animals can locate sound sources with decent accuracy and responsiveness by using their two ears associated with head rotations to avoid ambiguity (i.e., cone of confusion) [4]. Recently, SSL techniques based on a self-rotating bi-microphone array have been reported in the literature [5]–[9]. Single-SSL (SSSL) techniques have been well studied using different numbers of microphones, while for multi-sound-source-localization (MSSL) many reported techniques require large microphone arrays with specific structures, limiting them to be mounted on small robots. Pioneer work for MSSL assumed the number of sources to be known beforehand [10], [11]. Some of these approaches [12]–[14] are based on sparse component analysis (SCA) that requires the sources to be W-disjoint orthogonal [15] (i.e., in some time-frequency components, at most one source is active), thereby making them unsuitable for reverberant environments. Pavlidi et al. [16] and Loesch et al. [17] presented an SCA-based method to count and localize multiple sound sources but requires one sound source to be dominant over others in a time-frequency zone. Clustering methods have also been used to conduct MSSL [12]–[14], Catalbas et al. [18] presented an approach for MSSL by deploying four microphones at the corners of the room and the sound sources are required to be present within the boundary. The technique was limited to localize sound orientations in the two-dimensional plane using K-medoids clustering. The number of sound sources were calculated using the exhaustive elbow method, which is instinctive and computationally expensive. Traa et al. [19] presented an approach that converts the time-delay between the microphones in the frequency domain so as to model the phase differences in each frequency bin of short-time Fourier transform. Due to the linear relationship between phase difference and frequency, the data were then clustered using random sample consensus (RANSAC). In our
previous work [8], [9], we developed a SSSL technique based on an extended Kalman filter and a MSSL technique based on a cross-correlation approach, which was very computationally expensive.

The contributions of this paper include two novel MSSL approaches for identification of the number of sound sources as well as localizing them in a three-dimensional (3D) environment. The rotation of the bi-microphone array generates an Interaural Time Difference (ITD) signal with data points forming multiple discontinuous sinusoidal waveforms. In the first approach, a novel mapping mechanism is developed to convert the acquired ITD signal to an orientation domain. An unsupervised classification is then conducted using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN [20] is one of the most popular nonlinear clustering techniques. It can discover any arbitrary shaped clusters of densely grouped points in a data set and outperforms other clustering methods in the literature [21], [22].

The second presented novel approach for MSSL completes a sinusoidal ITD regression using a RANSAC-based method. Each of the sine waves in the ITD signal corresponds to a single sound source. The data points associated with each sine wave is separated by performing a repeated sinusoidal regression using RANSAC [23]. After a model is fitted in an iteration, the associated data points will be removed from the ITD signal before the next iteration starts. The azimuth and elevation angles of the sound source are then determined for each fitted model. A threshold is then selected to determine the number of sound sources with the qualifying number of data points.

Both simulations and experiments were conducted to test the proposed two approaches. The results show that both approaches are capable of correctly generating the number of sound sources and their 3D orientations in terms of azimuth and elevation angles. However, the RANSAC-based approach outperforms the DBSCAN-based approach on the identification of the number of sound sources, while the DBSCAN-based approach outperforms the RANSAC-based approach in the localization accuracy.

The rest of the paper is organized as follows. In Section [II-B] the mathematical calculation for the ITD signal generated by the self-rotating microphone array is presented. In Section [III] the mapping mechanism for regression is presented. Section [IV] presents the localization algorithm using DBSCAN clustering and Section [V] presents the RANSAC-based localization algorithm. Simulation results are presented and discussed in Section [VI] and Section [VII] concludes the paper.

II. PRELIMINARIES

A. Interaural Time Difference (ITD)

The ITD is the time difference between a sound signal arriving at two microphones and can be calculated using the cross-correlation technique [24], [25].

![Cross-correlation technique](image)

Consider a single stationary sound source and two spatially separated microphones placed in an environment. Let $y_1(t)$ and $y_2(t)$ be the sound signals captured by the microphones in the presence of noise, which are given by

$$
y_1(t) = s(t) + n_1(t),$$

$$
y_2(t) = \delta \cdot s(t + t_d) + n_2(t),$$

where $s(t)$ is the sound signal, $n_1(t)$ and $n_2(t)$ are real and jointly stationary random processes, $t_d$ denotes the time difference of $s(t)$ arriving at the two microphones, and $\delta$ is the signal attenuation factor due to different traveling distances. It is commonly assumed that $\delta$ changes slowly and $s(t)$ is uncorrelated with noises $n_1(t)$ and $n_2(t)$ [24]. Figure 1 shows the process of ITD estimation between signals $y_1(t)$ and $y_2(t)$, where $H_1(f)$ and $H_2(f)$ could be the scaling functions or pre-filters [24], which eliminate or reduce the effect of background noise and reverberations using various techniques [26]–[29].

The cross-correlation function of $y_1$ and $y_2$ is given by

$$R_{y_1,y_2}(\tau) = E\left[y_1(t) \cdot y_2(t - \tau)\right],$$

where $E[\cdot]$ represents the expectation operator. The time difference of $y_1$ and $y_2$, i.e., the ITD, is given by $T \triangleq \arg\max_\tau R_{y_1,y_2}$. The distance difference of the sound signal traveling to the two microphones is given by $d \triangleq T \cdot c_0$, where $c_0$ is the sound speed and is usually selected to be 345 m/s on the Earth surface.

Remark 1. As a matter of simplicity, the signal $d$ is referred as ITD in the context. ITD is the only cue used in this paper for the source counting and localization, generated without using any scaling functions nor pre-filters mentioned above.

B. Mathematical Model for ITD signal

Before discussing the multi-source ITD signal collected by the self-rotating bi-microphone array, the single source ITD signal is first modeled. In this paper, the location of a single sound source is defined in a spherical coordinate frame, whose origin is assumed to coincide with the center of a ground robot.

As shown in Figures 2 and 3, the left and right microphones, $L$ and $R$, collects the acoustic signal generated by the sound source $S$. Let $O$ be the center of the robot as well as the bi-microphone array. The sound source location is represented by $(D, \theta, \varphi)$, where $D$ is the distance between the source and the center of the robot, i.e., the length of segment $d$. Let $T$ be the time delay from the source to the microphones $L$ and $R$. The ITD is given by $T = \frac{2d}{c_0}$. The effect of background noise and reverberations is reduced using various techniques [26]–[29].
In the shaded triangle, $SOF$, shown in Figures 3 and 4, define $\alpha = \angle SOF$ and we have $\cos \alpha = \cos \theta \sin \psi$. Based on the far-field assumption [30], we have
\[ d \triangleq \hat{T} \cdot c_0 = 2b \cos \alpha = 2b \cos \theta \sin \psi. \] (4)

To avoid cone of confusion [41] in SSL, the binaural microphone array needs to be rotated with a nonzero angular velocity [11]. Without loss of generality, in this paper we assume a clockwise rotation of the microphone array on the horizontal plane while the robot itself does not rotate throughout the entire estimation process, which implies that $\varphi$ is constant.

The initial heading of the microphone array is configured to coincide with the heading of the robot, i.e., $\beta(t = 0) = 0$, which implies that $\varphi = \psi(0)$. As the microphone array rotates clockwise with a constant angular velocity, $\omega$, we have $\beta(t) = \omega t$ and due to Equation (3) we have
\[ \psi(t) = \varphi - \beta(t) = \varphi - \omega t. \] (5)

The resulting time-varying $d(t)$ due to Equation (4) is then given by
\[ d(t) = 2b \cos \theta \sin (-\omega t + \varphi). \] (6)

Because the microphone array rotates on the horizontal plane, $\theta$ does not change during the rotation for a stationary sound source. The resulting $d(t)$ is a sinusoidal signal with the amplitude $A \triangleq 2b \cos \theta$, which implies that
\[ \theta = \cos^{-1} \frac{A}{2b}. \] (7)

It can be seen from Equation (6), the phase angle of $d(t)$ is the azimuth angle of the sound source. Therefore, the localization of a stationary sound source equates the identification of the characteristics (i.e., the amplitude and phase angle) of the sinusoidal signal, $d(t)$.

The collection of the ITD signal for multiple sound sources (as shown in Figure 11) illustrates a group of multiple discontinuous sinusoidal waveforms, each corresponding to a single sound source, satisfying the amplitude-elevation and phase-azimuth relationship as mentioned above.

III. MODEL FOR MAPPING AND SINUSOIDAL REGRESSION

The signal $d(t)$ in Equation (6) is sinusoidal with its amplitude $A = 2b \cos \theta$ and phase angle $\varphi$ that corresponds to the azimuth angle of the sound source. Since the frequency, $\omega$, of $d(t)$ is the known rotational speed of the microphone array, the localization task (i.e., identifying $\theta$ and $\varphi$) is to estimate the amplitude and phase angle of $d(t)$, i.e., $A$ and $\beta$. Consider a general form of $d(t)$ expressed as
\[ d(t) = A_1 s_{\omega t} + A_2 c_{\omega t}, \] (8)
where $s_{\omega t} \triangleq \sin(\omega t)$ and $c_{\omega t} \triangleq \cos(\omega t)$, and we have
\[ A = \sqrt{A_1^2 + A_2^2}, \] (9)
and
\[ \varphi = \tan^{-1} \left( \frac{A_2}{A_1} \right). \] (10)

Consider the two data points, \( y_1 = d(t_1) \) and \( y_2 = d(t_2) \), collected at two distinct time instants \( t_1 \) and \( t_2 \), respectively, and we have
\[
\begin{bmatrix}
  y_1 \\
  y_2 
\end{bmatrix} = \begin{bmatrix}
  s_{\omega t_1} & c_{\omega t_1} \\
  s_{\omega t_2} & c_{\omega t_2}
\end{bmatrix} \begin{bmatrix}
  A_1 \\
  A_2 
\end{bmatrix},
\] (11)
which gives,
\[
\begin{bmatrix}
  A_1 \\
  A_2 
\end{bmatrix} = \frac{1}{s_{\omega(t_2-t_1)}} \begin{bmatrix}
  y_2c_{\omega t_1} - y_1c_{\omega t_2} \\
  y_1s_{\omega t_2} - y_2s_{\omega t_1}
\end{bmatrix}. \] (12)

According to (9) and (10), we have
\[
A = \sqrt{\left( \frac{y_2c_{\omega t_1} - y_1c_{\omega t_2}}{s_{\omega(t_2-t_1)}} \right)^2 + \left( \frac{y_1s_{\omega t_2} - y_2s_{\omega t_1}}{s_{\omega(t_2-t_1)}} \right)^2},
\] (13)
and
\[
\varphi = n\pi + \tan^{-1} \left( \frac{y_1s_{\omega t_2} - y_2s_{\omega t_1}}{y_2c_{\omega t_1} - y_1c_{\omega t_2}} \right), \] (14)
where \( n \) is an integer.

IV. DBSCAN-BASED MSSL

In the DBSCAN algorithm [21], a random point from the data set is considered as a core cluster point when more than \( m \) points (including itself) within a distance of \( \epsilon \) (epsilon ball) exists in its neighborhood. This cluster is then extended by checking all of the other points satisfying the \( \epsilon \) and \( m \) criteria thereby letting the cluster grow. A new arbitrary point is then chosen and the process is repeated. The point which is not a part of any cluster and having fewer than \( m \) points in its epsilon ball is considered as a "noise point". The DBSCAN technique is more suitable for applications with noise and performs better than the Kmeans method, which requires a prior knowledge of the number and the approximate centroid locations of clusters and can also fail in the presence of noisy data points.

The DBSCAN-Based MSSL technique consists of two stages. In the first stage, the data points of the ITD signal are mapped to the orientation domain. The data set consisting of all the data points in multi-source ITD signal contains not only inliers but also outliers, which produce undesired mapped locations. When the number of inliers is significantly greater than the outliers after a number of iterations, highly dense clusters will be formed. In the second stage, these clusters are detected using the DBSCAN technique by carefully selecting parameters \( m \) and \( \epsilon \). The number of clusters corresponds to the number of sound sources and the centroids of these clusters represent the locations of the sound sources.

The complete DBSCAN-based MSSL algorithm is described in [1]. Two points in the data set are selected randomly and mapped into the orientation domain by calculating the angles \( \theta \) and \( \varphi \) using Equations (6), (13), (7) and (14). A set \( S := \{ (\theta_1, \varphi_1), (\theta_2, \varphi_2), ..., (\theta_N, \varphi_N) \} \) of these mapped points is then created. The process for detection of clusters is then started. A point \( (\theta_i, \varphi_i) \) in \( S \) is randomly chosen and is decided to be a core cluster point or a noise point by checking the density-reachability criteria under the \( m\epsilon \)-condition [21].

The time complexity of Algorithm 1 is \( \mathcal{O}(ND + ND^2) \), where \( ND \) is the number of iterations for mapping and clustering. The number of iterations, \( ND \), needs to be selected large enough for the algorithm to work efficiently.

V. RANSAC-BASED MSSL

The RANSAC algorithm [23] is able to identify inliers (e.g., parameters of a mathematical model) in a data set that may contain a significantly large number of outliers. The input to the RANSAC algorithm includes a set of data, a parameterized model, and a confidence parameter \( \sigma_{conf} \). In each iteration, a subset of the original data is randomly selected and used to fit the predefined parameterized model. All other data points in the original data set are then tested against the fitted model. A point is determined to be an inlier of the fitted model, if it satisfies the \( \sigma_{conf} \) condition. The process is repeated by selecting another random subset of the data. After a fixed number of iterations, the parameters are then selected for the best fitting (with maximum inliers) estimated model.

The RANSAC-based MSSL method is described in Algorithm 2. It can be seen from Equation (9) that the signal \( d(t) \) generated by the self-rotating bi-microphone array is sinusoidal. Two points from the ITD signal are selected randomly and a sine wave with the given frequency (i.e., the angular speed of the rotation, \( \omega \)) is generated. The count represents the number of points whose distance to the fitted sine wave is less than \( \sigma_{conf} \), which is the threshold.
Algorithm 2 RANSAC-Based MSSL

1: Capture $d(t)$ for one full rotation of the bi-microphone array
2: Select $N_R, \sigma_{conf}$ and initialize $e = 0$
3: WHILE there are samples in $d$
4: \hspace{1em} FOR $j = 1$ to $N_R$ DO
5: \hspace{2em} Randomly choose $y_1$ and $y_2$ from $d$
6: \hspace{2em} Calculate $\hat{A}$ and $\hat{\phi}$ using Equations (13) and (14)
7: \hspace{2em} Calculate $\hat{d} = \hat{A} \sin(\omega t + \hat{\phi})$
8: \hspace{2em} Calculate $\text{count} = \text{number of points in } d \text{ fitting } \hat{d}$ with at least $\sigma_{conf}$
9: \hspace{2em} IF $e < \text{count}$
10: \hspace{3em} $A_K = \hat{A}, \phi_K = \hat{\phi}$ and $e = \text{count}$
11: \hspace{2em} END IF
12: \hspace{1em} END FOR
13: Calculate $\theta_K$ using Equation (7)
14: $e = 0$
15: Remove samples on $\hat{d}$ within $\sigma_{conf}$ from $d$
16: END WHILE

for a point to be considered inlier. Then the points in $d$ that belong to $\hat{d}$ according to the $\sigma_{conf}$ condition will be removed from $d$. This procedure is repeated for $N_R$ iterations and the parameters $A_K$ and $B_K$ are updated every time the number of inliers is greater than that in the previous iterations. This process is repeated until either all the points in $d$ are examined or $N_R$ iterations are completed. The time complexity of Algorithm 2 is $O(\frac{n}{\epsilon} \cdot N_R)$, where $n$ is the number of samples in the ITD. After the first few of $N_R$ iterations, most of the data points are removed. This results in $n$ to be very small number as compared to $N_R$. The iterations $N_R$ should be chosen large enough to ensure the probability that at least one of the sets of randomly selected points does not include an outlier.

The number of sound source is determined by carefully selecting a threshold, as shown in Figure 5. The confidence about the presence of a sound source is dependent on the $\text{count}$ value. The source with the maximum $\text{count}$ is considered to be qualified with 100 % confidence and the confidence values for other sources are calculated relatively.

VI. SIMULATION AND EXPERIMENTAL RESULTS

Audio Array Toolbox [31] is used to establish an emulated rectangular room using the image method described in [32]. The robot was placed in the origin of the room. The sound sources and the microphones are assumed omni-directional and the attenuation of the sound are calculated per the specifications in Table 1. A number of recorded speech signals available at [33] were used as sound sources to test the technique. Different number of sound sources were placed at various azimuth and elevation angles at a fixed distance of 5 m and the ITD signal was recorded by the rotating bi-microphone array with mics separated by a distance of 0.18 m (which is approximate distance between the ears of a human). The sound sources were separated by at least 20° in azimuth and at least 10° in elevation. The ITD value was calculated and recorded every 1° of rotation. Noise with a variance ($\sigma_{\text{noise}}$) of 0.001 was added to this ITD signal for the simulations in order to account for sensor noise. These simulations were run on a high performance cluster named Joker [34].

![Figure 5. Confidence on presence of sound source](image)

Table 1

| Parameter                 | Value             |
|---------------------------|-------------------|
| Dimension                 | 20 m x 20 m x 20 m|
| Reflection coefficient    | 0.5               |
| (walls, floor and ceiling)|                   |
| Sound speed               | 345 m/s           |
| Temperature               | 22°C              |
| Static pressure           | 29.92 mmHg        |
| Relative humidity         | 38 %              |

Figure 6. Experimental environment and the robotic platform based on a Kobuki turtlebot 2 with two MEMS microphones.
Numerous experiments were also conducted using a robotic platform in an indoor environment with $RT60 = 670$ ms, as shown in Figure 6. Figure 7 shows the impulse response of the room. The sound sources were kept at a distance of about 1 to 1.5 m from the center of the robot.

![Impulse response of the room](image)

Figure 7. Impulse response of the room reverberation showing secondary peaks representing the reflections from the floor and the walls.

![Robot and source locations](image)

Figure 8. Sound source locations in the simulation.

![Mapped data points](image)

Figure 9. Simulation results of the mapped data points in the orientation domain.

Figure 8 shows four sound sources in the simulation placed at $S_1(20^\circ, 50^\circ)$, $S_2(30^\circ, 150^\circ)$, $S_3(50^\circ, 200^\circ)$ and $S_4(60^\circ, 300^\circ)$ and a distance of 5 m from the robot placed at the origin. Figure 9 shows the estimation of the parameters $\theta$ and $\phi$ for every two points chosen from the multi-source domain.

### Table II

| Parameters       | For simulations | For experiments |
|------------------|-----------------|-----------------|
| $\sigma_{\text{conf}}$ | 0.015688 m      | 0.015688 m      |
| $N_D$            | 1000            | 1000            |
| $N_T$            | 6000            | 6000            |
| $\text{Threshold}$ | 10 %            | 7 %             |
| $\epsilon$       | 3$^\circ$       | 3$^\circ$       |
| $m$              | 40              | 40              |
| $\sigma_{\text{noise}}$ | 0.001 m         | -               |

### Table III

| No. of sources | MAE (Sim) $\varphi$ (deg) | MAE (Expt) $\varphi$ (deg) | MAE (Expt) $\theta$ (deg) | Avg (deg) |
|----------------|---------------------------|-----------------------------|----------------------------|-----------|
| 4              | 1.73                      | 3.20                        | 3.71                       | 3.58      |
| 3              | 0.38                      | 3.58                        | 3.17                       | 3.26      |
| 2              | 1.96                      | 2.18                        | 2.89                       | 2.92      |
| 1              | 0.94                      | 0.57                        | 2.27                       | 0.35      |

### Table IV

| No. of sources | MAE (Sim) $\varphi$ (deg) | MAE (Expt) $\varphi$ (deg) | MAE (Expt) $\theta$ (deg) | Avg (deg) |
|----------------|---------------------------|-----------------------------|----------------------------|-----------|
| 4              | 0.88                      | 1.82                        | 3.92                       | 1.20      |
| 3              | 2.51                      | 3.56                        | 2.33                       | 5.62      |
| 2              | 2.55                      | 3.01                        | 3.07                       | 3.88      |
| 1              | 1.61                      | 1.97                        | 2.15                       | 3.07      |

### Table V

| Estimated vs Actual number of sound source count for DBSCAN-based MSSL in the simulated environment. |
|-------------------------------------------------|
| $K$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|-----|----|----|----|----|----|----|----|----|----|
|     | 1  | 941| 56 | 0  | 0  | 0  | 0  | 0  | 0  |
|     | 2  | 11 | 902| 48 | 15 | 7  | 0  | 0  | 0  |
|     | 3  | 1  | 61 | 847| 47 | 38 | 5  | 0  | 0  |
|     | 4  | 12 | 79 | 126| 68 | 82 | 36 | 11 | 6  |
|     | 5  | 4  | 17 | 80 | 183| 393| 110| 6  | 3  |

Signal $d$ at each iteration. The parameters used for the RANSAC-based and DBSCAN-based algorithms are listed in Table III. The value $\sigma_{\text{conf}}$ was chosen to be 0.015688 m, which implies that all sound sources are assumed to be separated by at least 5$^\circ$ from each other. Tables III and IV show the simulation and experimental results of localization with the number of sound sources varying from one to four.

1000 Monte Carlo simulation runs were performed using the two proposed approaches, respectively, with specifications given in Table II. The simulation were run with $K = 1, 2, 3, 4, 5$ sources and the results of the source counting are listed in the Table V. The clustering result using DBSCAN, as shown in Figure 10.

Figure 11 shows the simulation result of a sample run.
where three sources were detected. The three sound sources were kept at $S_1(20, 10^\circ)$, $S_2(180^\circ, 20^\circ)$ and $S_3(3400^\circ, 30^\circ)$ and the estimated locations by the RANSAC-based algorithm are $S_1(24.31^\circ, 12.12^\circ)$, $S_2(181.85^\circ, 22.14^\circ)$ and $S_3(342.07^\circ, 33.08^\circ)$, respectively. Since the ITD signal is noisy any point very close ($\sigma_{\text{conf}} = 0.015688$) to any of $d_n$ was chosen to be on the ITD by the RANSAC algorithm. The value $\sigma_{\text{conf}, \text{desc}}$ can be chosen depending on possibility of sound sources to be close to each other and the noise level. The signal to noise ratio (SNR) of the measured signal $d$ was 18.94 dB. For a source to be considered as a qualified sound source, the threshold for the confidence that worked for us was 10% in simulation and 7% in experiments.

As shown in Figure 12, the average error of orientation localization with the DBSCAN-based algorithm is less as compared to the RANSAC-based algorithm, which, however, generates comparatively more accurate results for source counting, as shown in the Figure 13. In both simulations and experiments, the error of elevation angle estimation was found to be large for sources kept close to zero elevation, which coincide the conclusion in [9]. The performance of the localization and source counting using both the aforementioned techniques improves significantly by increasing the number of rotation of the bi-microphone array. The sound sources are assumed to be active during the rotation of the bi-microphone array with possible pauses such as in case of speech signals.

VII. Conclusion

Two novel techniques are presented for small autonomous unmanned vehicles (SAUVs) to perform multi-sound-source localization (MSSL) using a self-rotating bi-microphone array. The DBSCAN-based MSSL approach iteratively maps the randomly chosen points in the ITD signal to the orientation domain, leading to a data sets for clustering. These clusters are detected using the density based spatial clustering for application with noise (DBSCAN). The number of clusters gives the number of sound sources and the location of the centroid of these clusters determines the location of the sound sources. The second proposed technique uses random sample
consensus (RANSAC) to iteratively estimate parameters of a model using two randomly randomly chosen data points from the ITD signal data. It then uses a threshold to decide between the qualifying sound sources. The simulation and experimental results show the effectiveness of both approaches in identifying the number and the orientations of the sound sources.

REFERENCES

[1] J. Borenstein, H. Everett, and L. Feng, *Navigating mobile robots: systems and techniques*. A K Peters Ltd., 1996.

[2] D. V. Rahatkin, “Optimum sensor placement for microphone arrays,” Ph.D. dissertation, RUTGERS The State University of New Jersey - New Brunswick, 1998.

[3] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[4] H. Wallach, “On sound localization,” *The Journal of the Acoustical Society of America*, vol. 10, no. 4, pp. 270–274, 1939.

[5] S. Lee, Y. Park, and Y.-s. Park, “Three-dimensional sound source localization using inter-channel time difference trajectory,” *International Journal of Advanced Robotic Systems*, vol. 12, no. 12, p. 171, 2015.

[6] A. A. Handzel and P. Krishnaprasad, “Biomimetic sound-source localization,” *IEEE Sensors Journal*, vol. 2, no. 6, pp. 607–616, 2002.

[7] G. H. Eriksen, “Visualization tools and graphical methods for source localization and signal separation,” Master’s thesis, University of Oslo, Department of Informatics, 2006.

[8] X. Zhong, W. Yost, and L. Sun, “Dynamic binaural sound source localization with ITD cues: Human listeners,” *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2376–2376, 2015.

[9] D. Gala, N. Lindsay, and L. Sun, “Three-dimensional sound source localization for unmanned ground vehicles with a self-rotational two-microphone array,” in *Proceedings of the 5th International Conference of Control, Dynamic Systems, and Robotics (CDSR’18)*. Accepted (2018), June.

[10] L. Sun and Q. Cheng, “Indoor multiple sound source localization using a novel data selection scheme,” in *48th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2014, pp. 1–6.

[11] X. Zhong, L. Sun, and W. Yost, “Active binaural localization of multiple sound sources,” *Robotics and Autonomous Systems*, vol. 85, pp. 83–92, 2016.

[12] 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.

[13] M. Swartling, B. Sällberg, and N. Grbić, “Source localization for microphone arrays,” in *International Organization for Standardization (ISO), British, European and International Standards (BSEN)*, Noise emitted by machinery and equipment – Rules for the drafting and presentation of a noise test code, 12001: 1997 Acoustics.

[14] K. D. Donohue, “Audio array toolbox,” [Online] Available: http://vis.uky.edu/distributed-audio-lab/about/, 2017, Dec 22.

[15] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] K. D. Donohue, “Audio systems lab experimental data - single-track single-speaker speech,” [Online] Available: http://hpc.nmsu.edu/clusão/audio/Data/audioexpdata.htm , 2010, Feb 10.

[17] New Mexico State University, “High performance cluster, joker,” [Online] Available: http://https://hpc.nmsu.edu/, 2017, Feb 10.