Speaker Localization and Tracking with a Microphone Array on a Mobile Robot Using von Mises Distribution and Particle Filtering

Ivan Marković and Ivan Petrović

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 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.

I. 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 nonambiguous 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 having to respond to constantly changing acoustic conditions, make speaker localization and tracking a formidable problem.

Existing speaker localization strategies can be categorized in four general groups. The first group of algorithms refer to beamforming methods in which the array is steered to various locations of interest and searches for the peak in the output power [1], [2], [3], [4]. The second group includes beamforming methods based upon analysis of spatiospectral correlation matrix derived from the signals received at the microphones [5]. The third group relies on computational simulations of the physiologically known parts of the hearing system, e.g. binaural cue processing [6], [7], [8]. 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 [9], 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 [10], [11], [12], [13]. In mobile robotics, due to small microphone array dimensions, usually hyperbolae intersection is not calculated, only the angle (azimuth and/or elevation) is estimated [14], [15], [16], [17], [18].

Even though the TDOA estimation based methods are outperformed to a certain degree by several more elaborate methods [19], [20], [21], they still prove to be extremely effective due to their elegance and low computational costs. This paper proposes a new speaker localization and tracking method based on TDOA estimation, probabilistic measurement modelling based on von Mises distribution, and particle filtering. Speaker localization and tracking based on particle filtering was also used in [1], [31], [34], [37], but the novelty of this paper is the proposed measurement model used for a posteriori inference about the speaker location. The benefits of the proposed approach are that it solves the front-back ambiguity, increases the robustness by using all the available measurements, and localizes and tracks a speaker over the full range around the mobile robot, while keeping low computational complexity of TDOA estimation based algorithms.

The rest of the paper is organized as follows. Section II

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

Ivan Marković and Ivan Petrović are with the Faculty of Electrical Engineering, Department of Control and Computer Engineering, University of Zagreb, Unska 3, 10000 Zagreb, Croatia, {ivan.markovic@fer.hr, ivan.petrovic@fer.hr}

describes the implemented azimuth estimation method and the voice activity detector. Section III analyses Y and square microphone array geometries, while Section IV 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 V presents the conducted experiments. In the end, Section VI concludes the paper and presents future works.

II. 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 two microphones is described first.

A. 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:

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

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:

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

B. 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 [9], 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:

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

$$G_{i}^{n}(k) = \frac{\xi_{i}^{n}(k)}{1 + \xi_{i}^{n}(k)},$$
(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 [22]:

$$\xi_i^n(k) = \alpha_e [G_i^{n-1}(k)]^2 \gamma_i^{n-1}(k) + (1 - \alpha_e) \max\{\gamma_i^n(k) - 1, 0\},$$
(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 [23], [24]. 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 [25].

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]:

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

where α_{rev} is the reverberation decay, δ is the level of reverberation and $\lambda_{i,rev}^0(k) = 0$. Equation (7) can be seen as modelling the *precedence effect* [26], [27], in order to give less weight to frequencies where recently a loud sound was present.

Using just 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 [28], the main reason for this approach is that speech can exhibit both wideband 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:

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

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

C. 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 [22], [23], [29].

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:

$$p(X_{i}^{n}(k)|H_{0}^{n}(k)) = \frac{1}{\pi\lambda_{i}^{n}(k)} \exp\left(-\frac{|X_{i}^{n}(k)|^{2}}{\lambda_{i}^{n}(k)}\right)$$

$$p(X_{i}^{n}(k)|H_{1}^{n}(k)) = \frac{1}{\pi(\lambda_{i}^{n}(k) + \lambda_{i,x}^{n}(k))} \times \exp\left(-\frac{|X_{i}^{n}(k)|^{2}}{\lambda_{i}^{n}(k) + \lambda_{i,x}^{n}(k)}\right). \quad (9)$$

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

$$\Lambda_{i}^{n}(k) = \frac{p(X_{i}^{n}(k)|H_{1}^{n}(k))}{p(X_{i}^{n}(k)|H_{0}^{n}(k))} \\ = \frac{1}{1+\xi_{i}^{n}(k)} \exp\left(\frac{\gamma_{i}^{n}(k)\xi_{i}^{n}(k)}{1+\xi_{i}^{n}(k)}\right).$$
(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



Fig. 1: Recorded speech signal with corresponding scaled likelihood ratio

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 [30]. 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 [31].

D. Direction of Arrival Estimation

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

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

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):

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



Fig. 2: Direction of arrival angle transformation

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 $\theta_{ij} = 0^\circ$) is performed. The transformation is done with the following equation (angles φ_{24}^+ and θ_{24}^+ in Fig. 2):

$$\theta_{ij}^{\pm} = \alpha_{ij} \pm \varphi_{ij}$$

= $\operatorname{atan2}\left(\frac{y_j - y_i}{x_j - x_i}\right) \pm \operatorname{arccos}\left(\frac{c\Delta\tau_{ij}}{a_{ij}}\right).$ (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 $2 \cdot {M \choose 2}$ possible azimuth values. How to solve the front-back ambiguity and fuse the measurements is explained in Section IV.

III. 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 two microphone array configurations are compared as if placed on a circular surface with fixed 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 θ_{ij}^{\pm} depends solely on the estimation of TDOA. Therefore, it is reasonable to analyse the sensitivity of azimuth estimation to TDOA estimation error. Furthermore, it is shown that this sensitivity depends on the microphone array configuration.

Firstly, we define the error sensitivity of azimuth estimation to TDOA measurement, s_{ij} , as follows [32]:

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

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

$$s_{ij} = \frac{c}{a_{ij}} \frac{1}{|\sin(\theta_{ij} - \alpha_{ij})|}.$$
(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 analysed for circle radius r = 30 cm, thus yielding square side length a = 0.42cm and triangle side length a = 0.52 cm. 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. 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 $a_{ij} = r$, the length of the baseline connecting the microphones on the vertices with the microphone in the orthocenter, and



Fig. 3: Possible array placement scenarios



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

other for $a_{ij} = r\sqrt{3}$, the length of the baseline connecting the microphones on the vertices of the triangle. The first group has the largest error sensitivity value of 3.8 approximately, and the second group has the largest error sensitivity value of 2.2 approximately. For the square array, Fig. 4 shows also two groups of sensitivity curves: one for $a_{ij} = r\sqrt{2}$, the side length of the square, and the other for $a_{ij} = 2r$, the diagonal length of the square. The first group has the largest error sensitivity value of 3.3 approximately, and the second group has the largest error sensitivity value of 2.3 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 are utilising all microphone pair measurements to estimate azimuth, it is practical to compare joint error sensitivity (JES) curves, which we define as:

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



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



Fig. 6: Joint error sensitivity curves for both Y and square microphone array configurations

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 array, 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 [33], 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 [14].

IV. 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 predicting the evolution of the speaker state over time (system model), and second relating the noisy measurements to the speaker 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.

Before presenting the models and the algorithm in details, we describe in general major successive steps of the algorithm. The algorithm starts with an initialization step at which we assume that the speaker can be located anywhere around the mobile robot, i.e. we assume that the angle has a uniform distribution. Any further action is taken only if voice activity is detected by the VAD described in II-C. When the VAD condition is fulfilled, the algorithm proceeds with predicting the state of the speaker trough the sound source dynamics model described in Section IV-A. Once measurements are taken, a measurement model based on a mixture of von Mises distributions, described in Section IV-B, is constructed. Since this model is inherently multimodal, particle filtering approach, described in Section IV-C, is utilised to represent the pdf of such measurement model and to effectively estimate the speaker azimuth as the expected value this pdf.

A. Sound source dynamics model

The sound source dynamics is modelled by the well behaved Langevin motion model [34]:

$$\begin{bmatrix} \dot{x}_{k} \\ \dot{y}_{k} \end{bmatrix} = \alpha \begin{bmatrix} \dot{x}_{k-1} \\ \dot{y}_{k-1} \end{bmatrix} + \beta \begin{bmatrix} v_{x} \\ v_{y} \end{bmatrix},$$

$$\begin{bmatrix} x_{k} \\ y_{k} \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \delta \begin{bmatrix} \dot{x}_{k} \\ \dot{y}_{k} \end{bmatrix},$$
(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, $v_x, v_y \sim \mathcal{N}(0, \sigma_v)$ is the stohastic velocity disturbance, α and β are model parameters, and δ is the time between update steps.

Although the state of the speaker is defined in 2D by (18), which is found to describe well motion of the speaker [1], [31], [34], [37], when working with such a small microphone array

like in mobile robotics, due to high uncertainty in distance estimation only the speaker azimuth is determined [14], [32]. Therefore the system state, i.e. the speaker azimuth, is defined via the following equation:

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

B. 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 $\mathbf{z}_k = [\theta_{ij}^{\pm}, \ldots, \theta_{M,M-1}^{\pm}]_k, i \neq j, \{i, j\} = \{j, i\}$ 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 [35], [36]:

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

where $0 \le \theta_{ij} < 2\pi$ is the measured azimuth, $0 \le \theta_k < 2\pi$ 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:

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

$$p_{ij}\left(\theta_{ij,k}^{\pm}|\theta_{k},\kappa\right) = p_{ij}\left(\theta_{ij,k}^{+}|\theta_{k},\kappa\right) + p_{ij}\left(\theta_{ij,k}^{-}|\theta_{k},\kappa\right)$$
$$= \frac{1}{2\pi I_{0}(\kappa)}\exp\left[\kappa\cos\left(\theta_{ij,k}^{+}-\theta_{k}\right)\right] + \frac{1}{2\pi I_{0}(\kappa)}\exp\left[\kappa\cos\left(\theta_{ij,k}^{-}-\theta_{k}\right)\right]$$
(24)

Having all pairs modelled as a sum of two von Mises distributions, we propose a linear combination of all those



azimuth [°]

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

pairs to represent the microphone array measurement model. Such a model has the following multimodal pdf:

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

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.

By looking at (16), we can see that the error sensitivity is the greatest when the argument in the sine function is zero. This corresponds to a situation when speaker is located at the baseline of a microphone pair. Furthermore, we can see that the error sensitivity is the smallest when speaker is on a line perpendicular to the microphone pair baseline. Since we need the coefficients β_{ij} to give the least weight to a microphone pair in the former situation and the most weight to a microphone pair in the latter situation, it would be appropriate to calculate β_{ij} by inverting (16). However, we use scaled and inverted (16):

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

where the ratio c/a_{ij} is set to one, since it is constant, and the coefficients are scaled so as to never cancel out completely a possibly unfavourable pdf. We can also see that the mixture coefficients are a function of the estimated azimuth and that this form can only be applied after the first iteration of the algorithm.

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.

C. 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, multimodal distributions, and it has been shown in [1], [31], [34], [37] to be practical for sound source tracking. Moreover, in [1] it is successfully utilised to track multiple sound sources. Particle filtering 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 $\{\theta_k^p, w_k^p\}_{p=1}^P$ denote a random measure that characterises the posterior pdf $p(\theta_k | \mathbf{z}_k)$, where $\{\theta_k^p, p = 1, \dots, P\}$ is a set of particles with associated weights $\{w_k^p, p = 1, \dots, P\}$. The weights are normalised so that $\sum_p w_k^p = 1$. Then, the posterior density at k can be approximated as [38]:

$$p(\theta_k | \mathbf{z}_k) \approx \sum_{p=1}^{P} w_k^p \delta(\theta_k - \theta_k^p),$$
(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 [38]:

$$w_k^p \propto w_{k-1}^p p(\mathbf{z}_k | \boldsymbol{\theta}_k^p), \tag{28}$$

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

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 $w_k^p = 1/P$. In the SIR scheme, resampling is applied at each time index. Since we have $w_{k-1}^p = 1/P \ \forall p$, the weights are simply calculated from:

$$w_k^p \propto p(\mathbf{z}_k | \theta_k^p).$$
 (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 *KLDsampling* procedure proposed in [39]. 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



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

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):

$$\hat{\theta}_{k} = \mathbf{E}[\theta_{k}] = \operatorname{atan2}\left(\frac{\mathbf{E}[y_{k}]}{\mathbf{E}[x_{k}]}\right) = \operatorname{atan2}\left(\frac{\mathbf{E}[\sin(\theta_{k})]}{\mathbf{E}[\cos(\theta_{k})]}\right)$$
$$= \operatorname{atan2}\left(\frac{\sum_{p=1}^{P} w_{k}^{p} \sin(\theta_{k}^{p})}{\sum_{i=p}^{P} w_{k}^{p} \cos(\theta_{k}^{p})}\right), \quad (30)$$

where E[.] is the expectation operator.

D. Algorithm summary

In order to get a clear overview of the complete algorithm, we present its flowchart diagram in Fig. 9, and hereafter describe each step of the algorithm with implementation details.

Initialization: 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 $w_0^p = 1/P \ \forall p$, i.e. we assume that the angle has a uniform distribution.

Voice activity detection: In the speaker detection part a VAD is applied to recorded signals. Firstly, LR is calculated with (10), and then a binary-decision procedure is made through (11). If the VAD result is smaller the threshold η , i.e. no voice activity is detected, we proceed to a decision logic in which we either add white Gaussian noise (WGN) defined with $\mathcal{N}(\mu_c, \sigma_c^2)$ to particle positions, in order to account for speaker moving during a silence period, or if this state lasts longer than a given threshold I_c , the algorithm is reset and we simply go back to the initialization step. If the VAD result is larger then the threshold η , i.e. voice activity is detected, then the algorithm proceeds to speaker state prediction.

Prediction: In this step all the particles are propagated according to the motion model given by (18).

Measurement model construction: 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: $\{\theta_k^p, w_k^p\}_{p=1}^P \rightarrow \{\theta_k^p, 1/P\}_{p=1}^P$. The SIR algorithm is used (see [38]), 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 speaker detection step.

V. TEST RESULTS

The proposed algorithm is thoroughly tested by simulation and experiments with a microphone array composed of four microphones arranged in either Y or square geometry (depending on the experiment). The circle 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.



Fig. 9: Flowchart diagram of the proposed algorithm

Hereafter we present first an illustrative simulation and then the experimental results.

A. Simulation

In order to get a deeper insight into particle behaviour, in this section we present an illustrative simulation. We constructed a measurement vector \mathbf{z}_k similar to one that would be experienced during experiments. Six measurements were distributed close to the true value ($\theta = 45^\circ$), while the other six were their counterparts, thus yielding:

$$\mathbf{z}_{k} = \begin{bmatrix} \theta_{12}^{-} & \theta_{13}^{-} & \theta_{14}^{+} & \theta_{23}^{+} & \theta_{24}^{+} & \theta_{34}^{+} & \theta_{12}^{+} & \theta_{13}^{+} & \theta_{14}^{-} & \theta_{23}^{-} & \theta_{24}^{-} & \theta_{34}^{-} \end{bmatrix}$$

= $\begin{bmatrix} 42^{\circ} & 44^{\circ} & 45^{\circ} & 45^{\circ} & 46^{\circ} & 48^{\circ} & 135^{\circ} & 75^{\circ} & 225^{\circ} & 15^{\circ} & 315^{\circ} & 255^{\circ} \end{bmatrix}$
(31)

The algorithm was tested with such z_k for the first four iterations of the algorithm execution. The results are shown in Fig. 10 where particles before and after the resampling are shown. We can see that in the first step the particles are spread uniformly around the microphone array. After the first measurement, the particle weights are calculated and the particles are resampled according to their respective weights. This procedure is repeated throughout the next iterations, and we can see in Fig. 10 that the particles converge to the true azimuth value.

B. Experiments

The microphone array consisting of four omnidirectional microphones is placed on a Pioneer 3DX robot as shown in Fig. 16. Audio interface is composed of low-cost microphones, pre-amplifiers and external USB soundcard (whole equipment costing circa €150). All the experiments were done in realtime, yielding $L/F_s = 21.33$ ms system response time. Realtime multichannel signal processing for the Matlab implementation was realised with Playrec¹ utility, while RtAudio API² was used for the C/C++ implementation. The experiments were conducted in a classroom which has dimensions of $7 \,\mathrm{m} \times 7 \,\mathrm{m} \times 3.2 \,\mathrm{m}$, parquet wooden flooring, one side covered with windows and a reverberation time of 850 ms. During the experiments, typical noise conditions were present, like computer noise and air ventilation. In the experiments two types of sound sources were used; a WGN source and a single speaker uttering a test sequence. The parameter values used in all experiments are summed up in Tab. I.

The first set of experiments was conducted in order to qualitatively assess the performance of the algorithm. Two type of experiments were performed; one with a stationary robot and the other with a moving robot.

In the experiments with a stationary robot Y array configuration was used, and a loud white noise sound source, since it represents the best-case scenario in which all the frequency bins are dominated by the information about the sound source location. Two cases were analyzed. Figure 11 shows the first case in which a sound source moved around the mobile robot at a distance of 2 m making a full circle.

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

Figure 12 shows the results from the second case, where a sound source made rapid angle changes at a distance of 2 m under 0.5 s (initialization threshold was set to $I_c = 10$). 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.

In the experiment with a moving robot the speaker was stationary and uttered a sequence of words with natural voice amplitude. This experiment was also conducted with Y array configuration. The robot started at 3 m away from the speaker and moved towards the speaker until reaching distance of 1 m. Then the robot made a 180° turn and went back to the previous position, where it made another 180° turn. The results form this experiment are shown in Fig. 13.

The second set of experiments was conducted in order to quantitatively assess 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 [40] was used for leg tracking. The authors find it to be a good reference system in controlled conditions. Measurement accuracy of the LMS200 LRF is ± 35 mm, and due to determining the speaker location as the centre between the legs of the speaker, we estimate the accuracy of the ASJPDAF algorithm to be less than 0.5° . In the experiments, 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 percentage of samples that fall within $\pm 5^{\circ}$ from the ground truth azimuth, RMSE is calculated as deviation from the ground truth azimuth, while standard deviation is simply 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. II, 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

¹http://www.playrec.co.uk/



Fig. 10: Simulation results

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.

In [1], where an array of eight microphones was used and a beamforming approach, similar experiments were performed with an open and closed array configuration. For the open configuration, our algorithm shows smaller detection reliability of less than 4% on average, and larger RMSE of less than 2° on average. For the closed configuration, our algorithm shows the same detection reliability on average, and larger RMSE of less then 1° on average.

From the previous discussion we can see that the algorithm proposed in [1] shows better or equal performance, on average, in both detection reliability and RMSE. However, in [1] an array of eight, compared to four, microphones was used and a computationally more expensive approach was utilised. The beamforming approach is based on heuristically dividing the space around the mobile robot into a direction grid, steering the microphone array to all possible directions, and for each direction an expression like (8) is calculated for all microphone pairs. Although more complex, it does however have an advantage of being able to track multiple simultaneously talking speakers.

Exactly how much localization based on beamforming is more complex than our algorithm depends on the previously mentioned heuristic division, i.e. direction search grid resolution. For an example, our algorithm requires one calculation of (8) for all microphone pairs to estimate speaker azimuth, while a beamforming approach with a grid having angular resolution of 1° would require 360 calculations of (8) for all microphone pairs.

The third set of experiments was conducted in order to assess 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



Fig. 11: Tracking azimuth of a white noise source making a full circle



Fig. 12: Tracking azimuth of a white noise source making rapid angle changes

both array configurations. Figures 14 and 15 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. 15 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. 15 the algorithm gets back on track in a short time period.

VI. CONCLUSIONS AND FUTURE WORKS

Using a microphone array consisting of four omnidirectional microphones, a new audio interface for a mobile robot that successfully localizes and tracks a speaker was developed. Novelty of the proposed algorithm is in the method based



Fig. 13: Speaker tracking while robot is in motion

TABLE I: Values of parameters used in the implemented speaker localization algorithm

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

on a linear combination of probabilistically modelled time difference of arrival measurements, yielding a measurement model which uses 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. The major contribution of the proposed algorithm is that it solves the front-back ambiguity and a unique azimuth value is calculated in a robust and computationally undemanding manner. Indeed, the number of required cross-correlation evaluations is equal to the number of different microphone pairs. Moreover, by integrating a voice activity detector to the time difference of arrival estimation, operation under adverse noisy conditions is guaranteed up to the performance of the voice activity detector itself. The algorithm accuracy and precision was tested in real-time with

TABLE II: Experimental results of the algorithm performance for Y and square array configuration

	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	



Fig. 14: Speaker tracking compared to leg tracking (Y array)

a reliable ground truth method based on leg-tracking with a laser range finder.

Furthermore, two most common microphone array geometries were meticulously analysed and compared theoretically based on error sensitivity to time difference of arrival estimation and the robustness to microphone occlusion. 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.

ACKNOWLEDGMENT

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



Fig. 15: Speaker tracking compared to leg tracking (square array)



Fig. 16: The robot and the microphone array used in the experiments

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

- 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.

[4] B. Mungamuru, P. Aarabi, Enhanced Sound Localization., IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : a Publication of the IEEE Systems, Man, and Cybernetics Society 34 (3) (2004) 1526–40.

URL http://www.ncbi.nlm.nih.gov/pubmed/15484922

- [5] E. D. Di Caludio, R. Parisi, Multi Source Localization Strategies, in: Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.
- [6] J. Merimaa, Analysis, Synthesis, and Perception of Spatial Sound -Binaural Localization Modeling and Multichannel Loudspeaker Reproduction, Ph.D. thesis, Helsinki University of Technology (2006).
- [7] 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.
- [8] K. Nakadai, K. Hidai, H. G. Okuno, H. Kitano, Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots, Eurospeech - Scandinavia.
- [9] 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.
- [10] 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.
- [11] Y. T. Chan, K. C. Cho, A Simple and Efficient Estimator for Hyperbolic Location, IEEE Transactions on Signal Processing (1994) 1905–1915.
- [12] R. Bucher, D. Misra, A Synthesizable VHDL Model of the Exact Solution for Three-Dimensional Hyperbolic Positioning System, VLSI Design (2002) 507–520.
- [13] K. Doğançay, A. Hashemi-Sakhtsari, Target tracking by time difference of arrival using recursive smoothing, Signal Processing 85 (2005) 667– 679.
- [14] 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

- [15] 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.
- [16] 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.
- [17] 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.
- [18] J. Valin, F. Michaud, J. Rouat, D. Létourneau, Robust Sound Source Localization Using a Microphone Array on a Mobile Robot, in: Proceedings on the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2003, pp. 1228–1233. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.70. 4899\&rep=rep1\&type=pdf
- [19] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.
- [20] 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.
- [21] A. Badali, J.-M. Valin, F. Michaud, P. Aarabi, Evaluating Real-Time Audio Localization Algorithms for Artificial Audition in Robotics, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2009, pp. 2033–2038. doi:10.1109/IROS.2009.5354308. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=

5354308

- [22] Y. Ephraim, D. Malah, Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator, Speech and Signal Processing (1984) 1109–1121.
- [23] I. Cohen, B. Berdugo, Speech Enhancement for Non-Stationary Noise Environments, Signal Processing 81 (2001) 283–288.
- [24] I. Cohen, Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging, Speech and Audio Processing 11 (2003) 466–475.

- [25] Y. Ephraim, I. Cohen, Recent Advancements in Speech Enhancement, in: C. Dorf (Ed.), Circuits, Signals, and Speech and Image Processing, Taylor and Francis, 2006.
- [26] 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.
- [27] J. Huang, N. Ohnishi, X. Guo, N. Sugie, Echo Avoidance in a Computational Model of the Precedence Effect, Speech Communication 27 (1999) 223–233.
- [28] 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.
- [29] J. Sohn, N. S. Kim, W. Sung, A Statistical Model-Based Voice Activity Detection, IEEE Signal Processing Letters 6 (1) (1999) 1–3. doi:10.1109/97.736233. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber= 736233
- [30] 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.
- [31] E. A. Lehmann, A. M. Johansson, Particle Filter with Integrated Voice Activity Detection for Acoustic Source Tracking, EURASIP Journal on Advances in Signal Processing.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] K. V. Mardia, P. E. Jupp, Directional Statistics, Wiley, New York, 1999.
- [36] N. I. Fisher, Statistical Analysis of Circular Data, Cambridge University Press, 1996.
- [37] D. B. Ward, E. A. Lehmann, R. C. Williamson, Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment, IEEE Transactions on Speech and Audio Processing 11 (6) (2003) 826–836. doi:10.1109/TSA.2003.818112. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber= 1255469
- [38] 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.
- [39] D. Fox, Adapting the Sample Size in Particle Filter Through KLD-Sampling, International Journal of Robotics Research 22.
- [40] 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.