TDE Sign Based Homing Algorithm for Sound Source Tracking Using a Y-shaped Microphone Array

Sreejith T. M.†, Joshin P. K.†, Harshavardhan S., T. V. Sreenivas
Department of Electrical Communication Engineering,
Indian Institute of Science
Bangalore, India-560012
Email: {sreejithtm.tvm, joshinpk}@gmail.com, {harsha, tvsree}@ece.iisc.ernet.in

***Abstract*—Sound source localization, tracking and homing in play an important role in human-robot interaction. The frequent tracking of a moving acoustic source is computationally expensive for any real time application. In this paper, we propose a simple algorithm based on the sign of time delay of arrival estimation (STDE). Also a special Y-array microphone is designed which is well suited for real time tracking and homing in application. Compared to the other conventional planar microphone arrays, Y-array has the unique property of proximity detection of the acoustic source. The STDE algorithm along with the Y-array is used to estimate the gross source region successively for real time robotic homing in applications. Experiments conducted in anechoic and varechoic enclosures show that the proposed algorithm works accurately in low and moderate reverb conditions of upto RT60<800ms.**

Keywords*—real-time homing in, STDE, gross source region, voronoi region.

I. Introduction

Robots are becoming affordable devices that can be used in everyday life. For effective interaction with humans, the robots should be able to communicate in a similar way as human do. One of the fundamental requirements for a robot to interact with humans is the ability to locate and track humans as they move. If the robot can be steered towards the source of speech sound, the SNR of the received signal will get improved, there by improving the performance of any further signal processing algorithm implemented on the robot.

Existing sound source localization procedures may be broadly classified into three general categories: 1) steered beam-former based approaches, 2) platform effect based approaches, 3) TDE based approaches. The beam-forming based approach demands a large number of microphone for achieving higher spatial resolution. The problem of multiple sound source localization can be addressed using this method [1]. However, the requirement of large number of microphones makes this approach less suitable for robotic applications [2]–[4]. The platform effect based localization techniques require the knowledge of the shape of the platform and microphone locations. For instance, Head Related Transfer Function (HRTF) based methods of localization is an emulation of the localization technique of human's auditory system in which direction dependent filtering is done [5]. This type of localization techniques are platform specific and cannot be remounted from one platform to other without the knowledge of the HRTF data base of the new platform [6], [7]. The TDE based methods are widely used because they require lesser number of microphones with low computational complexity [7]–[10]. The basic idea behind the conventional TDE based algorithm is to find the intersection of the half-hyperboloids corresponding to the TDEs of different microphone pairs. But the sampling rate of the signal and the quasi-stationary segment of estimation affect the accuracy of TDEs.

Here, we propose a STDE based algorithm for gross source region estimation. Since we are considering only the sign of TDE, tracking algorithm complexity is reduced. Also, since the algorithm does not demand the precise estimation of TDE, the time frame of observation signal can be reduced considerably. A smaller time frame window in turn provides for quasi-stationarity (the assumption that the acoustic source is not moving with respect to the listener during the time frame of observation) to accommodate the relative motion between source and the robot.

An application scenario in which a robot is homing in on a single sound source (human) is addressed in this paper. Considering the height of a normal robot and a human speaker, the exact co-ordinate of the speaker need not be found for the homing in application. If the gross source region is estimated, robot can be steered towards the source. A successive approximation algorithm is developed, which is well suited for real time homing in application. This is accomplished by using a planar microphone array and the STDE algorithm. We consider a specific planar microphone array configuration and the STDE to divide the space around the robot using $\binom{n}{2}$ division planes (n is the number of microphones in the array and the division planes are perpendicular to the plane of the microphone array). The conventional planar array lack the ability for range detection though the direction of arrival can be estimated accurately [10]–[12]. The new microphone configuration provides for the unique feature of a "home-region" to detect the source proximity and stop the robot motion.

†These two authors contributed equally to the work presented.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

II. Estimating Gross Source Region Using STDE in a Single Speaker Environment

We define time delay estimate (TDE) as the difference in time of arrival (TOA) of the acoustic signal reaching a pair of microphone from a single sound source (speaker). The proposed algorithm for homing in makes use of the sign of TDE for the estimation of gross source region. The locus of an acoustic source having a particular TDE between two microphones M_1 and M_2 is shown in Fig. 1 .

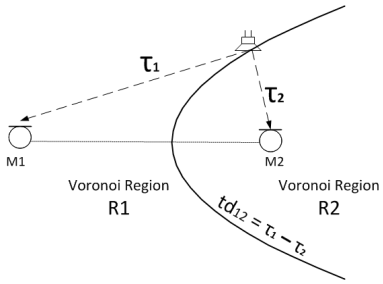


Fig. 1. TDE based voronoi region formation

Locus of the point in the 3-D space of (M_1, M_2) so that the difference of its distances from two fixed points is constant, is a half-hyberboloid. Since the distance traveled by the acoustic signal is directly proportional to the time taken to travel through the medium, locus of an acoustic source having a constant TDE from two microphones is a half-hyperboloid.

The gross source region can be defined as the set of points in the space of the microphone pair that obeys a particular condition on TDE. For example, in Fig. 1, two voronoi regions are formed.

$$R1 = \{(x, y, z) : td_{12} < threshold\} \quad (1)$$
$$R2 = \{(x, y, z) : td_{12} > threshold\} \quad (2)$$

where td_{12} is the TDE given by (5) and (x,y,z) are the rectangular co-ordinates of the points.

Here, for any non-zero value of the threshold, the boundaries separating the voronoi regions will be half-hyperboloid. As the value of TDE tends to zero, the eccentricity of the half-hyperbolod tends to infinity and the division boundary becomes a hyperplane as shown in Fig. 2.

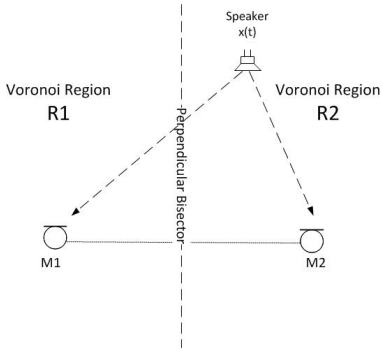


Fig. 2. STDE based voronoi region formation.

A. TDE Estimation using GCC-PHAT

GCC-PHAT [13] is used to calculate the TDEs. Let $x(t)$ be the sound signal produced by the speaker and $x_1(t)$ and $x_2(t)$ be the signals reaching the microphones M_1 and M_2 respectively. Let τ_1 and τ_2 be the time taken by the acoustic signal $x(t)$ to travel in their respective paths with attenuation α_1 and α_2 respectively. Then the signal model can be written as

$$x_1(t) = \alpha_1 x(t - \tau_1) \quad (3)$$
$$x_2(t) = \alpha_2 x(t - \tau_2) \quad (4)$$
$$td_{12} = \tau_1 - \tau_2 \quad (5)$$

The normalized cross power spectral density is given by

$$S_{12}(\omega) = \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)||X_2(\omega)|} \quad (6)$$

Where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of $x_1(t)$ and $x_2(t)$. Before applying Fourier transforms, zero padding is done to convert circular convolution into linear convolution. The GCC-PHAT is obtained by taking the inverse Fourier transform of $S_{12}(\omega)$

$$R_{12}(\tau) = \int_{-\infty}^{\infty} \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)||X_2(\omega)|} e^{jw\tau} dw \quad (7)$$

The cross correlation thus obtained is maximized over all the possible time delays to get TDE

$$td_{12} = \underset{\tau}{\operatorname{argmax}} \left\{ \int_{-\infty}^{\infty} \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)||X_2(\omega)|} e^{jw\tau} dw \right\} \quad (8)$$

A positive value of td_{12} implies that source is located in the right half space ($R2$) and a negative value implies that source is located in the left half space ($R1$). Thus,

$$STDE_{12} = sign(td_{12}) \quad (9)$$

III. Voronoi Region Formation Using Array of Microphones

A. A symmetric array case

A symmetric array using four microphones is shown in Fig. 3, where M_i, $i = 1, 2, 3, 4$ represents the four microphones.

The dotted lines (line-$m_i m_j$) are the perpendicular bisectors of the microphone pairs (M_i, M_j) which represent the locus of zero TDE points for the corresponding microphone pair. The STDEs for the four pairs of microphones are calculated to divide the space into eight voronoi regions. By analyzing these STDEs, we can determine as to which sector the acoustic source belongs. The GCC-PHAT [13] between the signals reaching all the pairs of microphones are used to compute the TDEs as explained in (8) .

The Robot (PC-BOT) considered has only forward motion and hence for source homing in, a microphone array with more number of voronoi regions towards the front can help the robot to steer more accurately and efficiently towards the source. Also in STDE based localization algorithm that uses a symmetric microphone array, it is not possible to design a stopping criteria for the robot when it falls within the proximity of the source. This motivates the necessity of a well designed microphone array for the homing in applications.

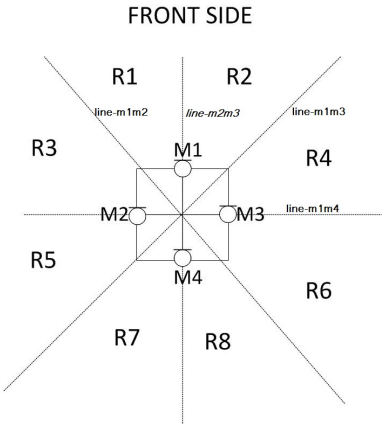


Fig. 3. Symmetric array of four microphones

B. 'Y' Shaped Microphone Array

Fig. 4 shows the microphone array arrangement in which M_i, $i = 1, 2, 3, 4$ represents the four microphone configuration. This is a co-planar geometry in which microphones are arranged in a non-linear manner, in a 'Y' shape. Using the STDE between each pair of microphones, the entire space of microphone array is divided into nine voronoi regions as shown in Fig. 4. This non-linear array is designed and configured such that it has the following advantages over the usual linear and non-linear array configurations in the literature.

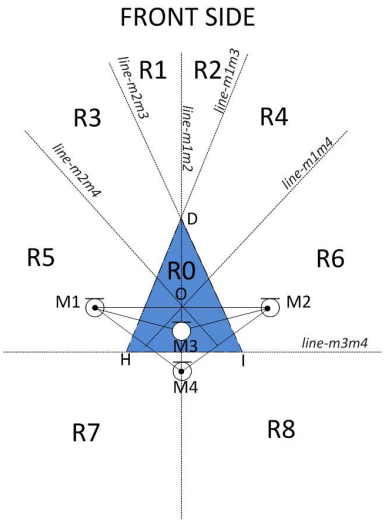


Fig. 4. Voronoi Region Formation of 'Y' array

Front Resolution: The geometric configuration of the microphones is such that the voronoi regions to the front of the robot are narrower compared to that at the back side. The sector resolution (SR) of different voronoi regions (SR_i; i = 0, 1, 2, ..., 9) are given by $SR_1 = SR_2 < SR_3 = SR_4 < SR_5 = SR_6 < SR_7 = SR_8$. This provides better sector resolution at the front side compared to a symmetric geometry (III-A), where the sector resolution is uniform. This makes the robot capable of switching its search process from coarse mode to fine mode as it heads towards the acoustic source, thereby improving the homing in time.

Home-region Formation: Fig. 4 shows ΔDHI, a triangle shaped home-region (R0) at the front side, which can be used to determine whether the source is in the proximity of the robot or not. R0 is defined as,

$$R0 = \{(x, y, z) : td_{13} > 0, td_{23} > 0, td_{34} < 0\} \qquad (10)$$

When the source is detected within this home-region, it indicates that the robot has approached close to the source and this can be taken as the criteria to stop the robot motion. The shape of this triangular search area can be easily designed and is described in section III-C.

Calibration Free Microphones: Home-region based stopping criteria doesn't make use of any source distance calculations or signal energy measurements. This avoids the requirement of calibrated microphones. Since the algorithm is based only on the STDE calculation, the microphones can be any type of other omnidirectional microphones and we need not change the design parameters. This makes the 'Y' array design robust to a wide range of microphones.

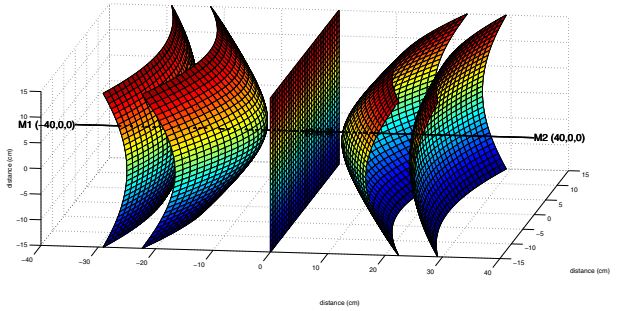


Fig. 5. Half-hyperboloid loci for five different TDE

Localization in 3-D Space Using Coplanar Array: The voronoi region for the co-planar microphone configuration, naturally does not differentiate the height co-ordinate of the source. This choice is well suited to the specific application, since the robot can move only in 2-D plane. Thus for each microphone pair of the Y-array, the voronoi region boundary in 3-D space is half-hyperboloid as shown in Fig. 5; i.e, with two microphones M_1, M_2 placed on the X-axis, equidistant from origin, five different loci are shown by considering five different TDEs . For a non-zero TDE, the locus will always be a hyperboloid lobe as explained in section II. Since we are using STDE based sector classification, the decision boundary in 3-D space will be a plane which is perpendicular to the horizontal plane of the microphone array and is free from any height dependent ambiguity.

C. Design of Microphone Geometry

We first position the microphones M_1, M_2 and M_3 (see Fig. 6) to form a triangular shaped home-region ΔDHI. The orthocenter of ΔABC is kept outside the triangle (orthocenter because of the use of STDE to form voronoi regions). To bring the orthocenter (D) outside ΔABC, the angle θ should be an acute angle i.e, the distance 'x' in Fig. 6 should be much less compared to the distance d/2. Also we have to fix the area of the home-region. In order to design the microphone array satisfying all these constraints, a relationship between

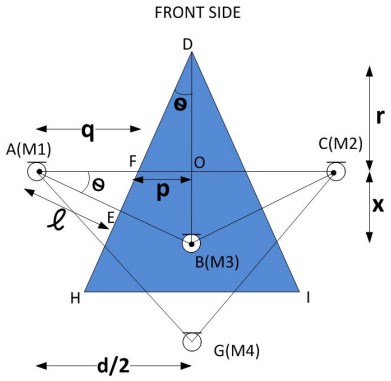


Fig. 6. Design of Microphones Geometry

'x', 'd' and 'r' is derived as shown below. In Fig. 6, the line DH intersects the lines AB and AC at the points E and F respectively. Also, the line DB intersects the line AC at the point O.

From ΔAOB, ΔAEF

$$\ell = \frac{\sqrt{\frac{d^2}{4} + x^2}}{2} \quad (11)$$

$$p = \frac{d}{2} - \frac{\ell}{\cos\theta} \quad (12)$$

$$\cos\theta = \frac{d}{\sqrt{4x^2 + d^2}} \quad (13)$$

From ΔDOF, ΔAOB,

$$\tan\theta = \frac{2x}{d} = \frac{p}{r} \quad (14)$$

Eliminating 'θ', 'ℓ' and 'p'

$$2rx = \frac{d^2}{4} - x^2 \quad (15)$$

or

$$x = r\left(\sqrt{1 + \frac{d^2}{4r^2}} - 1\right) \quad (16)$$

Thus, the stopping distance 'r' determines 'x', the position of M_3. Often '$\frac{d}{2}$' is small and hence 'x' is quite small for a reasonable 'r'.

We can position M_4 as shown in Fig. 4. The purpose of fourth microphone is to form the division lines line-m3m4, line-m1m4 and line-m2m4. Also, it determines the relative resolution between the voronoi regions (R_4, R_6) and (R_3, R_5). It is mounted on the perpendicular bisector of the microphones M_1 and M_2 (line-m1m2) such that it is within the top horizontal plane of the robot.

D. Homing in by Successive Estimation:

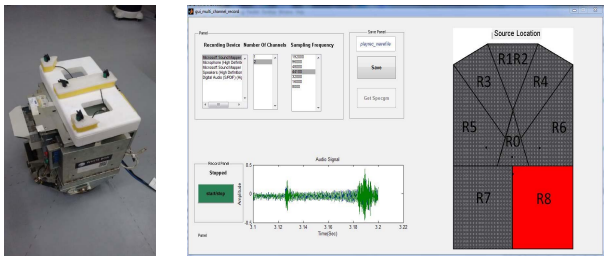
A quasi-stationary speech segment of 100ms is considered during which it is assumed that the relative motion between the robot and source is negligible. This assumption is reasonable as most of the human-robot interactive application requires low and moderate velocity robots. In each 100ms, the STDEs between all the six pairs of microphones are determined and the voronoi region of the acoustic source is estimated. The robotic wheel control is done in such a way that the robot turns to bring the acoustic source in the front-end voronoi regions R_1 and R_2. If the source is detected in the back-end regions R_7 and R_8, the robot is made to turn with maximum possible velocity and as the source falls in front R_1 and R_2 sectors, the turning velocity is reduced. Once the source falls in R_1 or R_2 region, the robot's turning velocity is made zero and it is made to move forward with high velocity. Once the home-region is detected, the wheel motion is stopped and hence the robot.

Since the robot wheel control cannot be updated for every 100ms, a maximum vote of ten estimates detected in a second is taken and is declared as the voronoi region for that one second. The wheel control is updated every one second based on this voted voronoi region.

IV. Experiments and Results

A. Performance Evaluation of Y-Array in Real Environment

The performance of the Y-array designed for homing in application is evaluated in real acoustic environment using Monte Carlo experiments. Four omnidirectional microphones are mounted on a wooden platform fixed at the horizontal top plane of the robot as shown in Fig. 7.



'Y' array GUI to display gross source region

Fig. 7. Experimental setup to evaluate the performance of 'Y' array

The orthocenter distance ('r' in Fig. 6) is chosen as 100 cm. The distance between microphones M_1 and M_2 ('d' in Fig. 6) is 30 cm. Using (16), the position of microphone M_3 is calculated as x = 1.1 cm. With this values of 'x', 'd', 'r', the home-region ΔHDI (see Fig 6) is characterized by the angles $\angle D = 8.38°$ and $\angle H = \angle I = 85.81°$. To evaluate the localization performance of the microphone array, the robot is kept stationary in a room where voronoi regions are marked on the floor. A single speaker environment is considered in which recorded human voice played through a mobile phone speaker is used as the acoustic source. Circles of different radii are drawn on the floor keeping the position of microphone M_3 as the centre. For instance, Fig.8 shows a typical source path keeping radius as 3 meters. The acoustic source is moved randomly along the arc of the circle drawn in each voronoi region. The experiment is conducted for 100 seconds along each arc and voronoi regions are calculated and noted in every 100 ms time frame. The percentage error is calculated as

$$Percentage\ error = \frac{No.\ of\ wrong\ detections}{Total\ no.\ detections} \times 100 \quad (17)$$

The experiment is repeated with circles of different radii (Table I). Also, home-region detection is tested by moving the acoustic source randomly within the triangular shaped voronoi

region R_0 for 100 seconds and calculating the percentage error as explained by (17) (Table II)

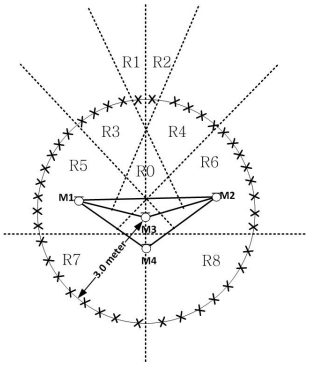


Fig. 8. Source positions for evaluating 'Y' array performance

The performance is evaluated under different reverberation conditions of the enclosure (Varechoic chamber). The results of the experiments conducted in anechoic chamber (zero reverberation) and varechoic chamber (RT60 = 800ms and 2000ms) are shown in Tables I and II.

Voronoi Region	Anechoic Chamber		Varechoic Chamber			
	error(%) (radius 1.5m)	error(%) (radius 2m)	RT60 = 800 ms		RT60 = 2000ms	
			error(%) (radius 1.5m)	error(%) (radius 3.0m)	error(%) (radius 1.5m)	error(%) (radius 3.0m)
R1R2	0.60	0.20	0.40	1.20	10.00	50.80
R3	0.20	0.30	0.50	7.40	2.3	74.60
R4	0.20	0.40	0.20	7.20	14.6	54.30
R5	0.10	0.30	0.50	5.20	6.3	90.10
R6	0.70	0.20	0.20	0.50	12.9	70.30
R7	0.10	0.20	0.30	1.00	7.2	52.20
R8	0.20	0.20	0.10	2.10	2.2	22.60

TABLE I. PERCENTAGE ERROR IN VORONOI REGION DETECTION

Home-region Detection	Anechoic Chamber error(%)	Varechoic Chamber RT60 = 800ms error(%)	Varechoic Chamber RT60 = 2000ms error(%)
	0.20	10.30	11.20

TABLE II. PERCENTAGE ERROR IN HOME-REGION DETECTION

B. Real-time Homing in Experiment:

The Y-array microphone arrangement is mounted on the horizontal top plane of the mobile robot (White box robotics: 914 PC-BOT). An acoustic source is placed in the same room where the robot is deployed. In low and moderate reverb condition, the robot successfully tracked the acoustic source. A demonstration video of the robot homing in on an acoustic source in real time is provided in [14].

V. CONCLUSIONS

In this paper, we addressed the problem of homing in on an acoustic source by a mobile robot using a novel 'Y' shaped microphone array design. The homing in is achieved by successive estimation of the gross region of the acoustic source by using the STDEs. The proposed approach is shown to perform accurately in both anechoic condition and reverb condition with RT60 upto 800ms.

REFERENCES

[1] Y. Cho, D. Yook, S. Chang, and H. Kim, "Sound source localization for robot auditory systems," *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 3, pp. 1663–1668, August 2009.

[2] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1526–1540, 2004.

[3] Y. Tamai, S. Kagami, Y. Amemiya, Y. Sasaki, H. Mizoguchi, and T. Takano, "Circular microphone array for robot's audition," in *Sensors, 2004. Proceedings of IEEE*. IEEE, 2004, pp. 565–570.

[4] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi, "Three ring microphone array for 3d sound localization and separation for mobile robot audition," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*. IEEE, 2005, pp. 4172–4177.

[5] M. Rothbucher, M. Durkovic, T. Habigt, H. Shen, and K. Diepold, "Hrtf-based localization and separation of multiple sound sources," in *RO-MAN, 2012 IEEE*, Sept 2012, pp. 1092–1096.

[6] J. Ferreira, C. Pinho, and J. Dias, "Implementation and calibration of a bayesian binaural system for 3d localisation," in *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference on*, Feb 2009, pp. 1722–1727.

[7] H. Li, T. Yosiara, Q. Zhao, T. Watanabe, and J. Huang, "A spatial sound localization system for mobile robots," in *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*, May 2007, pp. 1–6.

[8] "A model-based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, vol. 27, no. 4, pp. 199 – 209, 1999.

[9] Z. Linan, Y. Peng, S. Hao, and C. Lingling, "Sound source target localization system of mobile robot," in *Control Automation Robotics Vision (ICARCV), 2010 11th International Conference on*, Dec 2010, pp. 2289–2294.

[10] H. Sun, P. Yang, L. Zu, and Q. Xu, "An auditory system of robot for sound source localization based on microphone array," in *Robotics and Biomimetics (ROBIO), 2010 IEEE International Conference on*, Dec 2010, pp. 629–632.

[11] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. G. Okuno, "Multiple moving speaker tracking by microphone array on mobile robot," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[12] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robot. Auton. Syst.*, vol. 58, no. 11, pp. 1185–1196, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2010.08.001

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

[14] [Online]. Available: http://www.sagiisc.in/demos