A Two Stage Approach for Passive Sound Source Localization Based on the SRP-PHAT Algorithm

M.A.AWAD-ALLA
Ain Shams University
Department of Mechatronics
matef70@yahoo.com

AHMED HAMDY
Helwan University
Department of Electrical Engineering
ahmed_hamdi@h-eng.helwan.edu.eg

FARID A.TOLBAH
Ain Shams University
Department of Mechatronics
tolbah2@yahoo.com

MOATASIM A.SHAHIN
Badr University
Department of Mechatronics
m-shahin@outlook.com

M.A.ABDELAZIZ
Ain Shams University
Department of Automotive Engineering
Mohamed_abdelaziz@eng.asu.edu.eg

Abstract: This paper presents a viable solution for the Sound Source Localization (SSL) problem. The aim of this paper is to develop a computationally viable approach to find the coordinate location of a sound source with acceptable accuracy when using a compact microphone array that can be mounted on top of a small moving robot. The approach suggested in this paper uses the SRP-PHAT algorithms in its core and it comprises two stages: the first stage contracts the search space by estimating the Direction of Arrival (DoA) vector using the near field model which in turn is used to form a smaller search region, the second stage is to use the SRP-PHAT algorithm to search this contracted region for the source location. The AV16.3 corpus was used to test the approach by running extensive experiments. The results are reported and it proves the effectiveness of this approach

Key–Words: Sound Source Localization, Passive Acoustic Localization, SRP-PHAT, Circular Microphone Array, Region Contraction.

1 Introduction

Sound Source Localization (SSL) is an important part of a robot's auditory system. It is used by autonomous robots to locate a target based on acoustic signals gathered by microphones, this can help when other robot's systems, such as the vision system, are impaired, this can be due to bad lighting conditions or other reasons. The SSL system on the robot must locate the acoustic target with accuracy even if the acoustic signals are noisy, in addition, it must be able to work in a diverse environment. The robot auditory system is expected to be small enough to fit on the robot and to be economical, such constraints make it difficult to achieve the SSL requirements regarding its accuracy and robustness. The motivation of this work is to develop a robust, accurate, computationally non-intensive SSL system that can be used on a mobile robot to find the coordinates of a speech source in an indoor environment using a small microphone array.

SSL approaches can be categorized into three main categories [1]: approaches based on Time Difference of Arrival (TDOA), approaches based on high resolution spectral calculations and approaches based on maximizing a beamformer.

TDOA approaches are usually two step approaches that involve the estimation of TDOAs between the signals of pairs of microphones as the first step, then mapping these TDOAs to an acoustic source location using geometrical relations. TDOA based locators are widely used in localization applications because of their simplicity in implementation and their low computational burden, but such locators rely mainly on the accuracy of the TDOA estimation; a small error in TDOA estimates can lead to significant error in the location estimation.

Several efforts have been done and reported in the literature in order to overcome the limitations of the TDOA based locators such as [2, 3, 4, 5, 6, 7, 8, 9, 10] which focused on increasing the robustness of the locator to ambient noise and reverberation. However, it is very difficult to obtain, using computationally viable algorithms, accurate acoustic location especially when small size microphone arrays are used.

The second category is the MUSIC-based locators. The MUSIC algorithm is a high resolution spectral analysis algorithm that has been extensively used, and its derivatives [11, 12, 13, 14] , in speaker localization tasks. Originally it was intended for narrowband signals, but several modifications has extended its use to wideband signals such as those of audio signals. This class of algorithms, although having high resolution, suffer from the very high computational load. Even though there exists some efforts to reduce this computational burden, still all MUSIC-based algorithms need eigenvalue or singular value decomposition which are computationally extensive operations [15]. This computational limitations limit the use of such algorithms in commercial compact microphone arrays.

The beamforming based methods search among possible candidate locations for the location that maximizes a certain function. The most successful and used algorithm in this category is the SRP-PHAT algorithm which finds the location that maximizes the SRP-PHAT function [16]. This algorithm has proven to be robust to ambient noise and reverberation to a certain extent. The main limitation of algorithms in this category is the computational burden resulting from the search process. [17, 18, 19, 20, 21, 22, 23] are some of the efforts in the literature to improve the computational burden of the SRP-PHAT algorithm, however, they involve iterative optimization or statistical algorithms which can be complicated to implement. There are other limitations to the SRP-PHAT algorithm other than its computational burden that can affect the localization estimate and its resolution. high levels of noise and reverberation can lead to an unsatisfactory location estimates, moreover, discrete calculations involved in calculating the SRP-PHAT function can lead to a wrong location estimate; a wrong location can have slightly higher or similar SRP-PHAT value compared to that of the true location. The source of such errors is due to discrete calculations resulting from: low sampling frequency, using the FFT algorithm in GCC-PHAT estimation and interpolation, [24, 25, 26, 27] addressed these limitations.

In this paper a two-stage mixed near field/far field approach is adopted. First, the far field model is adopted to estimate the Direction of Arrival (DOA) of the acoustic source in the closed form, then an uncertainty bound is applied to this DOA to form, along with a predefined search radius, a search region. The SRP-PHAT algorithm is applied on this contracted search region to extract the coordinates of the acoustic location. This approach has several merits: the search region is contracted to a smaller one with a very high degree of confidence that the acoustic source lies in it, this speeds up the SRP-PHAT search. Moreover, it would be highly unlikely that in the contracted search region would exists several maxima, therefore, the peak power found in that region would be that of the true source location.

This paper is organized as follows: after this introduction, section 2 derives the closed form solution for estimating the Direction of Arrival (DoA) vector represented by the azimuth and elevation angles this will be used as the first stage in the proposed localization scheme to contract the search area. Section 3 Describes briefly the SRP-PHAT algorithm which is adopted in this paper as the search algorithm that will search for the acoustic source in the contracted area obtained in the previous section. Section 4 describes the proposed localization approaches and explains the theory and rational behind it. The results are then showed and analyzed in section 5. Finally section 6 concludes this paper.

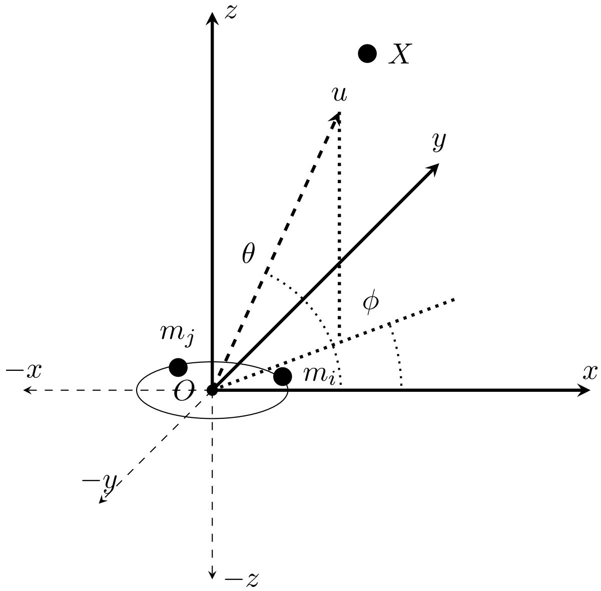

Figure 1: Localization system's geometry

2 Direction of Arrival (DoA) in the closed form

Consider a microphone array consisting of M microphone elements each at location $m(m_x, m_y, m_z)$ arranged in the xyz plane in any arbitrary geometry as shown in figure 1, it is required to estimate the direction vector $\overrightarrow{u}$ pointing at the acoustic source X. The Direction of Arrival (DoA) can be estimated from the Time Difference of Arrival (TDOA) between pairs of microphones, where there exist $M(M-1)/2$ pairs of microphones. let $\overrightarrow{u}$ be a unit direction vector pointing at the direction of the sound source:

$$\overrightarrow{u} = \begin{bmatrix} \cos(\theta)\cos(\phi) \\ \cos(\theta)\sin(\phi) \\ \sin(\theta) \end{bmatrix} = \begin{bmatrix} u_x \\ u_y \\ u_z \end{bmatrix} \tag{1}$$

Where ϕ is the *azimuth* and θ is the *elevation.*

The relationship between this direction vector and the TDOAs can be defined:

$$\tau_{ij}(\phi,\theta) = \frac{\overrightarrow{u}.(\overrightarrow{m_i} - \overrightarrow{m_j})}{c} \tag{2}$$

$$= \begin{bmatrix} \cos(\theta)\cos(\phi) \\ \cos(\theta)\sin(\phi) \\ \sin(\theta) \end{bmatrix} . \left(\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} - \begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} \right) . \frac{1}{c} \tag{3}$$

let $S = \left(\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} - \begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} \right)^T$ and $c\tau_{ij} = d$.

Rearranging equation 3:

$$\boldsymbol{d} = \boldsymbol{S}\overrightarrow{u} \tag{4}$$

Equation 4 is over-determined equation because there are $M(M-1)/2$ and only three variables, therefore, equation 4 can be solved in the closed-form using a simple least squares solution.

$$\overrightarrow{u} = -\boldsymbol{S}^{+}\boldsymbol{d} = -(\boldsymbol{S^T S})^{-1}\boldsymbol{S}^{T}\boldsymbol{d} \tag{5}$$

Where $\boldsymbol{S}^{+}$ is the pseudoinverse of $\boldsymbol{S}$.
After calculating the direction vector $\overrightarrow{u}$ the *azimuth* and *elevation* angles can be easily calculated:

$$\phi = atan2\left(\frac{u_y}{u_x}\right) \tag{6}$$

$$\theta = sin^{-1}\left(u_z\right) \tag{7}$$

Using the $atan2$ function in equation 6 allows for an efficient way to find theta in the $[-\pi, \pi]$ range provided that the microphone array has its elements distributed in the xy plane. in [28] they assumed that the elevation angle is equal to zero (for microphone array with all its elements in the xy plane) and estimated $cos(\theta) = 1$ and hence estimated the azimuth directly from the relations: $\phi = cos^{-1}(u_x)$ or $\phi = sin^{-1}(u_y)$. It was found that when assuming that the elevation is zero (which is not necessarily the case) the previous two relations will not yield the same azimuth angle this is because the elevation angle greatly influence the azimuth estimate. Using equation 6 provides better estimate of azimuth since no assumptions are made that the elevation is zero, but rather the elevation term $cos(\theta)$ will cancel each other.

The derivations of the previous equations can be found at [29] and [28].

3 The SRP-PHAT algorithm

The *Steered Response Power (SRP)* algorithm is a beamformer based algorithm that searches for the location that maximizes the SRP function among a set of candidate locations. The *Phase Transform (PHAT)* weighting function has been extensively used in literature and has been shown to work well in real environments, it provides robustness against noise and reverberation. The *SRP-PHAT* algorithms combines the benefits of the SRP beamformer and the robustness of the PHAT weighting function, making it one of the most used algorithms for the acoustic localization task. [30] showed that the SRP function is equivalent to summing all possible GCC combinations, therefore, the SRP-PHAT function can be written in terms of the *Generalized Cross Correlation (GCC)* as:

$$P_{PHAT}(\bar{X}) = \sum_{ij}^{M(M-1)/2} GCC - PHAT(\tau_{ij}(\bar{X})) \tag{8}$$

$$i = 1:M,\ j = 2:M,\ j > i$$

Where M is the number of microphones in the array. $\tau_{ij}(\bar{X})$ is the theoretical time delay between the signal received at microphone i and that at microphone j given the spatial location $\bar{X}$ and is calculated from the geometrical formula, given the spatial locations of the microphones $\bar{m} = [x, y, z]$ and the speed of sound c:

$$\tau_{ij}(\bar{X}) = \frac{||\bar{X} - \bar{m_i}|| - ||\bar{X} - \bar{m_j}||}{c} \tag{9}$$

and $GCC - PHAT(\tau_{ij}(\bar{X}))$ is the value of the *GCC-PHAT* function at the theoretical time delay $\tau_{ij}(\bar{X})$, The GCC-PHAT function can be computed in the frequency domain as:

$$GCC - PHAT_{ij}(\omega) = \psi_{PHAT}(\omega)S_i(\omega)S_j^*(\omega) \tag{10}$$

In equation 10 $S_i(\omega)$ and $S_j(\omega)$ are the acoustic signals in the frequency domain computed by applying the *Fast Fourier Transform (FFT)* to the time domain signals $s_i(\tau)$ and $s_j(\tau)$ recorded from microphones i and j respectively. * is the conjugate operator. ψ_{PHAT} is the PHAT weighting function, it is defined as the magnitude of the *Cross Power Spectrum* between the two microphones signals and can be written as:

$$\psi_{PHAT} = \frac{1}{|S_i(\omega)S_j^*(\omega)|} \tag{11}$$

Substituting 11 into 10 and converting to the time domain:

$$GCC - PHAT_{ij} = \mathcal{F}^{-1}\left[\frac{S_i(\omega)S_j^*(\omega)}{|S_i(\omega)S_j^*(\omega)|}\right] \tag{12}$$

Where $\mathcal{F}^{-1}$ is the *Inverse Fourier Transform*.

finally, a grid of candidate locations $\bar{X}$ is formed and used to evaluate the *SRP* function in equation 8. The candidate location that produces the highest "*Power*" is said to be the location of the sound source.

$$\bar{X} = \boldsymbol{argmax}(P_{PHAT}(\bar{X})) \qquad (13)$$

Finding $\bar{X}$ that maximizes the SRP-PHAT function in the previous equation is a computationally intense problem. The function has several local maxima and a fine grid has to be formed and searched over to get reliable results. In order to alleviate the computational burden of the grid search, optimization based techniques have been adopted and reported in the literature, however, there is no guarantee that these algorithms would find the global maximum of the function, moreover, due to factors such as excessive noise and reverberation or due to discrete calculations and interpolations involved in calculating the SRP-PHAT function, the global maximum of the function can deviate from the true location considerably.

4 Proposed Localization Approach

A two stage approach to the acoustic localization problem is suggested. The aim is to minimize the search area for the SRP-PHAT algorithm and increase the reliability and accuracy of the localization system especially when using low cost compact microphone arrays. The search area is minimized by estimating the DoA of the acoustic location and then forming a boundary around this estimated DoA according to the confidence level of this estimation along with the range of the microphone array. This can significantly reduce the number of maxima in the function since a majority of the original area has been eliminated as a possibility that the acoustic source originated from it. Therefore, the maximum found by the SRP-PHAT algorithm in this minimized area is most likely to be the only dominant peak and hence represents the true location, moreover, this is done in just one step, as the DoA can be estimated in the closed form, unlike optimization algorithms that can spend several iterations to find the peak, that if they did not get stuck in a local maxima. Figure 2 shows the idea of the localization approach.

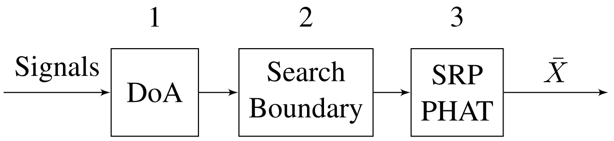

Figure 2: Proposed localization scheme

The DoA obtained from section 2, in the closed form, is an estimate of the true DoA, this is due to several reasons, such as:

- The TDOAs are inaccurate.
- The equations from the previous section are derived based on the *far-field* assumption, therefore, the closer the sound source is to the microphone array the more the error will be in the DoA estimate.
- Microphone array geometry can affect the DoA estimate.
- Low sampling frequency and discrete calculations.

Lets assume that the DoA estimate is contaminated with a zero-mean Gaussian noise ε with a standard deviation σ. The standard deviation is dependent on the inaccuracies in the system mentioned above and hence can be estimated by analyzing the system's errors or through experiments on the localization system.

$$\hat{\theta} = \theta - \varepsilon_1 \qquad (14)$$

$$\hat{\phi} = \phi - \varepsilon_2 \qquad (15)$$

By sampling N_1 points from the Normal distribution $\mathcal{N}_1(0, \sigma_1^2)$ and N_2 points from $\mathcal{N}_2(0, \sigma_2^2)$, where σ_1 and σ_2 are the standard deviations representing the errors in the azimuth and elevation respectively, $N_1 \times N_2$ permutations of "possible" azimuth and elevation angles are formed. Applying these angles to equation 16 assuming N_3 points of $r \in [0, r_{max}]$, where r_{max} is the acoustic range of the microphone array, a point cloud of $N_1 \times N_2 \times N_3 \quad xyz$ points is produced. Figure 3 shows an example of the search boundary produced from equation 16 in the 2D plane.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = r \times \begin{bmatrix} \sin(\phi)\cos(\theta) \\ \sin(\phi)\sin(\theta) \\ \cos(\phi) \end{bmatrix} \qquad (16)$$

5 Results

In order to evaluate the proposed approach the "Audio Visual AV16.3" corpus was used [31]. The audio corpus of the *AV16.3* is recorded in a meeting room context by means of two 0.1m radius *Uniform Circular Arrays (UCA)* with 8-microphone elements each at a sampling frequency of 16 kHz. Table 1 shows the xyz locations of the first of the two arrays, the reference point is the middle point between the two arrays.

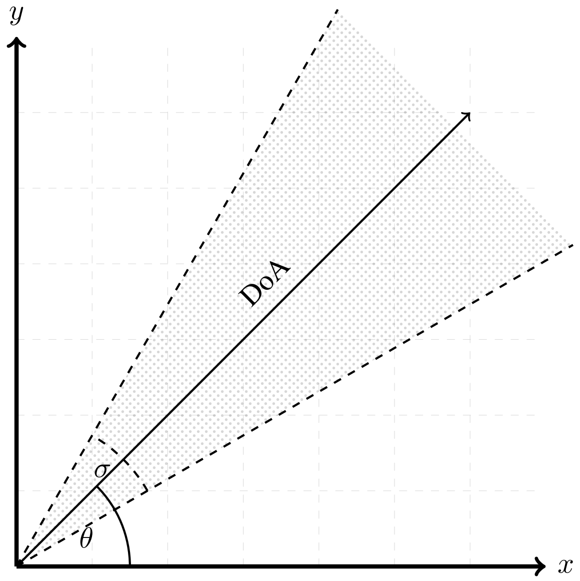

Figure 3: Search Boundary in the xy plane

The two UCAs are at plane $Z = 0$. The audio corpus consists of 8 annotated sequences in a variety of situations, for this work sequence "*seq01-1p-0000*" is used. It was recorded for the purpose of sound source localization evaluation, the recording spans over 217 seconds for a single speaker at 16 different locations, static at each location and recording 10 to 11 segments at each location.

Table 1: Microphone locations of AV16.3 first array

Microphone no.	X(m)	Y(m)
m1	-0.1	0.4
m2	-0.07071	0.32929
m3	0	0.3
m4	0.07071	0.32929
m5	0.1	0.1
m6	0.07071	0.47071
m7	0	0.5
m8	-0.07071	0.47071

Table 2: Algorithm input parameters

N	$r(m)$	$\sigma_1(rad)$	Window(mSec)	% overlap
1000	3	0.1	100	50

In the proposed approach the user is required to input only minimal settings namely: the number of points $N = N_1N_2N_3$ which will be used to fill the boundary area/volume (for 2D or 3D), the maximum range of the microphone array r and the standard deviations σ_1 and σ_2 that represent the error in the estimated azimuth and elevation. In addition to the frame window length and percentage overlap if required.

for the experiments presented here these settings are shown in table 2. Since the microphones in the UCA of the AV16.3 corpus are distributed along the x and y axes only ($z = 0$), it is impossible to calculate the elevation part of the DoA vector, therefore, the boundary area is formed using the azimuth only hence forming a 2D area represented by a triangle as shown in fig. 3, setting $\theta = 0$ in equation 16 x and y are calculated from $x = rcos(\phi)$ and $y = rsin(\phi)$ and the z component is appended as uniform random values covering from the floor to the ceiling of the room $z \in U(z_{min}, z_{max})$. The z component was added this way and was not ignored because experiments showed that it had significant effect on the overall localization results. In table 2 N is the *total* number of points that are distributed in the boundary volume; the distribution around the azimuth is a Gaussian with standard deviation σ_1 while the z values follow a uniform distribution from the floor to the ceiling of the room and has the same length N and was appended to the x and y values. The window used for the *Fourier analysis* is a *Hanning* window.

The measure used for the evaluation of the proposed approach is the *Root Mean Square Error (RMSE)* between the estimated location and the ground truth available in the AV16.3 corpus, the RMSE is calculated from:

$$RMSE = \sqrt{(x_{gt} - x_{est})^2 + (y_{gt} - y_{est})^2 + (z_{gt} - z_{est})^2} \quad (17)$$

Where subscripts *gt* and *est* stand for the ground truth and estimated values respectively.

Since the proposed approach uses random numbers to fill in the search boundary, it is expected that this approach would result in different results at each run, therefore, each experiment at each location was run 1000 times and the RMSE of each run was recorded and the variance of these $n = 1000$ runs was calculated from equation 18 and reported to show that the proposed approach has a low variance, i.e. will give consistent results at each run. Moreover, the minimum and maximum as well as the mean of these 1000 runs were reported and compared to the results of the conventional SRP-PHAT algorithm.

$$\sigma^2 = \frac{\sum (RMSE - \mu)^2}{n} \quad (18)$$

As mentioned, the results from the proposed approach were compared to those of the conventional

Table 3: Variance of the RMSE

	$\sigma^2 \times 10^{-3}$									
	Segments									
Locations	4.97	20.5	5.96	13.6	11.7	4.6	5.23	8.08	11.5	23.1
	30.4	24.2	38.2	26.7	36.9	63.3	31.7	28.8	13.9	16.5
	50.3	47.2	46.3	41.9	35	52	41.4	37.7	44.8	61.1
	30.5	7.31	36.3	30.4	32.2	12.5	18.1	39.5	62.9	32.4
	3.57	0.833	15.1	32.7	44.2	43.6	48.4	1.47	5.42	31.7
	10.9	10.5	13.2	9.84	11.7	15.2	7.44	10.4	32.1	4.04
	22.8	39.3	42	22.2	23.2	21.4	37.7	31.6	43.6	17.7
	22.2	28.3	19.9	19.1	9.89	16.1	15.9	5.66	6.88	47
	31.6	28.9	18.2	16.7	8.83	40.8	26.9	32.7	10.9	-
	30.6	7.06	33.9	45.5	34.7	107	24.2	28.5	38.7	33.3
	34.1	264	65.6	50.2	21	42.2	39.6	44.8	84.4	12.6
	24.1	53.3	56.5	50.1	35.1	41.7	43.6	28.1	30.4	61.9
	7.78	70.4	86.6	33.7	8.52	8.54	9.53	19.6	43.5	20.3
	18.3	18.5	14.1	23.7	8.93	16.7	24.3	22.8	30.2	17.7
	65	48.6	42.7	36.5	30.9	29.8	27.1	36.5	31.6	48.3
	58.7	37	68.5	88.7	83.6	56.2	53.3	93.8	68.3	94.6

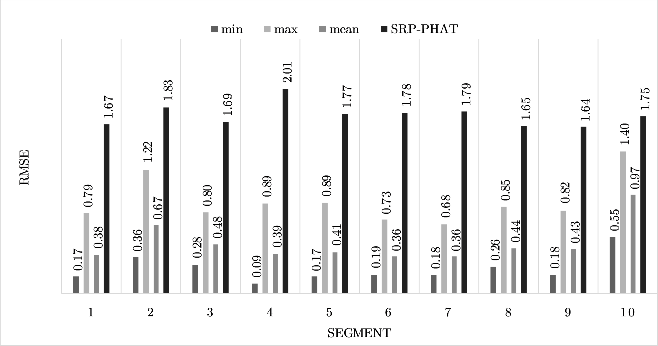

Figure 4: Results for location 1

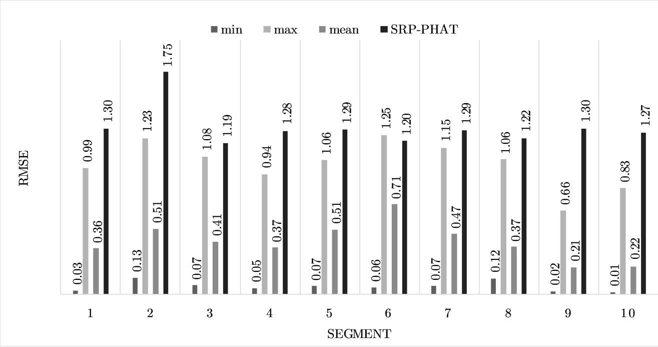

Figure 5: Results for location 2

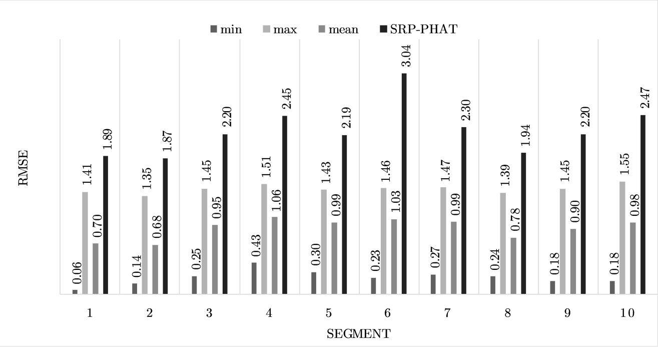

Figure 6: Results for location 3

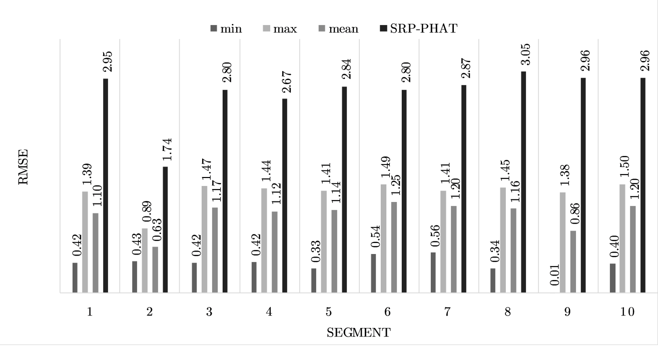

Figure 7: Results for location 4

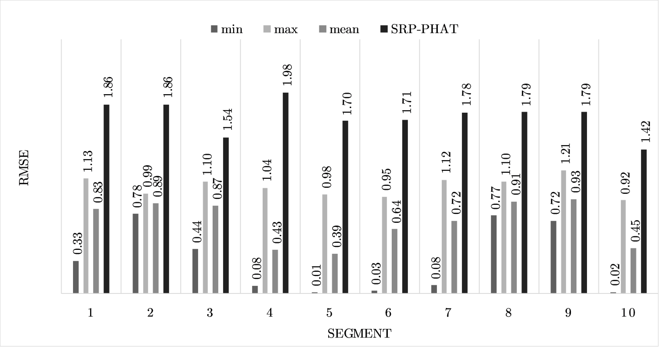

Figure 8: Results for location 5

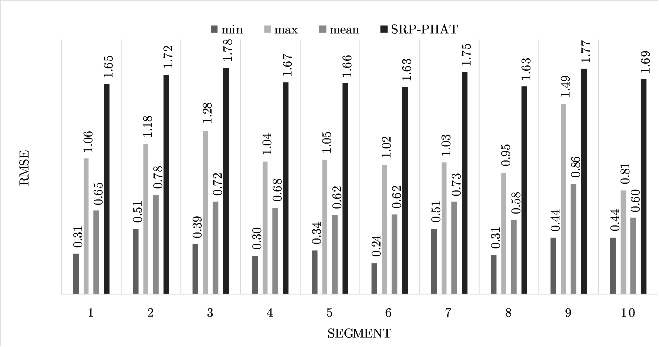

Figure 9: Results for location 6

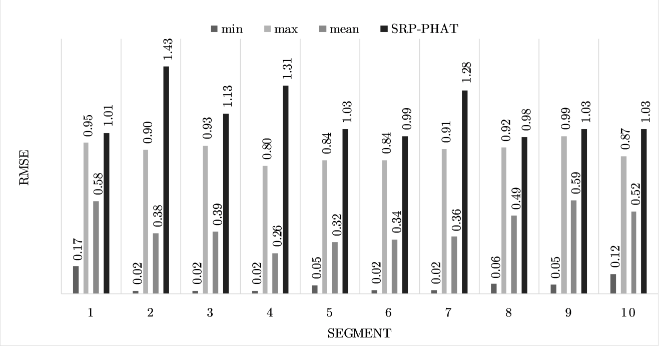

Figure 10: Results for location 7

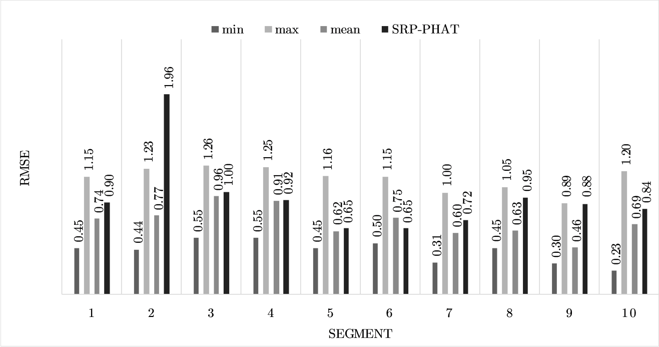

Figure 11: Results for location 8

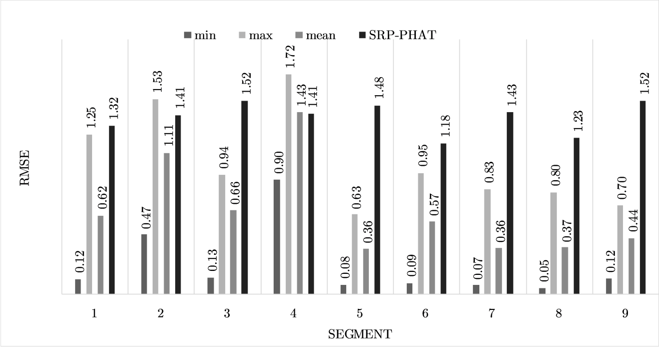

Figure 12: Results for location 9

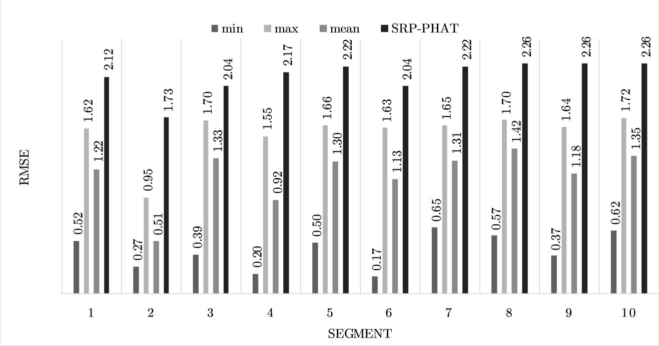

Figure 13: Results for location 10

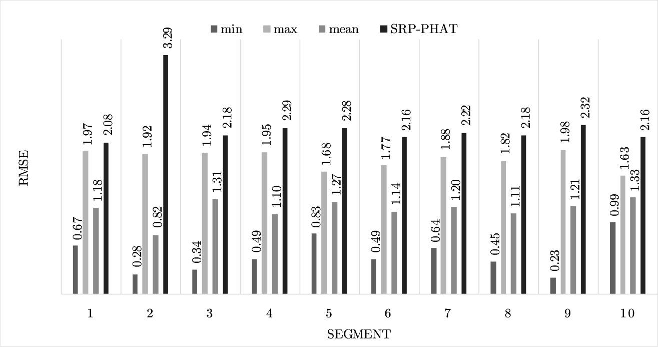

Figure 14: Results for location 11

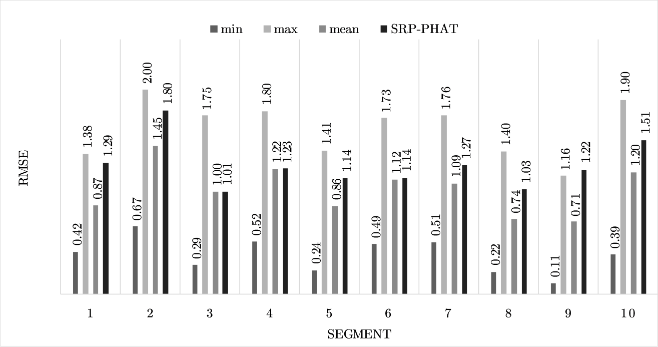

Figure 15: Results for location 12

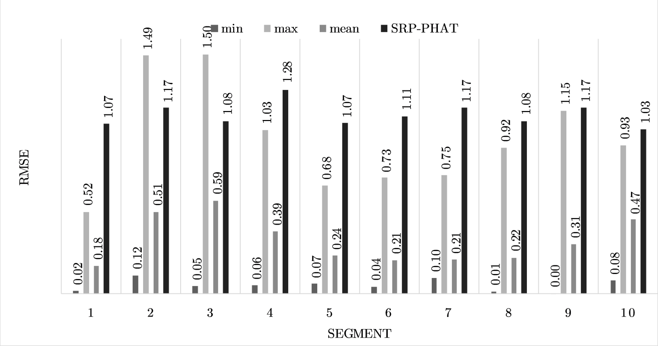

Figure 16: Results for location 13

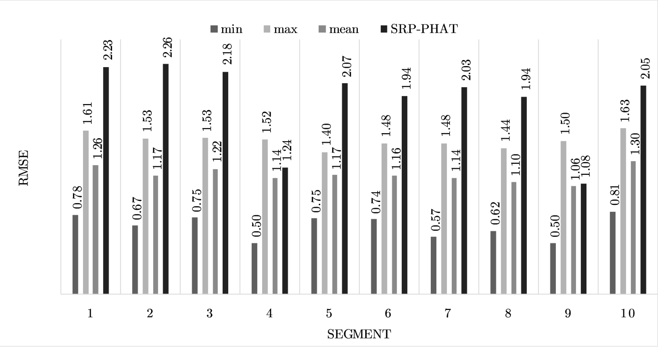

Figure 17: Results for location 14

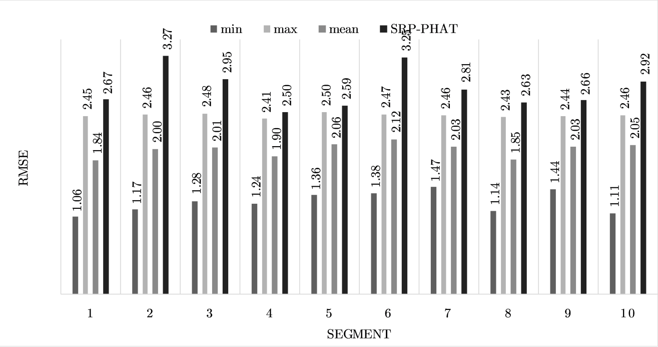

Figure 18: Results for location 15

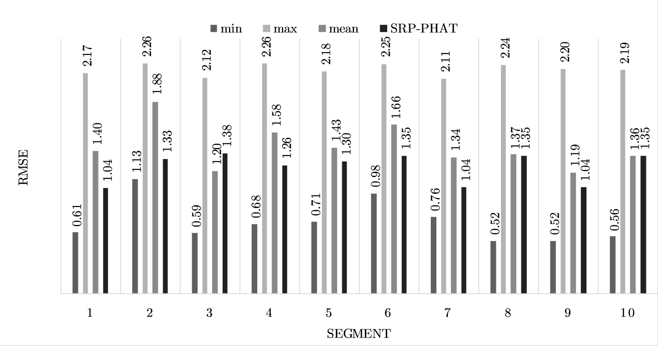

Figure 19: Results for location 16

SRP-PHAT algorithm. The settings of the SRP-PHAT algorithm were the same as our proposed approached, i.e. the same window and overlap. A mesh-grid of 125000 points in 3D was formed and fed to the SRP-PHAT algorithm for the search process, no optimizations were used in this search process. The conference room of the AV16.3 audio corpus was $8.2m \times 3.6m \times 2.4m$, hence the mesh-grid was formed by taking 50 linearly spaced points along the x, y and z axes creating a $50 \times 50 \times 50$ grid.

The variance resulted from the experiments on all 16 locations and 10 segments for each location is reported in table 3. From the table it is clear that the proposed approach has low variance.

Fig. 4 to fig. 19 show the results for each of the 16 locations separately. Each figure compares the minimum, maximum and mean RMSE of the proposed approach to that of the conventional SRP-PHAT algorithm. As it is clear from the graphs that the proposed approach yields lower RMSE results, in the majority of cases, than the conventional SRP-PHAT algorithm, even when comparing the maximum RMSE value. In locations 8,9,12,13,14 and 16 in fig. 11, 12, 15,16, 17 and 19 it was noticed that the RMSE of the SRP-PHAT algorithm was in some segments lower than the *maximum* RMSE of the proposed approach, this can be attributed to the low number of points of the proposed approach as compared to the high number of points used for the SRP-PHAT algorithm, this and the value of σ can affect the results; if σ is unrealistically small, i.e. has an overoptimistic, a search boundary can be formed where the true source location point lies on the boundaries or even outside the search area, and because of the Gaussian assumption, the points near the edge of the boundary are less represented that those around the mean DoA.

6 Conclusion

This paper presented a robust approach for the Sound Source Localization (SSL) problem. The proposed approach was developed with the aim to work with compact microphone arrays at low sampling frequencies and low computational burden, hence making it suitable to be used with small mobile robots. The proposed approach is based on a mixed far field/near field model where as a first step the DoA vector is estimated in the closed form using the far field model, then, a search boundary is formed based on the DoA vector, the expected error in the DoA estimates and the microphone array range, finally, based on the near field model, this boundary is searched using the conventional SRP-PHAT algorithm to find the source location. The AV16.3 corpus was used to evaluate the proposed approach, extensive experiments have been carried out to verify the reliability of the approach. The results showed that the proposed approach was successful in obtaining good results compared to the conventional SRP-PHAT algorithm eventhough only 1000 points were used for the search process as opposed to 125000 used by the SRP-PHAT algorithm. Minimum user input is required to run the algorithm, namely, the number of points to fill the search boundary, the microphone array range and the expected error in the DoA estimation. Obviously, by increasing the number of points in the search boundary the resolution will increase but so will the number of functional evaluations and hence the computational burden, but it was shown that even while using small number of points good results can be obtained. The expected error in the DoA estimation σ_1 and σ_2 depends on factors related to the microphone array system as well as factors related to the environment. The number of microphone elements in the array, their types and the array's geometry are some of the factors that affect the DoA estimation, moreover, factors such as the sampling frequency and discrete calculations and others contribute to this error and hence affects the values of σ_1 and σ_2. In addition, noise and revereberation and other environmental factors obviously affects σ_1 and σ_2 and cannot be easily predicted. All these factors make the calculation of σ_1 and σ_2 rather a difficult task. In this work σ_1 was figured by observing some experiments and figuring out the DoA error of each experiment. Finally, it should be mentioned that in the experiments carried out in this paper no efforts have been done to improve the SNR of the signals except for a simple second order band pass filter (300Hz-6kHz) and this was applied to the proposed approach and to the SRP-PHAT algorithm. It is expected that using some further denoising techniques would further improve the results, moreover, it is possible to use some optimization techniques to search for the peak power in the search boundary instead of performing a point by point search.

References:

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

[2] P. G. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Applications of Signal Processing to Audio and Acous-*

tics, 1997. 1997 IEEE ASSP Workshop on, pp. 4–pp, IEEE, 1997.

[3] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 495250, 2003.

[4] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.

[5] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 509–519, 2004.

[6] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 157–160, 2007.

[7] H. He, J. Lu, L. Wu, and X. Qiu, "Time delay estimation via non-mutual information among multiple microphones," *Applied Acoustics*, vol. 74, no. 8, pp. 1033–1036, 2013.

[8] J. Thyssen, A. Pandey, and B. J. Borgström, "A novel time-delay-of-arrival estimation technique for multi-microphone audio processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 21–25, IEEE, 2015.

[9] H. He, J. Chen, J. Benesty, Y. Zhou, and T. Yang, "Robust multichannel tdoa estimation for speaker localization using the impulsive characteristics of speech spectrum," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 6130–6134, IEEE, 2017.

[10] J. Choi, J. Kim, and N. S. Kim, "Robust time-delay estimation for acoustic indoor localization in reverberant environments," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 226–230, 2017.

[11] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 769285, 2003.

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

[13] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," Institute of Electrical and Electronics Engineers, 2009.

[14] J. Liang and D. Liu, "Passive localization of mixed near-field and far-field sources using two-stage music algorithm," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 108–120, 2010.

[15] F. Grondin and J. Glass, "Svd-phat: A fast sound source localization method," *arXiv preprint arXiv:1811.11785*, 2018.

[16] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1, pp. 375–378, IEEE, 1997.

[17] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 499–508, 2004.

[18] H. Do, H. F. Silverman, and Y. Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. I–121, IEEE, 2007.

[19] H. Do and H. F. Silverman, "A fast microphone array srp-phat source location implementation using coarse-to-fine region contraction (cfrc)," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pp. 295–298, IEEE, 2007.

[20] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[21] H. Do and H. F. Silverman, “Stochastic particle filtering: A fast srp-phat single source localization algorithm,” in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA’09. IEEE Workshop on*, pp. 213–216, Citeseer, 2009.

[22] B. Lee and T. Kalker, “A vectorized method for computationally efficient srp-phat sound source localization,” in *12th International workshop on acoustic echo and noise control (IWAENC 2010)*, 2010.

[23] M. Cobos, A. Marti, and J. J. Lopez, “A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling,” *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[24] G. Jacovitti and G. Scarano, “Discrete time techniques for time delay estimation,” *IEEE Transactions on signal processing*, vol. 41, no. 2, pp. 525–533, 1993.

[25] D. L. Maskell and G. S. Woods, “The estimation of subsample time delay of arrival in the discrete-time measurement of phase delay,” *IEEE Transactions on Instrumentation and Measurement*, vol. 48, no. 6, pp. 1227–1231, 1999.

[26] K. K. Sharma and S. D. Joshi, “Time delay estimation using fractional fourier transform,” *Signal processing*, vol. 87, no. 5, pp. 853–865, 2007.

[27] S. Tervo and T. Lokki, “Interpolation methods for the srp-phat algorithm,” *Proc. of 11th IWAENC*, 2008.

[28] P. Schober, *Source Localization with a Small Circular Array*. PhD thesis, JOHANNES KEPLER UNIVERSITY LINZ, 2017.

[29] J. Smith and J. Abel, “Closed-form least-squares source location estimation from range-difference measurements,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[30] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000.

[31] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “Av16. 3: an audio-visual corpus for speaker localization and tracking,” in *International Workshop on Machine Learning for Multimodal Interaction*, pp. 182–195, Springer, 2004.