Speaker Localization and Tracking in Mobile Robot Environment Using a Microphone Array ⋆

Ivan Marković **Ivan Petrović** *

* *Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia (e-mail: ivan.markovic85@gmail.com, ivan.petrovic@ fer.hr)*

Abstract: In this paper a method for speaker localization and tracking is proposed based on Time Difference of Arrival estimation enhanced with so called tuned phase transform. The localization method is based on Pseudo-linear estimator, and Y-shaped array for spatial sampling is proposed and compared to square array. The tracking is realized with Recursive Least-Squares algorithm. At the end, results recorded on a mobile robot are presented, showing that the developed audio interface and algorithm can localize and track speaker in mobile robot environment.

1. INTRODUCTION

In everyday life humans rely greatly on hearing as a complementary tool for understanding the world around us. Hearing, as one of the traditional five senses, elegantly supplements other senses as being omnidirectional, not limited by physical obstacles and the absence of light.

The problem of endowing systems with hearing depends, of course, on the field of utility. This paper's field of utility is mobile robotics and as it will be shown, several facts make auditory systems for mobile robots more challenging then, for e.g., conference rooms.

Existing source localization strategies are usually divided in three categories: those based upon maximizing the steered response power of a beamformer (a lot of great work with this method in mobile robotics was done by Valin et al. [2006]), techniques adopting high-resolution spectral estimation concepts, and approaches employing Time Difference Of Arrival (TDOA) information. An overview of many of these methods can be found in Brandstein et al. [2001], and Chen et al. [2006].

This paper proposes a new sound source localization method based on TDOA estimation using a microphone array of 4 microphones arranged in a specific Y-geometry. The proposed algorithm is based on Pseudo-linear estimation (PLE) of the sound source location, which enables us to solve the front-back ambiguity, increase the robustness by using all the available measurements, and to localize and track speaker over the full range around the mobile robot.

The main purpose of this algorithm is to provide the speaker location to other mobile robot systems. For an example, when tracking humans with laser sensors it could be useful to know which detected person is currently speaking, or we could be getting the robot's attention simply by means of voice, just as we would call a waiter in restaurant. Advantages of knowing from where a recorded sound comes from are numerous.

⋆ This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 036-0363078-3018.

2. TDOA ESTIMATION

The main idea behind TDOA-based locators is a two step one. Firstly, TDOA estimation of the speech signals relative to pairs of spatially separated microphones is performed. Secondly, this data is used to infer about speaker location. The TDOA estimation algorithm for 2 microphones is described first.

A windowed frame of L samples is considered. In order to determine the delay $\Delta\tau_{ij}$ in the signal captured by two different microphones (i and j), it is necessary to define a coherence measure which will yield an explicit global peak at the correct delay. Cross-correlation is the most common choice, since we have at two spatially separated microphones (in an ideal homogeneous, dispersion-free and lossless scenario) two identical time-shifted signals. Cross-correlation is defined by the following expression:

$$R_{ij}(\Delta\tau) = \sum_{n=0}^{L-1} x_{m_i}[n] x_{m_j}[n - \Delta\tau], \tag{1}$$

where x_{m_i} is the signal received by microphone i and $\Delta\tau$ is the correlation lag in samples. As stated earlier, R_{ij} is maximal when $\Delta\tau$ is equal to the delay between the two received signals.

The most appealing property of cross-correlation is the ability to perform calculation in frequency domain, thus significantly lowering the computational intensity of the algorithm. Since we are dealing with signal frames, we can only estimate the cross-correlation:

$$\hat{R}_{ij}(\Delta\tau) = \sum_{k=0}^{L-1} X_{m_i}(k) X^*_{m_j}(k) e^{j2\pi \frac{k\Delta\tau}{L}}, \tag{2}$$

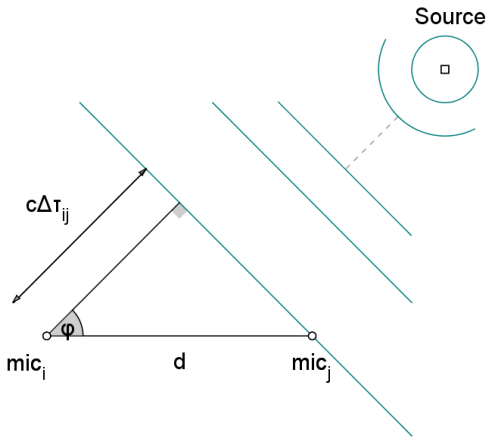

Fig. 1. Computing DOA angle from TDOA

where $X_{m_i}(k)$ is the discrete Fourier Transform (DFT) of $x_{m_i}[n]$ and $(.)^*$ denotes complex-conjugate. We are windowing the frames with rectangular window and no overlap. Therefore, before applying Fourier transform to signals x_{m_i} and x_{m_j}, it is necessary to zero-pad them with at least $2L$ zeros since we want to perform linear, not circular convolution.

A major limitation of cross-correlation given by (2) is that the correlation between adjacent samples is high, which has an effect of wide cross-correlation peaks. Therefore, an appropriate weighting should be used.

2.1 Spectral weighting

The problem of wide peaks in unweighted, i.e. generalized, cross-correlation (GCC) can be solved by whitening the spectrum of signals prior to computing the cross-correlation. The most common weightning function is Phase Transform (PHAT) which, if having constant SNR ration on all frequency bins, yields Maximum Likelihood Estimator (MLE). What PHAT function ($\psi_{\text{PHAT}} = 1/|X_{m_i}(k)||X^*_{m_j}(k)|$) does, is that it whitens the cross-spectrum of signals x_{m_i} and x_{m_j}, thus giving a sharpened peak at the true delay. The main drawback of the generalized cross-correlation with PHAT weighting is that it equally weights all frequency bins regardless of the SNR. Using just the PHAT weighting poor results were obtained and we concluded that the effect of the PHAT function should be tuned down, which lead us to GCC-PHAT-β:

$$\hat{R}_{ij}^{\text{PHAT}-\beta}(\Delta\tau) = \sum_{k=0}^{L-1} \frac{X_{m_i}(k)X^*_{m_j}(k)}{(|X_{m_i}(k)||X^*_{m_j}(k)|)^\beta} e^{j2\pi\frac{k\Delta\tau}{L}}. \quad (3)$$

where $0 < \beta < 1$ is the tuning parameter. As it was explained and shown by Donohue et al. [2007], the main reason for this approach is that speech can exhibit both wide-band and narrow-band characteristics. For example, if uttering the word "shoe", "sh" component acts as a wide-band signal and voiced component "oe" as a narrow-band signal. It was proposed in the latter article that a value of β between 0.5 and 0.6 should be taken. We chose to work with $\beta = 0.5$.

2.2 Direction of Arrival Estimation

The TDOA between microphones i and j: $\Delta\tau_{ij}$ can be found by locating the peak in the cross-correlation:

$$\Delta\tau_{ij} = \arg\max_{\tau} \hat{R}_{ij}^{\text{PHAT}}(\Delta\tau). \quad (4)$$

Once TDOA estimation is performed, it is possible to compute the position of the sound source through series of geometrical calculations. It is assumed that the distance to the source is much larger than the array aperture, i.e. we assume the so called far-field scenario. Although this might not always be the case, being that human-robot interaction is actually a mixture of far-field and near-field scenarios, this mathematical simplification is still a reasonable one. Fig. 1 illustrates the case of a 2 microphone array with a source in the far-field. Using the cosine law we can state the following:

$$\varphi = \pm \arccos\left(\frac{c\Delta\tau_{ij}}{d}\right), \quad (5)$$

where d is the distance between the microphones, $c = 344$ m/s is the speed of sound, and φ is the Direction of Arrival (DOA) angle.

Next, we conducted experiments with two microphones separated at distance $d = 0.5$ m and sampling frequency $F_s = 48$ kHz to test the effects of PHAT-β. The experiments were conducted over the range of 180° from the baseline of two microphones, uttering the word "Test". Beneficial effects of PHAT-β can be clearly seen from Fig. 2. In our experiments we did not experience absolute azimuth errors larger than 7°, except when getting close to $\varphi = 0°, 180°$. The best estimation results occur close to where $\varphi = 90°$. Similar phenomenon was reported also by Nakadai et al. [2002]. The reason for this kind of behavior lies in the non-linearity of the Acos function; when getting close to 0° and 180° small changes in TDOA error result in high azimuth error changes. This problem can be alleviated by using more than 2 microphones in a specific array geometry as will be shown in section 4.

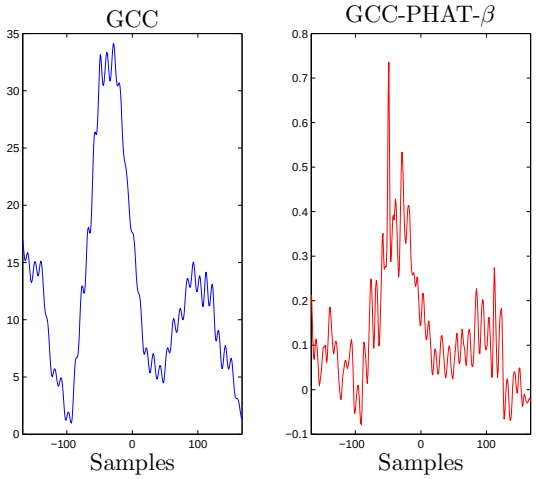

Fig. 2. Cross-correlation with no weighting (left figure) and PHAT-β weighting (right figure)

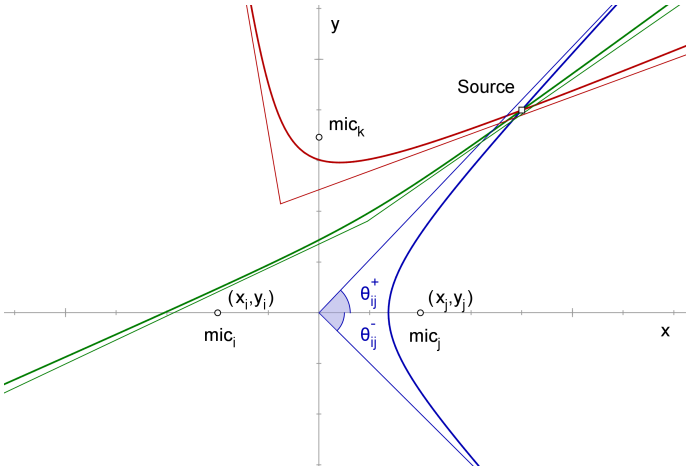

Fig. 3. Source location as an intersection of hyperbolic curves

3. HYPERBOLIC POSITION ESTIMATION

Performing sound source localization in mobile robot environment is specific in a few ways: the dimension of the array must be reasonably small, just as the number and dimension of the microphones, plus the algorithm must yield a closed-form unique solution for the sake of real-time application. No ± ambiguities on any of the 2-D axis is allowed. There are several methods that provide great results when locating a speaker in a conference room where microphones are mounted on a wall (only the + solution on the y axis is considered), but these methods are not practical for a mobile robot since the speaker can be located anywhere around the robot from 0° to 360°. Also, the fact that acoustical surrounding is not constant and that the speaker is always located outside of the microphone array (see Fig.6 and Fig. 7) makes sound source localization in mobile robot environment more challenging than localization in a single closed space.

Instead of searching for hyperbolae intersection as a location of the speaker, we turned to hyperbolae approximation with its asymptotes and searched for their intersection, giving us smaller variance over the bearing estimation and instead of four TDOA estimates, only three are needed for closed form solution. If we define R_{m_i} and R_{m_j} as the distances of the sound source from microphones i and j, it can easily be shown that a microphone pair (i, j) defines possible speaker locations in a form of hyperbolic curve:

$$\sqrt{(x-x_i)^2+(y-y_i)^2}-\sqrt{\left(x-x_j\right)^2+\left(y-y_j\right)^2}= \\ =R_{m_i}-R_{m_j}=R_{m_{ij}}=c\Delta\tau_{ij} \tag{6}$$

where (x_i, y_i) represent microphone coordinates and (x, y) sound source coordinates. Having more than one microphone pair enables us to calculate the speaker position as the intersection of the hyperbolic curves (Fig. 3).

As it was proposed by Drake et al. [2004], the bearing angles $\theta_{ij}^{\pm}$ of the hyperbolic asymptotes with respect to the baseline of a pair of microphones i and j, located at (x_i, y_i) and (x_j, y_j), are calculated as follows:

$$\theta_{ij}^{\pm}=\operatorname{atan2}\left(\frac{y_j-y_i}{x_j-x_i}\right)\pm\arccos\left(\frac{c\Delta\tau_{ij}}{\|(x_j,y_j)-(x_i,y_i)\|}\right). \tag{7}$$

3.1 Sound Source Localization Using Hyperbolic Asymptotes

What we get with (7) is bearing of the sound source, which is actually the asymptote of the corresponding microphone pair hyperbola. This is where the far-field assumption comes in handy again, the larger the distance of the sound source from the microphone array, hyperbolas' eccentricity becomes smaller giving better approximation with its own asymptotes (see Fig. 3).

Having N microphones, (7) will yield 2 $\binom{N}{2}$ hyperbolic asymptotes, each representing a possible bearing line of the sound source. Which bearing lines will be utilized and how exactly, will be explained in detail in section 4.1. First we need to determine the source location from the available bearing lines which emanate from the midpoints of microphone pairs defined by:

$$\mathbf{m}_{ij}=\frac{1}{2}\begin{bmatrix}x_i+x_j\\ y_i+y_j\end{bmatrix}_{2\text{x}1}. \tag{8}$$

To triangulate the feasible bearing a Pseudolinear Estimator (PLE) is used (Drake et al. [2004]) based on Least Squares (LS) to estimate the source location. The sound source PLE is given by the following relation:

$$\hat{r}_{\text{PLE}}=(\mathbf{A}^{\text{T}}\mathbf{A})^{-1}\mathbf{A}^{\text{T}}\mathbf{b}, \tag{9}$$

where

$$\mathbf{A}=\begin{bmatrix}\sin\varphi_{12} & -\cos\varphi_{12}\\ \vdots & \vdots\\ \sin\varphi_{ij} & -\cos\varphi_{ij}\end{bmatrix}_{\binom{N}{2}\text{x}2}, \\ \mathbf{b}=\begin{bmatrix}\begin{bmatrix}\sin\varphi_{12} & -\cos\varphi_{12}\end{bmatrix}_{1\text{x}2}\mathbf{m}_{12}\\ \vdots\\ \begin{bmatrix}\sin\varphi_{ij} & -\cos\varphi_{ij}\end{bmatrix}_{1\text{x}2}\mathbf{m}_{ij}\end{bmatrix}_{\binom{N}{2}\text{x}1}. \tag{10}$$

Here $\{\mathbf{m}_{ij}, \varphi_{ij}\}$ is the list of all feasible bearing lines of microphones i and j, with $\varphi_{ij} \in \{\theta_{ij}^{-}, \theta_{ij}^{+}\}$.

4. Y-ARRAY VS. SQUARE ARRAY

We chose to work with the Y-shaped array instead of the square shaped array. The first reason for this lies in the fact that the Y-shaped array positions the microphones in such a way that no two microphone-pair baselines are parallel. Having baselines with the maximum variety of different orientations maximizes the probability that the impinging source wave will be coming from $\varphi = 90°$ angles. This can be best seen from Fig. 4 and Fig. 5, on which a 15° beam is emanating from the midpoint of microphone pairs. We can see that 12 regions vs. 8 regions, in favor of the Y-shaped array, are covered around the array. This is so because in the case of the square array, two couples

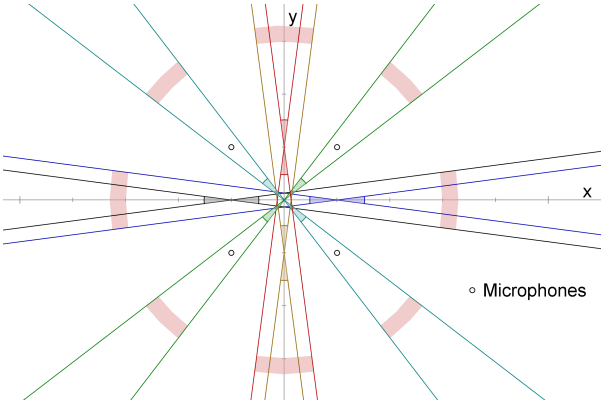

Fig. 4. Square Array Microphone Baseline Orientation Variety

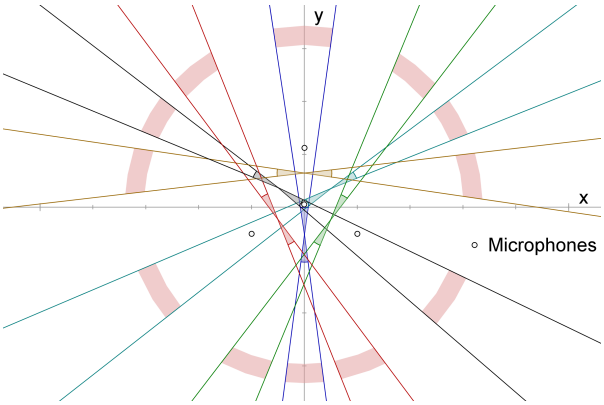

Fig. 5. Y-array Microphone Baseline Orientation Variety

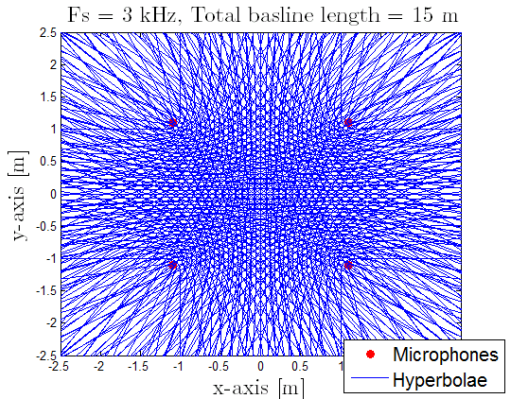

Fig. 6. Square Array Resolution Map

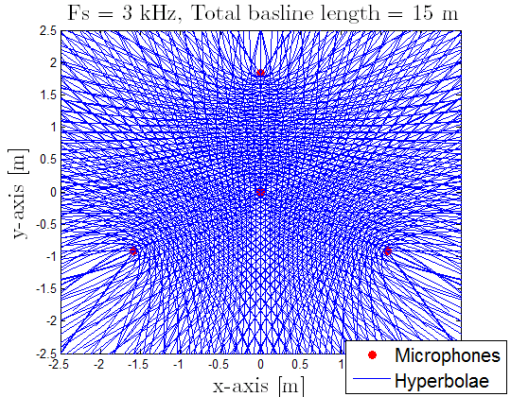

Fig. 7. Y-Array Resolution Map

of parallel microphone baselines are used for the same direction (positive and negative headings of the x and y axis.).

The second reason is that the Y-shaped array (Fig. 7) has a slightly superior resolution map than the square array (Fig. 6) due to the greater incidence of hyperbolae intersections.

4.1 Source Bearing Angle Decision Making

As stated in section 3.1, each bearing angle φ_{ij} has two possible values. If we look at the Fig. 8 we can see that there are 12 different cells with different bearing angle combinations. The microphones were paired in such way that the positive bearing angles are emanating outside of the array (how exactly microphones were paired can be seen from Fig. 8). For an example, if the source is located somewhere in the cell area 1, then we would use $\{\theta^+_{12}, \theta^+_{13}, \theta^+_{14}, \theta^+_{34}, \theta^-_{23}, \theta^-_{24}\}$ in (9) to determine the location of the speaker.

The decision procedure is as follows: we need to calculate 12 instances of (9) for all 12 different cells, then the one where the speaker is located will have the largest range.

4.2 TDOA estimation using N microphones

Using an array of N microphones makes it possible to compute $\binom{N}{2}$ different cross-correlations, of which only $N-1$ are independent. Since we presume far-field case and constant speed planar wave front propagation, for TDOA values following relation holds:

$$\Delta\tau_{ij} = \Delta\tau_{kj} - \Delta\tau_{ki}. \tag{11}$$

For an example, when dealing with 4 microphones, there are 3 independent TDOA measurements, other 3 TDOA estimates can be derived from (11). In our algorithm only those blocks which satisfy 12 constraints in the form of $\Delta\tau_{ij} = \Delta\tau_{kj} - \Delta\tau_{ki} < \delta$ are considered valid and are processed further (δ is a parameter set empirically as an integer multiple of the sampling period $1/F_s$).

Having decided on the TDOA estimation algorithm, array geometry and hyperbolic localization procedure, next logical step was to test the system behavior through simulations. The simulations were performed through Monte Carlo runs. Microphones were placed at the vertices of equilateral triangle (side length $a = 0.6$ m), and the fourth microphone at the orthocenter. Speaker was located at $(x, y) = (1.5, 1.5)$ m, each TDOA measurement was corrupted with Gaussian noise of standard deviation $\sigma = 6 \cdot 1/F_s$, since that was the largest case of error we experienced during two microphone recordings.

Fig. 9 shows results of the simulation. Unfortunately, the measurement uncertainty of range is too great for range estimation to be of practical use, but results for azimuth proved to be encouraging. Histogram of azimuth estimation is shown in Fig. 10 with corresponding mean and variance.

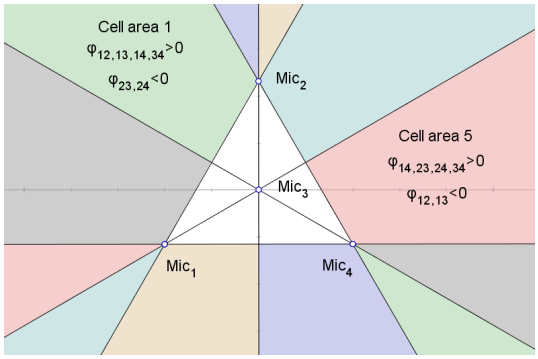

Fig. 8. Y-Array cell distribution with according to different bearing angle combinations

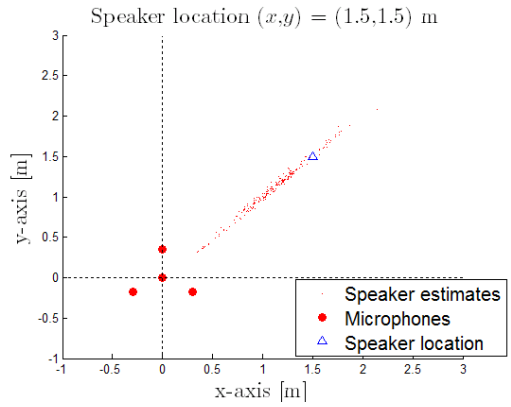

Fig. 9. Speaker location Monte Carlo runs

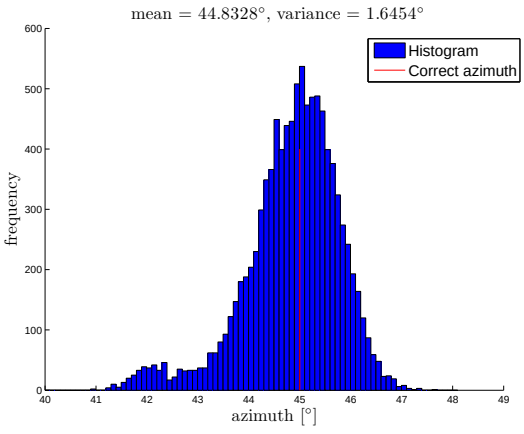

Fig. 10. Distribution of azimuth estimation values

5. TRACKING ALGORITHM

Recursive smoothing algorithm is used for speaker tracking. It has been shown and proposed in Doğançay et al. [2005] that its performance is almost identical to that of Kalman tracker, or even in some cases better. The target trajectory is estimated by utilizing the following kinematic equation, which represents a constant-acceleration motion model:

$$\mathbf{s}_k = \mathbf{s}_0 + \mathbf{v}_0 t_k + \frac{1}{2}\mathbf{a}t_k^2 = \mathbf{M}_k \boldsymbol{\xi}, \tag{12}$$

where $\mathbf{s}_0$ and $\mathbf{v}_0$ are the target location and velocity at t_0, respectively, and $\mathbf{a}$ is the constant target acceleration, $\mathbf{M}_k$ is the 2x6 time matrix:

$$\mathbf{M}_k = \begin{bmatrix} 1 & 0 & t_k & 0 & \frac{1}{2}t_k^2 & 0 \\ 0 & 1 & 0 & t_k & 0 & \frac{1}{2}t_k^2 \end{bmatrix}, \tag{13}$$

and $\boldsymbol{\xi}$ is the 6x1 target motion parameter vector:

$$\boldsymbol{\xi} = \begin{bmatrix} \mathbf{s}_0 \\ \mathbf{v}_0 \\ \mathbf{a} \end{bmatrix}. \tag{14}$$

Given $K \geq 3$ location estimates $\hat{\mathbf{s}}_k$, the unknown vector $\boldsymbol{\xi}$ can be estimated from:

$$\begin{bmatrix} \mathbf{M}_0 \\ \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_{K-1} \end{bmatrix} \boldsymbol{\xi} \approx \begin{bmatrix} \hat{\mathbf{s}}_0 \\ \hat{\mathbf{s}}_1 \\ \\ \hat{\mathbf{s}}_{K-1} \end{bmatrix}. \tag{15}$$

To track a speaker, (15) can be solved by using the Recursive Least Squares (RLS) algorithm:

$$\hat{\boldsymbol{\xi}}_k = \boldsymbol{\Phi}_k^{-1} \boldsymbol{\phi}_k, \tag{16}$$

where

$$\boldsymbol{\Phi}_k = \lambda \boldsymbol{\Phi}_{k-1} + \mathbf{M}_k^{\mathrm{T}} \mathbf{M}_k, \; k = 0, 1, \ldots \tag{17}$$

$$\boldsymbol{\phi}_k = \lambda \boldsymbol{\phi}_{k-1} + \mathbf{M}_k^{\mathrm{T}} \hat{\mathbf{s}}_k, \; k = 0, 1, \ldots \tag{18}$$

and $0 < \lambda < 1$ is the exponential forgetting factor. Note that the inverse matrix $\boldsymbol{\Phi}_k^{-1}$ can be calculated in advance since it is deterministic and independent of the location estimates $\hat{s}_k$. The smoothed location estimates are given by:

$$\hat{\mathbf{s}}_k^{\mathrm{RLS}} = \mathbf{M}_k \hat{\boldsymbol{\xi}}_k. \tag{19}$$

By making λ small, the tracking refresh rate can be improved at the expense of increased estimation variance. The final bearing angle is calculated from location coordinates given by (19).

6. EXPERIMENTAL RESULTS

The array used for experiments is composed of 4 omnidirectional microphones arranged in the Y geometry. Three microphones are placed on the vertices of equilateral triangle having side length $a = 0.6$ m, and the fourth microphone is placed at the orthocenter of the triangle. The microphone array is placed on a Pioneer 2DX robot as shown on Fig. 13. Audio interface is composed of low-cost microphones, pre-amplifiers and external USB sound-card (whole equipment costing $\sim$ 150 euro). Sampling frequency was $F_s = 48$ kHz, 16-bit precision, block length $L = 1024$ samples, and rectangular window was used with zero-padding approach.

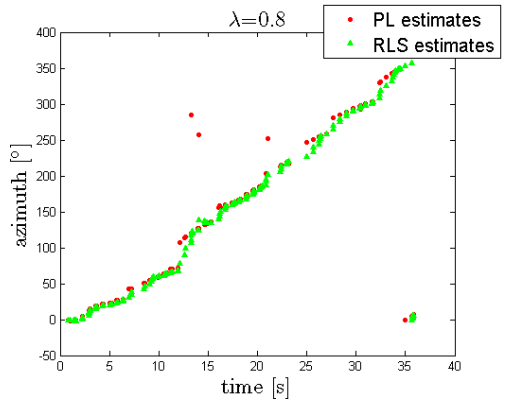

Fig. 11. Experiment results of azimuth estimation (speaker making a full circle)

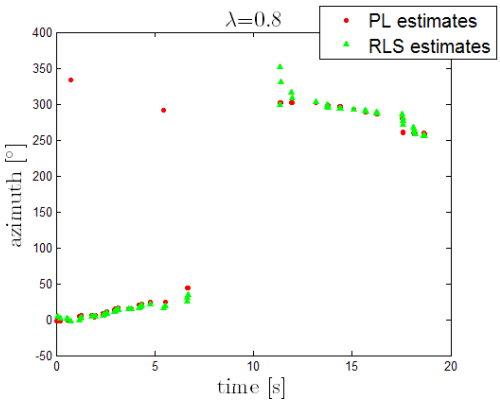

Fig. 12. Experiment results of azimuth estimation (speaker making a rapid angle change)

Two tests were performed. In first experiment speaker walked around the robot making a full circle, uttering "Test, one, two, three" continuously. The results of the experiment are shown in Fig. 11, from which can be seen that the algorithm successfully localizes and tracks the speaker. Raw data has few outliers, but the tracking resolved with RLS solves that problem.

In the second experiment, speaker uttered "Test, one, two, three" while moving from 0°-45°, changed the angle while keeping quiet, then continued repeating the sentence while moving from 315°-270° approximately. The situation in the second experiment is more likely to occur in real-life settings, since it is a reasonable assumption that speakers will have pauses while walking around the robot. As it can be seen from Fig. 12, in this case the algorithm also manages to track the speaker and eliminate outliers successfuly. The RLS estimates have a mild overshoot at the beginning of a new angle value, due to the rapid change.

7. CONCLUSION

We have implemented an audio interface for a mobile robot that accurately localizes and tracks speaker in 2-D over the full range around the robot.

The TDOA estimation used is shown to be robust to both reverberation and noise. Furthermore, PL estimation algorithm along with the RLS tracking approach proved to be an efficient tool in precise localization. However, our system is not yet capable of tracking multiple speakers and estimating the speaker range. Still, we find this to be the first step towards developing a functional audio interface, with final step being the full integration with other mobile robot systems.

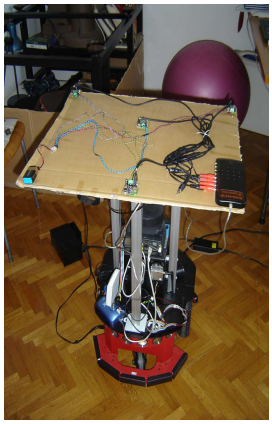

Fig. 13. Audio interface mounted on the robot

REFERENCES

M. Brandstein, and D. Ward. *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

Jingdong Chen, Jacob Benesty, and Yiteng (Arden) Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on Applied Signal Processing*, pages 1–19. 2006.

Kevin D. Donohue, Jens Hannemann, and Henry G. Dietz. Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environmnents. *Signal Processing*, volume 87, pages 1677–1691. 2007.

Kutluyil Doğançay, and Ahmad Hashemi-Sakhtsari. Target tracking by time difference of arrival using recursive smoothing. *Signal Processing*, volume 85, pages 667–679. 2005.

Samuel P. Drake, and Kutluyil Doğançay. Geolocation by time difference of arrival using hyperbolic asymptotes. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 361–364. 2004.

P.J. Moore, I.A. Glover, and C.H. Peck. An impulsive noise source position locator. *University of Bath*, 2002.

Kazuhiro Nakadai, Hiroshi G. Okuno, and Hirokai Kitano. Real-time sound source localization and separation for robot audition. *Proceedings of 2002 International Conference on Spoken Language Processing*, pages 193–196. 2002.

Jean-Marc Valin, François Michaud, and Jean Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 2006.