Speaker Localization and Tracking in Mobile Robot Environment Using von Mises Distribution Probabilistic Sensor Modelling and Particle Filtering

Ivan Marković[a,*], Ivan Petrović[a]

[a] *University of Zagreb Faculty of Electrical Engineering and Computing, Department of Control and Computer Engineering, Zagreb, Croatia*

Abstract

This paper deals with the problem of localizing and tracking a moving speaker over the full range around the mobile robot. The problem is solved by taking advantage of the phase shift between signals received at spatially separated microphones. The proposed algorithm is based on estimating the time difference of arrival by maximizing the weighted cross-correlation function in order to determine the azimuth angle of the detected speaker. The cross-correlation is enhanced with an adaptive signal-to-noise estimation algorithm to make the azimuth estimation more robust in noisy surroundings. A post processing technique is proposed in which each of these microphone-pair determined azimuths are further combined into a mixture of the von Mises distributions, thus producing a practical probabilistic representation of the microphone array measurement. It is shown that this distribution is inherently multimodal and that the system at hand is non-linear. Therefore, particle filtering is applied for discrete representation of the distribution function. Furthermore, two most common microphone array geometries are analysed and exhaustive experiments were conducted in order to qualitatively and quantitatively test the algorithm and compare the two geometries. Also, a voice activity detection algorithm based on the before mention signal-to-noise estimator was implemented and incorporated into the existing speaker localization system. The results show that the algorithm can reliably and accurately localize and track a moving speaker.

Keywords: Speaker localization, Microphone array, von Mises distribution, Particle filtering.

1. Introduction

In biological lifeforms hearing, as one of the traditional five senses, elegantly supplement other senses as being omnidirectional, not limited by physical obstacles, and absence of light. Inspired by these unique properties, researchers strive towards endowing mobile robots with auditory systems to further enhance human–robot interaction, not only by means of communication but also, just as humans do, to make intelligent analysis of the surrounding environment. By providing speaker location to other mobile robot systems, like path planning, speech and speaker recognition, such system would be a step forward in developing a fully functional human–aware mobile robots.

The auditory system must provide robust and non-ambiguous estimate of the speaker location, and must be updated frequently in order to be useful in practical tracking applications. Furthermore, the estimator must be computationally non-demanding and possess a short processing latency to make it practical for real-time systems. The afore mentioned requirements and the fact of an auditory system being placed on a mobile platform, thus changing the acoustical conditions on operating basis, make speaker localization and tracking a formidable problem.

Existing speaker localization strategies can be categorized in four general groups. The first group of algorithms refers to beamforming methods in which the array is steered to various locations of interest and searches for the peak in the output power [1–3]. The second group includes beamforming methods based upon analysis of spatiospectral correlation matrix derived from the signals received at the microphones [4]. The third group relies on computational simulations of the physiologically known parts of the hearing system, e.g. binaural cue processing [5–7]. The fourth group of localization strategies is based on estimating the Time Difference of Arrival (TDOA) of the speech signals relative to pairs of spatially separated microphones and then using that information to infer about the speaker location. Estimation of the TDOA and speaker localization from TDOA are two separate problems. The former is usually calculated by maximizing the weighted cross-correlation function [8], while the latter is commonly known as multilateration, i.e. hyperbolic positioning, which is a problem of calculating the source location by finding the intersection of at least two hyperbolae [9–12]. In mobile robotics, due to small microphone array dimensions, generally hyperbolae intersection is not calculated, only the angle (azimuth and/or elevation) is estimated [13–16].

Even though the TDOA estimation based methods are outperformed to a certain degree by several more elaborate methods [17, 18], they still prove to be extremely effective due to

*Corresponding author at: University of Zagreb Faculty of Electrical Engineeing and Computing, Department of Control and Computer Engineering, Zagreb, Croatia, Phone: +385 1 6129 561, Fax: +385 1 6129 809

Email addresses: ivan.markovic@fer.hr (Ivan Marković), ivan.petrovic@fer.hr (Ivan Petrović)

URL: http://act.rasip.fer.hr/people-opis.php?id=199 (Ivan Marković), http://act.rasip.fer.hr/people-opis.php?id=1 (Ivan Petrović)

their elegance and low computational costs. This paper proposes a new speaker localization method based on TDOA estimation using an array of 4 microphones. The proposed algorithm uses particle filtering and von Mises distribution for probabilistic modelling of the microphone pair measurements, which solves the front-back ambiguity, increases the robustness by using all the available measurements, and localizes and tracks speaker over the full range around the mobile robot. The main contribution of this paper is the proposed measurement model to be used for *a posteriori* inference about the speaker location.

The rest of the paper is organized as follows. Section 2 describes the implemented azimuth estimation method and the voice activity detector. Section 3 analyses Y and square microphone array geometries, while Section 4 defines the framework for the particle filtering algorithm, introduces the von Mises distribution, the proposed measurement model, and describes in detail the implemented algorithm. Section 5 presents the conducted experiments. In the end, Section 6 concludes the paper and presents future works.

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.

2.1. Principle of TDOA

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:

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

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

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

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

where $X_i(k)$ and $X_j(k)$ are the discrete Fourier Transforms (DFTs) of $x_i[n]$ and $x_j[n]$, and $(.)^*$ denotes complex-conjugate. We are windowing the frames with rectangular window and no overlap. Therefore, before applying Fourier transform to signals x_i and x_j, it is necessary to zero-pad them with at least L zeros, since we want to calculate linear, and not circular convolution.

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

2.2. 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 weighting function is the Phase Transform (PHAT) which, as it has been shown in [8], under certain assumptions yields Maximum Likelihood (ML) estimator. What PHAT function ($\psi_{\text{PHAT}} = 1/|X_i(k)||X_j^*(k)|$) does, is that it whitens the cross-spectrum of signals x_i and x_j, thus giving a sharpened peak at the true delay. In the frequency domain, GCC-PHAT is computed as:

$$\hat{R}_{ij}^{\text{PHAT}}(\Delta\tau) = \sum_{k=0}^{L-1} \frac{X_i(k)\, X_j^*(k)}{|X_i(k)||X_j(k)|} e^{j2\pi\frac{k\Delta\tau}{L}}. \tag{3}$$

The main drawback of the GCC with PHAT weighting is that it equally weights all frequency bins regardless of the signal-to-noise ratio (SNR), thus making the system less robust to noise. To overcome this issue, as proposed in [1], a modified weighting function based on SNR is incorporated into GCC framework.

Firstly, a gain function for such modification is introduced (this is simply a Wiener gain):

$$G_i^n(k) = \frac{\xi_i^n(k)}{1+\xi_i^n(k)}, \tag{4}$$

where $\xi_i^n(k)$ is the *a priori* SNR at the i th microphone, at time frame n, for frequency bin k and $\xi_i^0 = \xi_{min}$. The *a priori* SNR is defined as $\xi_i^n(k) = \lambda_{i,x}^n(k)/\lambda_i^n(k)$, where $\lambda_{i,x}^n(k)$ and $\lambda_i^n(k)$ are the speech and noise variance, respectively. It is calculated by using the *decision-directed estimation* approach proposed in [19]:

$$\xi_i^n(k) = \alpha_e[G_i^{n-1}(k)]^2\gamma_i^{n-1}(k) + (1-\alpha_e)\max\{\gamma_i^n(k)-1, 0\}, \tag{5}$$

where α_e is the adaptation rate, $\gamma_i^n = |X_i^n(k)|^2/\lambda_i^n(k)$ is the *a posteriori* SNR, and $\lambda_i^0(k) = |X_i^0(k)|^2$.

In stationary noise environments, the noise variance of each frequency bin is time invariant, i.e. $\lambda_i^n(k) = \lambda_i(k)$ for all n. But if the microphone array is placed on a mobile robot, most surely due to robot's changing location, we will have to deal with non-stationary noise environments. An algorithm used to estimate $\lambda_i^n(k)$ is based on *minima controlled recursive averaging* (MCRA) developed in [20, 21]. The noise spectrum

is estimated by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability. Speech absence in a given frame of a frequency bin is determined by the ratio between the local energy of the noisy signal and its minimum within a specified time window. The smaller the ratio in a given spectrum, more probable the absence of speech is. Further improvement can be made in (4) by using a different spectral gain function [22].

To make the TDOA estimation more robust to reverberation, it is possible to modify the noise estimate $\lambda_i^n(k)$ to include a reverberation term $\lambda_{i,\text{rev}}^n(k)$:

$$\lambda_i^n(k) \mapsto \lambda_i^n(k) + \lambda_{i,\text{rev}}^n(k), \tag{6}$$

where $\lambda_{i,\text{rev}}^n$ is defined using reverberation model with exponential decay [1]:

$$\lambda_{i,\text{rev}}^n(k) = \alpha_{\text{rev}}\lambda_{i,\text{rev}}^{n-1}(k) + (1-\alpha_{\text{rev}})\delta|G_i^{n-1}(k)X_i^{n-1}(k)|^2, \tag{7}$$

where α_{rev} is the reverberation decay, δ is the level of reverberation and $\lambda_{i,\text{rev}}^0(k) = 0$. Equation (7) can be seen as modelling the *precedence effect* [23, 24], in order to give less weight to frequencies where recently a loud sound was present.

Using just the PHAT weighting poor results were obtained and we concluded that the effect of the PHAT function should be tuned down. As it was explained and shown in [25], 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.

Based on the discussion above, the enhanced GCC-PHAT-β has the following form:

$$\hat{R}_{ij}^{\text{PHAT-}\beta\text{e}}(\Delta\tau) = \sum_{k=0}^{L-1} \frac{G_i(k)X_i(k)G_j(k)X_j^*(k)}{\left(|X_i(k)||X_j(k)|\right)^\beta} e^{j2\pi\frac{k\Delta\tau}{L}}. \tag{8}$$

where $0 < \beta < 1$ is the tuning parameter.

2.3. Voice Activity Detector

At this point it would be practical to devise a way of discerning if the processed signal frame contains speech or not. This method would prevent misguided interpretations of the TDOA estimation due to speech absence, i.e. estimation from signal frames consisting of noise only. Implemented Voice Activity Detector (VAD) is a statistical model-based one, originating from methods proposed in [19, 20].

Basically, two hypotheses are considered; $H_0^n(k)$ and $H_1^n(k)$, indicating respectively, speech absence and presence in the frequency bin k of the frame n. Observing DFT $X_i(k)$ of the signal at microphone i, the DFT coefficients are modelled as complex Gaussian variables. Accordingly, the conditional probability density functions (pdfs) of $X_i(k)$ are given by:

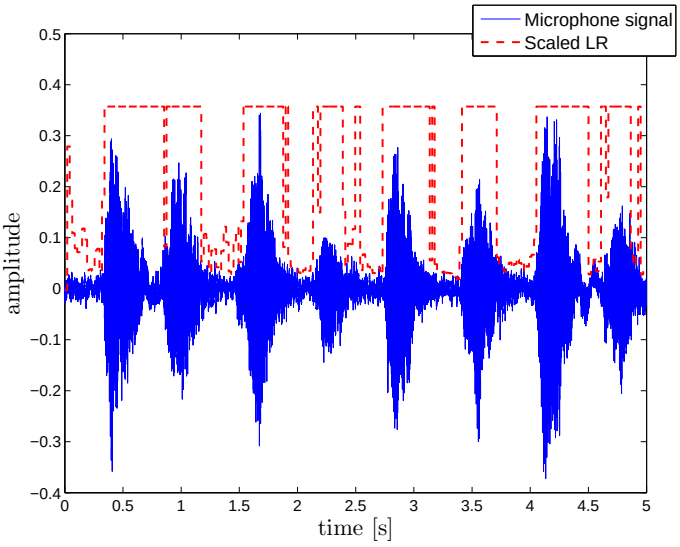


Figure 1: Recorded speech signal with corresponding scaled Likelihood Ratio

$$\begin{aligned} p(X_i^n(k)|H_0^n(k)) &= \frac{1}{\pi\lambda_i^n(k)}\exp\left(-\frac{|X_i^n(k)|^2}{\lambda_i^n(k)}\right) \\ p(X_i^n(k)|H_1^n(k)) &= \frac{1}{\pi(\lambda_i^n(k)+\lambda_{i,x}^n(k))}\times \\ &\quad \times\exp\left(-\frac{|X_i^n(k)|^2}{\lambda_i^n(k)+\lambda_{i,x}^n(k)}\right). \end{aligned} \tag{9}$$

Likelihood Ratio (LR) of the frequency bin k is given by:

$$\begin{aligned} \Lambda_i^n(k) &= \frac{p(X_i^n(k)|H_1^n(k))}{p(X_i^n(k)|H_0^n(k))} \\ &= \frac{1}{1+\xi_i^n(k)}\exp\left(\frac{\gamma_i^n(k)\xi_i^n(k)}{1+\xi_i^n(k)}\right). \end{aligned} \tag{10}$$

Figure 1 shows recorded speech and its scaled LR. It can be seen that the algorithm is successful in discriminating between speech and non-speech regions. The rise in LR value at the beginning of the recording is due to training of the SNR estimator. Finally, a binary-decision procedure is made based on the geometric mean of LRs:

$$\frac{1}{2L}\sum_{k=0}^{2L-1}\log\Lambda_i^n(k) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{11}$$

where a signal frame is classified as speech if the geometric mean of LRs exceed a certain threshold value η. This method can be further enhanced by calculating mean of optimally weighted LRs [26]. Also, instead of using a binary-decision procedure, VAD output can be a parameter based on SNR indicating the level of signal corruption, thus effectively informing a tracking algorithm to what extent measurements should be taken into account [27].

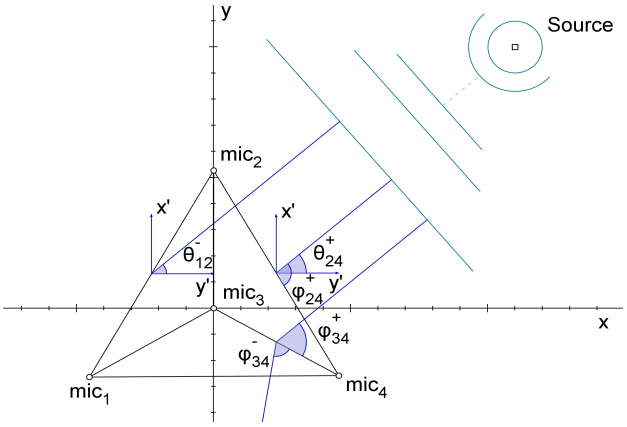


Figure 2: DOA angle transformation

2.4. Direction of Arrival Estimation

The TDOA between microphones i and j can be found by locating the peak in the cross-correlation:

$$\Delta\tau_{ij} = \arg\max_{\Delta\tau} \hat{R}_{ij}^{\mathrm{PHAT}-\beta e}(\Delta\tau). \tag{12}$$

Once TDOA estimation is performed, it is possible to compute the azimuth 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. Thus the expanding acoustical wavefront is modelled as a planar wavefront. 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. Using the cosine law we can state the following (Fig. 2):

$$\varphi_{ij} = \pm \arccos\left(\frac{c\Delta\tau_{ij}}{a_{ij}}\right), \tag{13}$$

where a_{ij} is the distance between the microphones, c is the speed of sound, and φ_{ij} is the Direction of Arrival (DOA) angle.

Since we will be using more than two microphones one must make the following transformation in order to fuse the estimated DOAs. Instead of measuring the angle φ_{ij} from the baseline of the microphones, transformation to azimuth θ_{ij} measured from the x axis of the array coordinate system (bearing line is parallel with the x axis when $\theta_{ij} = 0°$) is performed. The transformation is done with the following equation (angles φ_{24}^{+} and θ_{24}^{+} in Fig. 2):

$$\begin{aligned} \theta_{ij}^{\pm} &= \alpha_{ij} \pm \varphi_{ij} \\ &= \operatorname{atan2}\left(\frac{y_j - y_i}{x_j - x_i}\right) \pm \arccos\left(\frac{c\Delta\tau_{ij}}{a_{ij}}\right). \end{aligned} \tag{14}$$

At this point one should note the following:

- under the far-field assumption, all the DOA angles measured anywhere on the baseline of the microphones are equal, since the bearing line is perpendicular to the expanding planar wavefront (angles θ_{12}^{-} and θ_{24}^{+} in Fig. 2)
- front-back ambiguity is inherent when using only two microphones (angles φ_{34}^{-} and φ_{34}^{+} in Fig. 2).

Having M microphones, (14) will yield $2 \cdot \binom{M}{2}$ possible azimuth values. How to solve the front-back ambiguity and fuse the measurements is explained in Section 4.

3. Microphone Array Geometry

The authors find that microphone arrangement on a mobile robot is also an important issue and should be carefully analysed. If we constrain the microphone placement in 2D, then two most common configurations present:

- square array - four microphones are placed on the vertices of a square. The origin of the reference coordinate system is at the intersection of the diagonals
- Y array - three microphones are places on the vertices of an equilateral triangle, and the fourth is in the orthocenter which represents the origin of the reference coordinate system.

The dimensions of the microphone array depend on the type of the surface it is placed on. In this paper the two microphone array configurations will be compared as if they were placed on a circular surface with radius r (see Fig. 3). Hence, both arrays are defined by their respective square and triangle side length a, which is equal to $a = r\sqrt{2}$ and $a = r\sqrt{3}$, respectively.

Estimation of TDOA is influenced by the background noise, channel noise and reverberation, and the goal of (8) is to make the respective estimation as insensitive as possible to these influences. Under assumption that the microphone coordinates are measured accurately, we can see from (14) that the estimation of azimuth $\theta_{ij}^{\pm}$ depends solely on the estimation of TDOA. Therefore, it is reasonable to analyse the sensitivity of azimuth estimation to TDOA estimation error. As it will be shown, this sensitivity depends on the microphone array configuration.

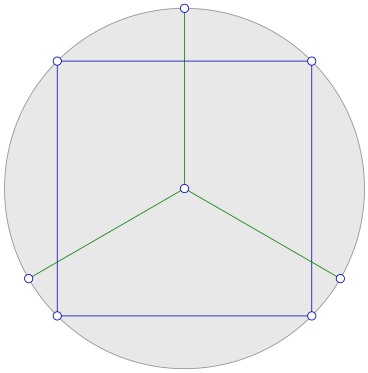


Figure 3: Possible array placement scenarios

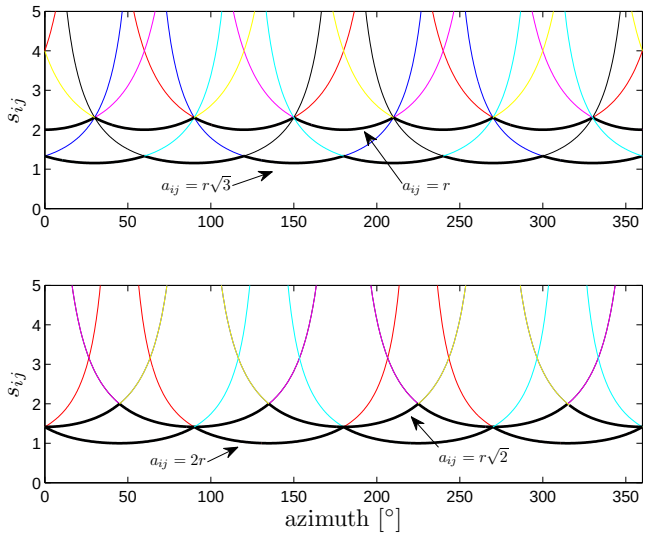


Figure 4: Error sensitivity of azimuth estimation for Y (upper plot) and square array (bottom plot)

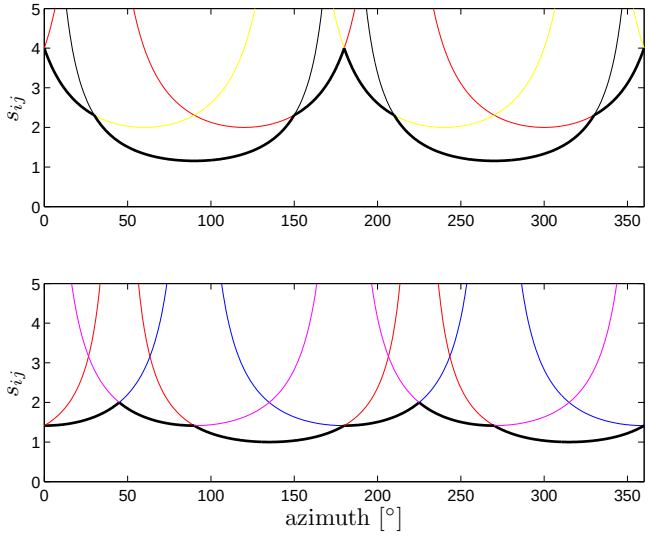


Figure 5: Error sensitivity of azimuth estimation for Y (upper plot) and square array (bottom plot) with one microphone occluded

Firstly, we define the error sensitivity of azimuth estimation to TDOA measurement, s_{ij}, as follows [28]:

$$s_{ij} = \frac{\partial \theta_{ij}}{\partial(\Delta\tau_{ij})}. \tag{15}$$

By substituting (13) and (14) into (15) and applying simple trigonometric transformations, we gain the following expression:

$$s_{ij} = \frac{c}{a_{ij}} \frac{1}{|\sin(\theta_{ij} - \alpha_{ij})|}. \tag{16}$$

From (16) we can see that there are two means by which error sensitivity can be decreased. The first is by increasing the distance between the microphones a_{ij}. This is kept under constraint of the robot dimensions and is considered to be fixed. The second is to keep the azimuth θ_{ij} as close to 90° relative to α_{ij} as possible. This way we are ensuring that the impinging source wave will be parallel to the microphones baseline. This condition could be satisfied if all the microphone pair baselines have the maximum variety of different orientations.

For the sake of the argument, let us set $c = 1$. Furthermore, both configurations will be analysed for equal dimensions (radius r will be equal for both configurations). The error sensitivity curves s_{ij}, as a function of azimuth θ_{ij}, for Y and square array are shown in Fig. 4.

We can see from Fig. 4 that the distance between the microphones a_{ij} mostly contributes to the offset of the sensitivity curves, and that the variety of orientations affects the effectiveness of angle coverage. For Y array, Fig. 4 shows two groups of sensitivity curves; one for $a_{ij} = r$ and other for $a_{ij} = r\sqrt{3}$. Former having the largest error sensitivity value of 2.3 approximately, and latter having the largest error sensitivity value of 1.3 approximately. For the square array, Fig. 4 shows also two groups of sensitivity curves; one for $a_{ij} = r\sqrt{2}$ and the other for $a_{ij} = 2r$. Former having the largest error sensitivity value of 2 approximately, and latter having the largest error sensitivity value of 1.4 approximately. From the above discussion we can see that the Y-array maximises baseline orientation variety, while the square array maximises total baseline length (this length is defined as sum of all the distances between the microphones and is in favour by factor 1.2 for square array). This type of analysis can also be easily made for bigger and more complex microphone array systems in order to search for the best possible microphone placements.

A possible scenario is that one of the microphones gets occluded and its measurement is unavailable or completely wrong. For Y array we have selected that one of the microphones on the vertices is occluded, since this is the most probable case, and for the square array it makes no difference, since the situation is only symmetrical for any microphone. Robustness of error sensitivity with respect to microphone occlusion is shown in Fig. 5 for both Y and square array, from which it can be seen that the result is far worse for Y array. This is logical, since we removed from the configuration two microphone pairs with largest baseline lengths. From the above discussion we can conclude that the square array is more robust to microphone occlusion.

Since we will be utilising all microphone pair measurements to estimate azimuth, it is practical to compare joint error sensitivity (JES) curves, which we define as:

$$\text{JES} = \sum_{\{i,j\}} s_{ij}, \ \forall \{i,j\} \ \text{ microphone pairs.} \tag{17}$$

Figure 6 shows both JES curves for Y and square array. We can see that there are two different peaks for both configurations. The peaks for Y array come from the fact that it has two different baseline lengths. The same applies for square ar-

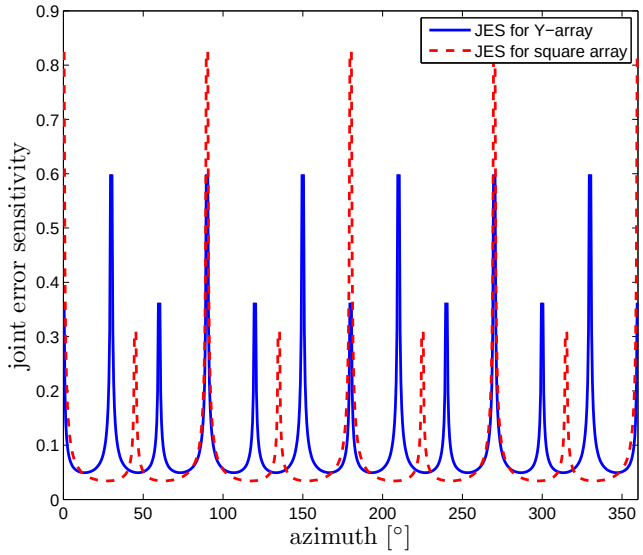


Figure 6: Joint error sensitivitiy curves for both Y and square microphone array configurations

ray, which additionally has the largest peak due to the fact that baselines of two couples of microphone pairs cover the same angle.

To conclude, we can state the following; although Y array configuration places microphones in such a way that no two microphone-pair baselines are parallel (thus ensuring maximum orientation variety), square array has larger total baseline length, yielding smaller overall error sensitivity and greater robustness to microphone occlusion.

Furthermore, when considering microphone placement on a mobile robot from a practical point of view, square array has one more advantage. If the microphones are placed on the body of the robot (as opposed to the top of the robot, e.g. the head), problem occurs for Y array configuration considering the placement of the fourth microphone (the one in the orthocenter). However, the advantages of Y-array should not be left out when considering tetrahedra microphone configurations (see [29], for e.g.). Also if the two configurations are analysed with both having the same total baseline length, Y array would prove to have superior angle resolution [13].

4. Speaker Localization and Tracking

The problem at hand is to analyse and make inference about a dynamic system. For that, two models are required: one describing the evolution of the speaker's state over time (system model), and second relating the noisy measurements to the speaker's state (measurement model). We assume that both models are available in probabilistic form. Thus, the approach to dynamic state estimation consists of constructing the *a posteriori* pdf of the state based on all available information, including the set of received measurements, which are further combined due to circular nature of the data, as a mixture of von Mises distributions.

4.1. Model of the sound source dynamics

The sound source dynamics is modelled by the well behaved Langevin motion model [30]:

$$\begin{aligned} \begin{bmatrix} \dot{x}_k \\ \dot{y}_k \end{bmatrix} &= \alpha \begin{bmatrix} \dot{x}_{k-1} \\ \dot{y}_{k-1} \end{bmatrix} + \beta \begin{bmatrix} \upsilon_x \\ \upsilon_y \end{bmatrix}, \\ \begin{bmatrix} x_k \\ y_k \end{bmatrix} &= \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \delta \begin{bmatrix} \dot{x}_k \\ \dot{y}_k \end{bmatrix}, \end{aligned} \tag{18}$$

where $[x_k, y_k]^{\mathbf{T}}$ is the location of the speaker, $[\dot{x}_k, \dot{y}_k]^{\mathbf{T}}$ is the velocity of the speaker at time index k, $\upsilon_x, \upsilon_y \sim \mathcal{N}(0, \sigma_\upsilon)$ is the stohastic velocity disturbance, α and β are model parameters, and δ is the time between update steps.

The system state, i.e. the speaker azimuth, is calculated via the following equation:

$$\theta_k = \operatorname{atan2}\left(\frac{y_k}{x_k}\right). \tag{19}$$

4.2. The von Mises distribution based measurement model

Measurement of the sound source state with M microphones can be described by the following equation:

$$\mathbf{z}_k = \mathbf{h}_k(\theta_k, n_k), \tag{20}$$

where $\mathbf{h}_k(.)$ is a non-linear function with noise term n_k, and $\mathbf{z}_k = [\theta^{\pm}_{ij}, \ldots, \theta^{\pm}_{M,M-1}]_k, i \neq j, \{i, j\} = \{j, i\}$ is the measurement vector defined as a set of azimuths calculated from (14). Working with M microphones gives $N = \binom{M}{2}$ microphone pairs and $2N$ azimuth measurements.

Since $\mathbf{z}_k$ is a random variable of circular nature, it is appropriate to model it with the von Mises distribution. The von Mises distribution with its pdf is defined as [31, 32]:

$$p(\theta_{ij}|\theta_k, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp[\kappa \cos(\theta_{ij} - \theta_k)], \tag{21}$$

where $0 \leq \theta_{ij} < 2\pi$ is the measured azimuth, $0 \leq \theta_k < 2\pi$ is the mean direction, $\kappa > 0$ is the concentration parameter and $I_0(\kappa)$ is the modified Bessel function of the order zero. Bessel function of the order m can be represented by the following infinite sum:

$$I_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k \, (x)^{2k+|m|}}{2^{2k+|m|} k! \, (|m| + k!)^2}, \quad |m| \neq \frac{1}{2}. \tag{22}$$

Mean direction θ_k is analogous to the mean of the normal Gaussian distribution, while concentration parameter is analogous to the inverse of the variance in the normal Gaussian distribution. Also, circular variance can be calculated and is defined as:

$$\vartheta^2 = 1 - \frac{I_1(\kappa)^2}{I_0(\kappa)^2}, \tag{23}$$

where $I_1(\kappa)$ is the modified Bessel function of order one.

According to (14), a microphone pair $\{i, j\}$ measures two possible azimuths θ^+_{ij} and θ^-_{ij}. Since we cannot discern from a single microphone pair which azimuth is correct, we can say,

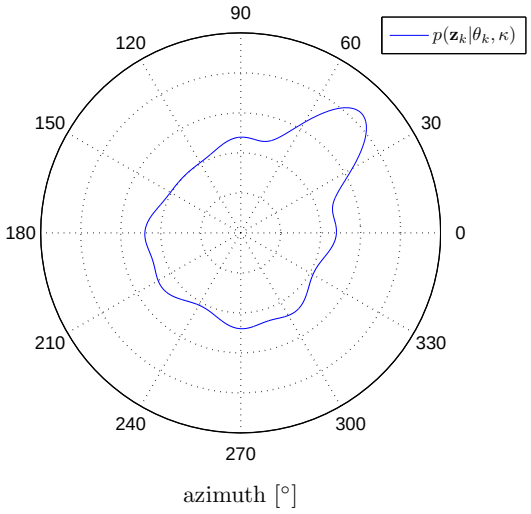


Figure 7: A mixture of several von Mises distributions wrapped on a unit circle (most of them having a mode at 45°)

from a probabilistic point of view, that both angles are equally probable. Therefore, we propose to model each microphone pair as a sum of two von Mises distributions, yielding a bimodal pdf of the following form:

$$\begin{aligned} p_{ij}\left(\theta^{\pm}_{ij,k}|\theta_k,\kappa\right) &= p_{ij}\left(\theta^{+}_{ij,k}|\theta_k,\kappa\right) + p_{ij}\left(\theta^{-}_{ij,k}|\theta_k,\kappa\right) \\ &= \tfrac{1}{2\pi I_0(\kappa)} \exp\left[\kappa\cos\left(\theta^{+}_{ij,k}-\theta_k\right)\right] + \\ &+ \tfrac{1}{2\pi I_0(\kappa)} \exp\left[\kappa\cos\left(\theta^{-}_{ij,k}-\theta_k\right)\right] \end{aligned} \tag{24}$$

Having all pairs modelled as a sum of two von Mises distributions, we propose a linear combination of all those pairs to represent the microphone array measurement model. Such a model has the following multimodal pdf:

$$p(\mathbf{z}_k|\theta_k,\kappa) = \frac{1}{2\pi I_0(\kappa)} \sum_{\{i,j\}=1}^{N} \beta_{ij} p_{ij}\left(\theta^{\pm}_{ij,k}|\theta_k,\kappa\right), \tag{25}$$

where $\sum \beta_{ij} = 1$ is the mixture coefficient. These mixture coefficients are selected so as to minimise the overall error sensitivity. As it has been shown, the error sensitivity is function of the azimuth. The goal of the coefficients β_{ij} is to give more weight in (25) to the most reliable pdfs. Therefore, we propose the following form of the coefficients:

$$\beta_{ij} = \frac{0.5 + \left|\sin(\theta_{k-1} - \alpha_{ij})\right|}{1.5}. \tag{26}$$

It is obvious that the mixture coefficients are function of the estimated azimuth and that this form can only be applied after the first iteration of the algorithm. Also, the coefficients are scaled so as to never cancel out completely a possibly unfavourable pdf. A careful reader will note that the origin of the form of (26) comes from the discussion in Section 3, resulting with (16).

The model (25) represents our belief in the sound source azimuth. A graphical representation of the analytical (25) is shown in Fig. 7. Of all the $2N$ measurements, half of them will measure the correct azimuth, while their counterparts from (14) will have different (not equal) values. So, by forming such a linear opinion pool, pdf (25) will have a strong mode at the correct azimuth value.

4.3. Particle filtering

From a Bayesian perspective, we need to calculate some degree of belief in the state θ_k, given the measurements $\mathbf{z}_k$. Thus, it is required to construct the pdf $p(\theta_k|\mathbf{z}_k)$ which bears multimodal nature due to TDOA based localization algorithm. Therefore, particle filtering algorithm is utilised, since it is suitable for non-linear systems and measurement equations, non-Gaussian noise, and multimodal distributions. This method represents the posterior density function $p(\theta_k|\mathbf{z}_k)$ by a set of random samples (particles) with associated weights and computes estimates based on these samples and weights. As the number of samples becomes very large, this characterisation becomes an equivalent representation to the usual function description of the posterior pdf, and the particle filter approaches the optimal Bayesian estimate.

Let $\{\theta^p_k, w^p_k\}^P_{p=1}$ denote a random measure that characterises the posterior pdf $p(\theta_k|\mathbf{z}_k)$, where $\{\theta^p_k,\ p = 1,\ldots,P\}$ is a set of particles with associated weights $\{w^p_k,\ p = 1,\ldots,P\}$. The weights are normalised so that $\sum_p w^p_k = 1$. Then, the posterior density at k can be approximated as [33]:

$$p(\theta_k|\mathbf{z}_k) \approx \sum_{p=1}^{P} w^p_k \delta(\theta_k - \theta^p_k), \tag{27}$$

where $\delta(.)$ is the Dirac delta measure. Thus, we have a discrete weighted approximation to the true posterior, $p(\theta_k|\mathbf{z}_k)$.

The weights are calculated using the principle of *importance resampling*, where the proposal distribution is given by (18). In accordance to the Sequential Importance Resampling (SIR) scheme, the weight update equation is given by [33]:

$$w^p_k \propto w^p_{k-1} p(\mathbf{z}_k|\theta^p_k), \tag{28}$$

where $p(\mathbf{z}_k|\theta^p_k)$ is calculated by (25), thus replacing θ_k with particles θ^p_k.

The next important step in the particle filtering is the *resampling*. The resampling step involves generating a new set of particles by resampling (with replacement) P times from an approximate discrete representation of $p(\theta_k|\mathbf{z}_k)$. After the resampling all the particles have equal weights, which are thus reset to $w^p_k = 1/P$. In the SIR scheme, resampling is applied at each time index. Since we have $w^p_{k-1} = 1/P\ \forall p$, the weights are simply calculated from:

$$w^p_k \propto p(\mathbf{z}_k|\theta^p_k). \tag{29}$$

The weights given by the proportionality (29) are, of course, normalised before the resampling step. It is also possible to perform particle filter size adaptation through the *KLD-sampling*

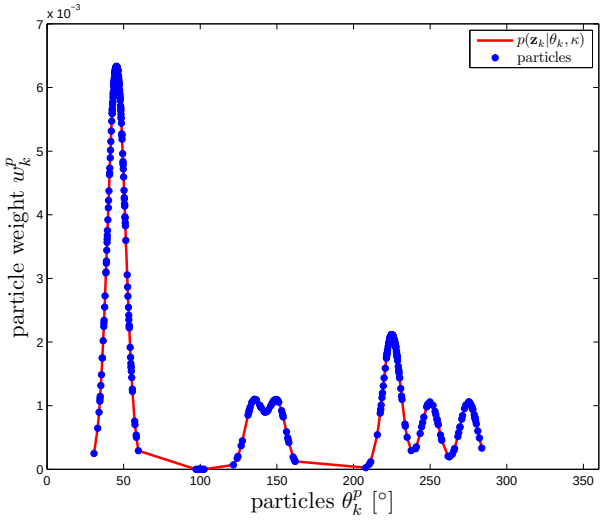


Figure 8: An unwraped discrete representation of the true $p(\mathbf{z}_k|\theta_k,\kappa)$

procedure proposed in [34]. This would take place before the resampling step in order to reduce the computational burden.

At each time index k and with M microphones, a set of $2N$ azimuths is calculated with (14), thus forming measurement vector $\mathbf{z}_k$ from which an approximation of (25) is constructed by pointwise evaluation (see Fig. 8) with particle weights w_k^p calculated from (29) and (25).

The θ_k is estimated, simply, as expected value of the system's state (19):

$$\begin{aligned}\hat{\theta}_k &= \mathrm{E}[\theta_k] = \operatorname{atan2}\left(\frac{\mathrm{E}[y_k]}{\mathrm{E}[x_k]}\right) = \operatorname{atan2}\left(\frac{\mathrm{E}[\sin(\theta_k)]}{\mathrm{E}[\cos(\theta_k)]}\right)\\ &= \operatorname{atan2}\left(\frac{\sum_{p=1}^{P} w_k^p \sin(\theta_k^p)}{\sum_{i=p}^{P} w_k^p \cos(\theta_k^p)}\right), \end{aligned} \tag{30}$$

where $\mathrm{E}[.]$ is the expectation operator.

4.4. Algorithm summary

As it was stated in section 4, the particle filtering algorithm follows the SIR scheme. The main idea is to spread the particle set $\{\theta_k^p, w_k^p\}_{p=1}^P$ in all possible directions, take the measurements $\mathbf{z}_k$, resample the particles with the highest probability and estimate the azimuth $\hat{\theta}_k$ from their respective weights. After a few steps, most particles will accumulate around the true azimuth value and track the sound source following the motion model given by (18). If at the particular time step k no valid measurements are available (outlier or no voice activity is detected), a Gaussian noise is added to spread the particles to cover a larger area. If this state lasts longer than a given time period, the algorithm is reset and the particles are again spread in all possible directions.

Initialization step: At time instant $k = 0$ a particle set $\{\theta_0^p, w_0^p\}_{p=1}^P$ (velocities $\dot{x}_0, \dot{y}_0$ set to zero) is generated and distributed accordingly on a unit circle. Since the sound source can be located anywhere around the robot, all the particles have equal weights $w_0^p = 1/P\ \forall p$.

Prediciton step: If there is voice activity detected and the current measurement is valid, all the particles are propagated according to the motion model given by (18). Otherwise, all the particles are corrupted with Gaussian noise, $\mathcal{N}(\mu_c, \sigma_c^2)$. If this state lasts longer than a certain threshold I_c, the algorithm resets to initialization step.

Weight computation: Upon receiving TDOA measurements, DOAs are calculated from (14) and for each DOA a bimodal pdf is constructed from (24). To form the proposed measurement model, all the bimodal pdfs are combined to form (25). The particle weights are calculated from (29) and (25), and normalized so that $\sum_{p=1}^{P} w_k^p = 1$.

Azimuth estimation: At this point we have the approximate discrete representation of the posterior density (25). The azimuth is estimated from (30).

Resampling: This step is applied at each time index ensuring that the particles are resampled respective to their weights. After the resampling, all the particles have equal weights: $\{\theta_k^p, w_k^p\}_{p=1}^P \rightarrow \{\theta_k^p, 1/P\}_{p=1}^P$. We use the *Systematic resampling* algorithm (see [33]), but particle size adaptation is not performed, since we have a modest number of particles required for this algorithm. When the resampling is finished, the algorithm loops back to the prediction step.

The algorithm testing was performed by simulation with a constructed measurement vector $\mathbf{z}_k$ similar to one that would be experienced during experiments. Six measurements were distributed close to the true value ($\theta = 45°$), while the other six were their counterparts. Fig. 9 show first four steps of the algorithm execution. The figures show particles before and after the resampling. We can see that the particles converge to the true azimuth value.

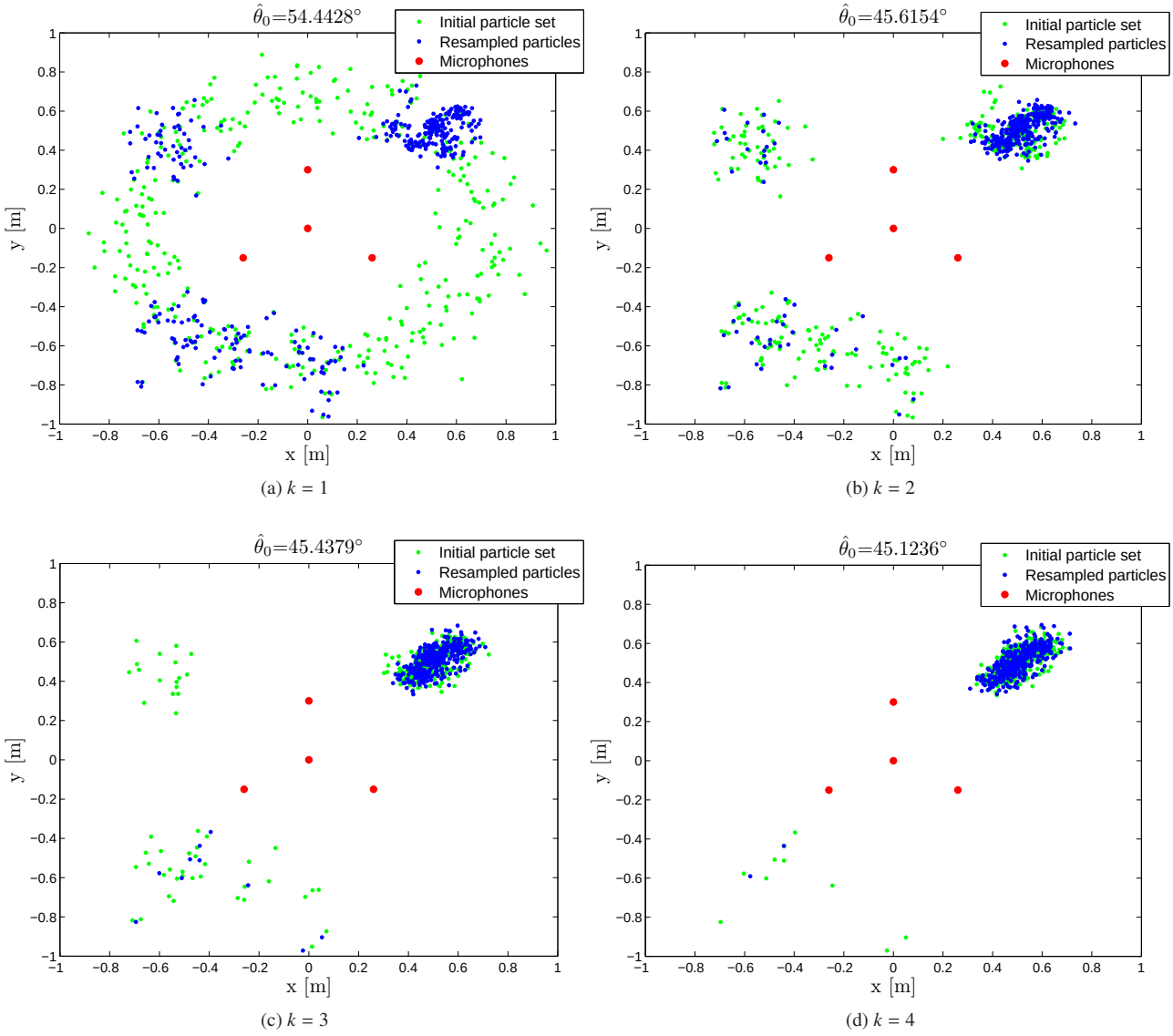
5. Experiments

The microphone array used for experiments is composed of 4 omnidirectional microphones arranged in either Y or square geometry (depending on the experiment). The circle's radius for both array configurations was set to $r = 30$ cm, yielding side length of $a = 0.52$ cm for Y array and $a = 0.42$ cm for square array. The microphone array is placed on a Pioneer 3DX robot as shown in Fig. 14. Audio interface is composed of low-cost microphones, pre-amplifiers and external USB soundcard (whole equipment costing cca. €150). All the experiments were done in real-time, yielding $L/F_s = 21.33$ ms system response time. Real-time multichannel signal processing for the Matlab implementation was realised with Playrec[1] utility, while for the C/C++ implementation, RtAudio API[2] was used. The parameter values used in all experiments are summed up in Tab. 2.

The first set of experiments was conducted in order to qualitatively asses the performance of the algorithm. In these experiments Y array configuration was used and two scenarios

[1] http://www.playrec.co.uk/

[2] http://www.music.mcgill.ca/~gary/rtaudio/



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

Figure 9: Simulation results

were analyzed. Figure 10 shows the first scenario in which a white noise source moved around the mobile robot making a full circle. Figure 11 shows the results from the second scenario, where a white mnoise source made rapid angle changes under 0.5 s ($I_c = 10$ in this case). Both experiments were repeated with smaller array dimensions (a=30 cm), resulting in smaller angle resolution, and no significant degradations to the algorithm were noticed. Performance in adverse noise conditions was also tested in a way that a loud white noise source was present simultaneously to speaker uttering. The algorithm was able to localize the speaker as long as it was louder than the noise source.

The second set of experiments was conducted in order to quantitatively asses the performance of the algorithm. In order to do so, a ground truth system needed to be established. The Pioneer 3DX platform on which the microphone array was placed is also equipped with SICK LMS200 laser range finder (LRF). Adaptive Sample-Based Joint Probabilistic Data Association Filter (ASJPDAF) for multiple moving objects developed in [35] was used for leg tracking. The authors find it to be a good reference system in controlled conditions. Basically, a human speaker walked around the robot uttering a sequence of words, or carried a mobile phone for white noise experiments, while the ASJPDAF algorithm measured range and azimuth from the LRF scan.

In this set of experiments three parameters were calculated: detection reliability, root-mean-square error (RMSE) and standard deviation. To make comparison possible, the chosen parameters are similar to those in [1]. The detection reliability is defined as the percent of samples that fall within ±5° from the ground truth azimuth, RMSE is calculated as deviation from the ground truth azimuth, while standard deviation is simply

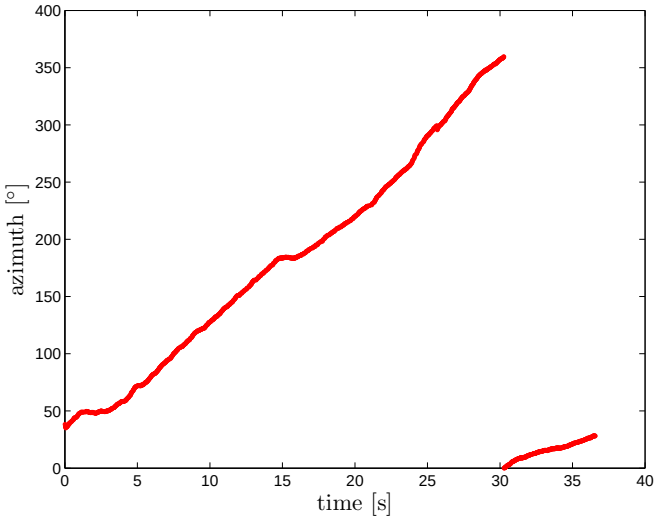


Figure 10: Azimuth estimation for a white noise source making a full circle

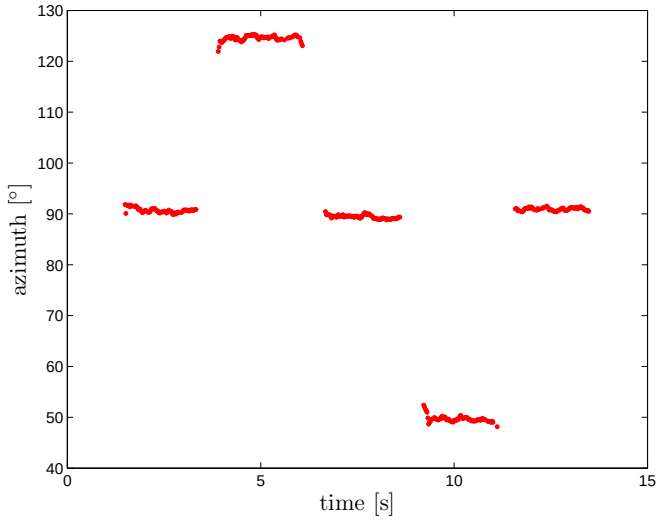


Figure 11: Azimuth estimation for a speaker source making rapid angle changes

Table 1: Results of the second group of experiments

	Y-array		Square array	
Range	W. noise	Voice	W. Noise	Voice
	Detection [%]			
1.50 [m]	97.43	98.93	99.43	97.71
2.25 [m]	97.71	92.86	98.00	96.0
3.00 [m]	94.57	86.86	96.00	91.43
	RMSE [°]			
1.50 [m]	1.90	2.20	1.72	2.19
2.25 [m]	1.61	3.07	1.99	2.83
3.00 [m]	2.38	4.58	1.80	3.95
	Std. deviation [°]			
1.50 [m]	0.96	1.59	0.94	1.36
2.25 [m]	1.10	2.78	1.04	2.30
3.00 [m]	1.65	3.85	1.14	3.01

Table 2: Values of parameters used in the implemented speaker localization algorithm

Signal processing	
L = 1024	rectangular window (no overlap)
F_s = 48 kHz	16-bit precision
SNR Estimation	
$\alpha_{\text{rev}} = 0.85$	$\delta_{\text{rev}} = 0.8$
$\alpha_e = 0.9$	
Voice activity detection	
$\eta = 1$	
Cross-correlation	
$\beta = 0.8$	$c = 344$ m/s
Particle filter	
$\alpha = 0.1$	$\beta = 0.04$
$\delta = L/F_s$	$P = 360$
$\kappa = 20$	$\sigma_v^2 = 0.1$ m/s
$\mu_c = 0$	$\sigma_c^2 = 0.02$
$I_c = 50$	

the deviation of the measured set from its mean value.

The experiments were performed at three different ranges for both the Y and square array configurations, and, furthermore, for each configuration voice and white noise source were used. The white noise source was a train of 50 element 100 ms long bursts, and for the voice source speaker uttered: "Test, one, two, three", until reaching the number of 50 words in a row. In both configurations the source changed angle in 15° or 25° intervals, depending on the range, thus yielding in total 4150 sounds played. The results of the experiments are summed up in Tab. 1, from which it can be seen (for both array configurations) that for close interaction the results are near perfect. High detection rate and up to 2° error and standard deviation rate at distance of 1.5 m are negligible. In general, for both array configurations performance slowly degrades as the range increases. With the range increasing the far-field assumption does get stronger, but the angular resolution is lower, thus resulting in higher error and standard deviation. Concerning different array configurations, it can be seen that square array shows better results in all three parameters, on average up to 2.3% in detection, 0.4° in RMSE, and 0.4° in standard deviation.

The third set of experiments was conducted in order to asses

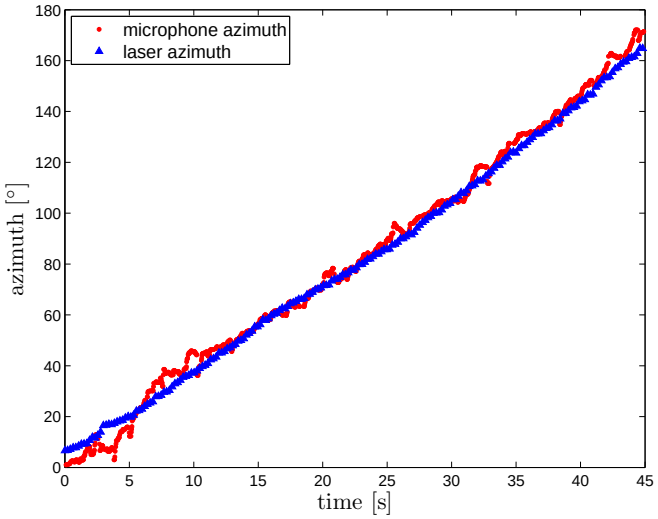


Figure 12: Speaker tracking compared to leg tracking (Y array)

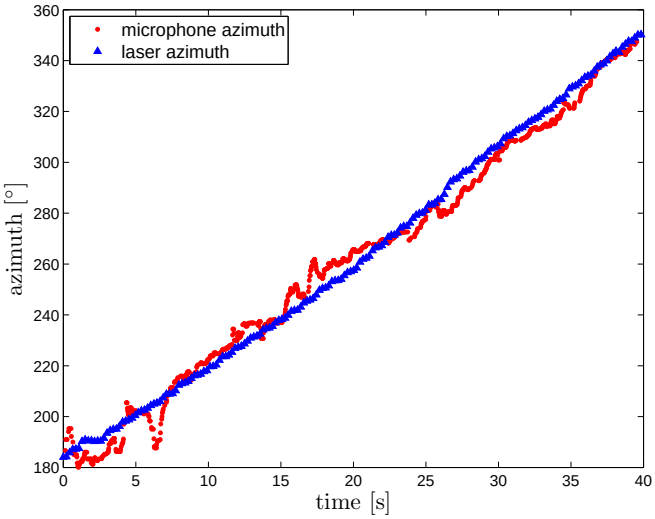


Figure 13: Speaker tracking compared to leg tracking (square array)

the tracking performance of the algorithm. A speaker made a semicircle at approximately 2 m range around the robot uttering: "Test, one, two, three", while at the same time legs were tracked using LRF. The experiment was made for both array configurations. Figures 12 and 13 show the azimuth measured with the leg tracker and with the microphone array arranged in the Y and square configurations, respectively. It can be seen that the square array, in this case, shows bigger deviations from the laser measured azimuth than the Y array does. In Fig. 13 at 6.3 seconds, one of the drawbacks of the algorithm can be seen. It is possible that at an occasion, erroneous measurements might outnumber the correct ones. In this case, wrong azimuths will be estimated for that time, but as can be seen in Fig. 13 the algorithm will get back on track in a short time period.

6. Conclusions and Future Works

Using a microphone array consisting of 4 omnidirectional microphones, an audio interface for a mobile robot that successfully localizes and tracks a speaker was developed. The concept is based on a linear combination of probabilistically modelled Time Difference of Arrival measurements. The measurement model uses the proposed von Mises distribution for Direction of Arrival analysis and for derivation of an adequate azimuth estimation method. In order to handle the inherent multimodal and non-linear characteristics of the system, a particle filtering approach was utilised.

All this resulted with a reliable and elegant algorithm that was tested in real-time with an accurate and precise ground truth method based on leg-tracking with a Laser Range Finder. The implemented Voice Activity Detection algorithm, based on adaptive noise estimation technique, enables the algorithm to function under adverse noise conditions.

Furthermore, two most common microphone array geometries were meticulously analysed. They were compared theoretically based on error sensitivity to Time Difference of Arrival estimation and the robustness to microphone occlusion. Moreover, all the algorithm verification experiments were conducted with both microphone array geometries and the results are summed up in a tabular form. The analysis and experiments showed square array having several advantages over the Y array configuration, but from a practical point of view these two configurations have similar performances.

In order to develop a functional human-aware mobile robot system, future works will strive towards the integration of the proposed algorithm with other systems like leg tracking, robot vision etc. The implementation of a speaker recognition algorithm and a more sophisticated voice activity detector would further enhance the audio interface. Also, by utilising a time difference of arrival estimation method that is capable of tracking multiple speakers, further capabilities of the proposed measurement model could be researched.

Acknowledgment

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under grant No. 036-0363078-3018. The authors would like to thank Srećko Jurić-Kavelj for providing his ASJPDAF leg-tracking algorithm and for his help around the experiments.

References

[1] J.-M. Valin, F. Michaud, J. Rouat, Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering, Robotics and Autonomous Systems 55 (3) (2007) 216–228.

[2] J. H. DiBiase, H. F. Silverman, M. S. Brandstein, Robust Localization in Reverberant Rooms, in: Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

[3] Y. Sasaki, Y. Kagami, S. Thompson, H. Mizoguchi, Sound Localization and Separation for Mobile Robot Tele-operation by Tri-concentric Microphone Array, Digital Human Symposium, 2009.

[4] E. D. Di Caludio, R. Parisi, Multi Source Localization Strategies, in: Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

Figure 14: The robot and the microphone array used in the experiments

[5] J. Merimaa, Analysis, Synthesis, and Perception of Spatial Sound - Binaural Localization Modeling and Multichannel Loudspeaker Reproduction, Ph.D. thesis, Helsinki University of Technology (2006).

[6] J. F. Ferreira, C. Pinho, J. Dias, Implementation and Calibration of a Bayesian Binaural System for 3D Localisation, in: Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics, 2009, pp. 1722–1727.

[7] K. Nakadai, K. Hidai, H. G. Okuno, H. Kitano, Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots, Eurospeech - Scandinavia, 2001.

[8] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Transactions on Acoustics Speech and Signal Processing 24 (4) (1976) 320327.

[9] M. S. Brandstein, J. E. Adcock, H. F. Silverman, A Closed Form Location Estimator for Use in Room Environment Microphone Arrays, IEEE Transactions on Speech and Audion Processing (2001) 45–56.

[10] Y. T. Chan, K. C. Cho, A Simple and Efficient Estimator for Hyperbolic Location, IEEE Transactions on Signal Processing (1994) 1905–1915.

[11] R. Bucher, D. Misra, A Synthesizable VHDL Model of the Exact Solution for Three-Dimensional Hyperbolic Positioning System, VLSI Design (2002) 507–520.

[12] K. Doğançay, A. Hashemi-Sakhtsari, Target tracking by time difference of arrival using recursive smoothing, Signal Processing 85 (2005) 667–679.

[13] I. Marković, I. Petrović, Speaker Localization and Tracking in Mobile Robot Environment Using a Microphone Array, in: J. Basañez, Luis; Suárez, Raúl ; Rosell (Ed.), Proceedings book of 40th International Symposium on Robotics, no. 2, Asociación Española de Robótica y Automatización Tecnologías de la Producción, Barcelona, 2009, pp. 283–288.
URL `http://crosbi.znanstvenici.hr/prikazi-rad?lang=EN\&rad=389094`

[14] T. Nishiura, M. Nakamura, A. Lee, H. Saruwatari, K. Shikano, Talker Tracking Display on Autonomous Mobile Robot with a Moving Microphone Array, in: Proceedings of the 2002 International Conference on Auditory Display, 2002, pp. 1–4.

[15] J. C. Murray, H. Erwin, S. Wermter, A Recurrent Neural Network for Sound-Source Motion Tracking and Prediction, IEEE International Joint Conference on Nerual Networks (2005) 2232–2236.

[16] V. M. Trifa, G. Cheng, A. Koene, J. Morén, Real-Time Acoustic Source Localization in Noisy Environments for Human-Robot Multimodal Interaction, 16th IEEE International Conference on Robot and Human Interactive Communication (2007) 393–398.

[17] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

[18] J. Chen, J. Benesty, Y. A. Huang, Time Delay Estimation in Room Acoustic Environments: an overview, EURASIP Journal on Applied Signal Processing (2006) 1–19.

[19] Y. Ephraim, D. Malah, Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, Speech and Signal Processing (1984) 1109–1121.

[20] I. Cohen, B. Berdugo, Speech Enhancement for Non-Stationary Noise Environments, Signal Processing 81 (2001) 283–288.

[21] I. Cohen, Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging, Speech and Audio Processing 11 (2003) 466–475.

[22] Y. Ephraim, I. Cohen, Recent Advancements in Speech Enhancement, in: C. Dorf (Ed.), Circuits, Signals, and Speech and Image Processing, Taylor and Francis, 2006.

[23] J. Huang, N. Ohnishi, N. Sugie, Sound Localization in Reverberant Environment Based on the Model of the Precedence Effect, IEEE Transactions on Instrumentation and Measurement 46 (1997) 842–846.

[24] J. Huang, N. Ohnishi, X. Guo, N. Sugie, Echo Avoidance in a Computational Model of the Precedence Effect, Speech Communication 27 (1999) 223–233.

[25] K. D. Donohue, J. Hannemann, H. G. Dietz, Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments, Signal Processing 87 (2007) 1677–1691.

[26] S.-I. Kang, J.-H. Song, K.-H. Lee, J.-H. Park Yun-Sik Chang, A Statistical Model-Based Voice Activity Detection Technique Employing Minimum Classification error Technique, Interspeech (2008) 103–106.

[27] E. A. Lehmann, A. M. Johansson, Particle Filter with Integrated Voice Activity Detection for Acoustic Source Tracking, EURASIP Journal on Advances in Signal Processing (2007).

[28] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, N. Sugie, A Model-Based Sound Localization System and its Application to Robot Navigation, Robotics and Autonomous Systems 27 (1999) 199–209.

[29] B. Kwon, G. Kim, Y. Park, Sound Source Localization Methods with Considering of Microphone Placement in Robot Platform, 16th IEEE International Conference on Robot and Human Interactive Communication (2007) 127–130.

[30] J. Vermaak, A. Blake, Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments, in: Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing, 2001.

[31] K. V. Mardia, P. E. Jupp, Directional Statistics, Wiley, New York, 1999.

[32] N. I. Fisher, Statistical Analysis of Circular Data, Cambridge University Press, 1996.

[33] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking, Signal Processing 50 (2001) 174–188.

[34] D. Fox, Adapting the Sample Size in Particle Filter Through KLD-Sampling, International Journal of Robotics Research 22 (2003).

[35] S. Jurić-Kavelj, M. Seder, I. Petrović, Tracking Multiple Moving Objects Using Adaptive Sample-Based Joint Probabilistic Data Association Filter, in: Proceedings of 5th International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2008), 2008, pp. 99–104.