Binaural Bearing Only Tracking of Stationary Sound Sources in Reverberant Environment

Ingo Kossyk[1], Michael Neumann[2] and Zoltan-Csaba Marton[1]

***Abstract*— In this work we present a framework for the estimation of the Cartesian position of stationary sound sources in reverberant environments and under the influence of heavy clutter based on binaural bearing measurements. We employ a particle filter (PF) on binaural measurements to estimate the position of the sound sources in a bearing only tracking (BOT) formulation and investigate how the estimation accuracy can be improved in reverberant environments by applying a gating method that is inspired by the precedence effect. We evaluate the interaural coherency in order to identify time frequency units of the received signals that show a high linear dependency and therefore are potentially dominated by the direct sound emitted by sound sources. We use a particle filter for state estimation and lay out the theoretical model for state representation, propagation and estimation. The feasibility of the presented methods is evaluated in simulations and we give first results of tracking performance when applied to real world binaural localization measurements of a sound source in a typical reverberant scenario. Our results show that gating the binaural bearing measurements with the interaural coherency can improve localization accuracy to a large degree.**

I. INTRODUCTION

The challenge in BOT problems is that the Cartesian position of potential sound sources has to be estimated without knowing the distance to the target by taking into account the control input, position and orientation of the robot and bearing measurements. Furthermore, binaural measurements of sound source localization are heavily affected by reflections in the environment - regarded as reverberation in this work - and by disturbances caused by arbitrary noise sources and false detections which we will regard as clutter. Overlapping sounds of multiple sources can lead to the detection of phantom sound sources. Therefore, the challenge of solving the BOT problem in binaural sound localization is that all these effects combined lead to a high variance of the probability distribution and to a high degree of clutter in the bearing only measurements.

Interaural time difference (ITD) and interaural intensity difference (IID) can be evaluated in order to localize sounds in the horizontal plane, while binaural localization in the median plane is possible when IID and ITD are analyzed together with the filter effects that are introduced by reflections in the pinna, head and torso of the binaural receiver [1], [2], [4]. In the literature there exist several approaches to computational binaural sound localization, for example [3], [4], [8], [9], [16]. Keyrouz et al. [16] presented a method that is based on applying principal component analysis reduction techniques to a known dataset of head related transfer functions (HRTF) and used a reduced HRTF representation in order to localize sound sources by applying inverse filtering with the reduced data and cross correlation. It was also shown that localization is possible with machine learning methods by Deleforge et al. in [8], [9] by means of binaural manifold learning. In the here presented work we rely on the cross channel approach which is proposed by MacDonald in [3]. Experiments have shown that humans localize sound sources mainly according to the directional information contained in the direct sound. This effect has been experimentally investigated in [10], [11] and is known as the precedence effect. Promising results have been shown for using models of the precedence effect in computational auditory scene analysis systems in [5]. Multitarget tracking solutions with PF have been presented in several works, for example by solving the joint probabilistic data association problem (JPDA) [15], [17] or by estimating the joint multitarget probability density (JMPD) [18]. In this work we rely on the latter. Bearing only tracking problems of multiple moving targets with PFs have been investigated specially in the application of radar detection and possible solutions are given in [14], [19]. The signal model of the PF we present in this work is similar to earlier works, as for example presented in [22], where its feasibility has been shown with the application on localization of moving sound sources (for example with GCC [23], AEDA [24]) in reverberant environments with a stationary and widely spaced microphone array. Another

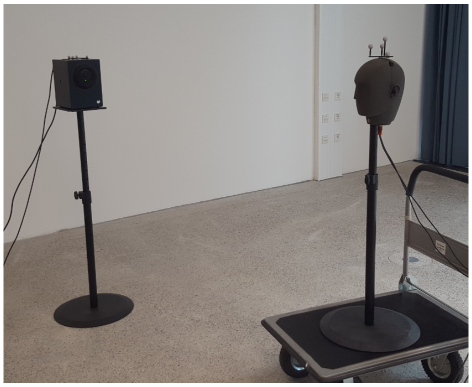

Fig. 1: Experimental setup in reverberant environment with the binaural dummy head KU 100 on a mobile platform and a coaxial speaker for the simulation of a stationary point source. The ground truth was acquired with an optical tracking system.

[1] Ingo Kossyk and Zoltan-Csaba Marton are with the institute of Robotics and Mechatronics at German Aerospace Center (DLR e.V.), Department of Perception and Cognition, Münchner Strasse 20, 82234 Wessling, Germany ingo.kossyk@dlr.de, zoltan.marton@dlr.de

[2] Michael Neumann is with the TU Ilmenau, Ehrenbergstrasse 29, 98693 Ilmenau, Germany michael.neumann@tu-ilmenau.de

more recent work by Evers et al. describes an approach of BOT acoustic tracking of moving speakers for robot audition based on an extended Kalman filter. In the aforementioned work the authors evaluate the performance of their proposed system with simulations of direction of arrival measurements directly without taking into account real world conditions [20]. Our contribution in this work is that we investigated a PF formulation when applied to *bearing only* measurements that result from *binaural sound source localization* with a moving observer and that we evaluate it with *real data* when combined with methods that can improve accuracy and robustness in *reverberant environments*.

In this work we will first lay out the theoretical foundations regarding binaural sound source localization in reverberant environments. See Fig. 2 for an overview of the presented methods. We will explain how we solve the BOT problem with a PF and give the formulations of the state space representation, the state propagation and sensor model. Next, we will show the results of the evaluation of our proposed methods in simulation and with *real world measurements* in a usual *reverberant environment*. Finally, we will conclude with a discussion of the results and possible fields of application. We will also give an outlook on possible and necessary improvements and shortly discuss interesting insights which we gathered during experiments.

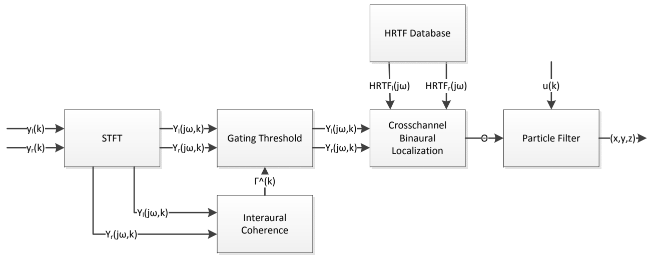

Fig. 2: Structure and dataflow of the localization methods.

II. THEORETICAL FOUNDATIONS

A. Binaural Sound Source Localization

In the literature the commonly used model for binaural localization of sound sources relies on the evaluation of the Interaural Time Difference (ITD) and Interaural Intensity Difference (IID) between the two ears of a binaural sensor. These features are mainly accounted to the localization in the azimuthal plane of the binaural sensor. When combined with the spectral filtering effects due to reflections on the torso, head and pinna of the ears ITD and ILD form a head related transfer function (HRTF) which makes localization in the median plane possible, too. In this work we assume a robotic system with two microphones which are placed in the ear canals of artificial pinna on the right and left side of the robots head. In the following all formulations are given in the frequency domain and we describe the angle dependent binaural signal model under anechoic conditions as:

$$Y_k(j\omega,k) = HRTF_m(j\omega,\theta) * S(j\omega,k) + V(j\omega,k) \\ m \in [l,r] \tag{1}$$

$S(jw,k)$ is the Short Time Fast Fourier Transformed (STFT) monoaural source signal of time instance k in dependency of complex frequency $j\omega$. $HRTF_m(jw,\theta); m \in [l,r]$ are the head related transfer functions of the left and the right ear which are unique for every possible bearing θ of a sound source relative to the binaural sensor. Here θ is an azimuth and elevation pair in spherical coordinates for each possible direction of arrival of a wavefront relative to the sensor. $Y_m(j\omega,k); m \in [l,r]$ are the signals that are received at the left and right ear at each time instance. $V(j\omega,k)$ is an additional noise term due to the characteristics of the sensor. We use the binaural sound source localization algorithm as proposed by MacDonald [3] in order to estimate the filter pair and therefore bearing of the measured sound source at time instance k.

B. Interaural Coherence

The interaural coherence (IC) is inspired by the precedence effect which has been experimentally investigated in human hearing [10], [11]. The precedence effect describes the phenomena that human listeners localize sounds mainly due to the directional information contained in the direct sound wavefronts and directional information of reflections and diffuse sounds is suppressed. In reverberant environments the signal model of Eq. (1) changes because sounds received at the left and right ear are composed of the direct wave front and delayed wavefronts due to reflections in the environment. Furthermore, the delayed wave fronts are arriving the binaural sensor from arbitrary directions and are filtered with different head related transfer functions, hence:

$$Y_m(j\omega,t) = BRTF_m(j\omega,\theta) * S(j\omega,k) + V(j\omega,k) \\ m \in [l,r] \tag{2}$$

Here, the binaural room transfer functions $BRTF_m(j\omega,\theta); m \in [l,r]$ are modeled as the combination of a Room Transfer Function (RTF) and the angular dependent HRTFs for direct wavefronts and reflections. The signal that is received at the ears is a superposition of the direct sound filtered with a HRTF and time delayed reflections that have been filtered with HRTFs of arbitrary bearings depending on the position of the observer and the source. Late reflections in reverberant environments are assumed to be diffuse. The superposition of a received sound with early and late reflections leads to a decrease in the coherency of the signals that are received by both ears. Therefore, valid and mostly undisturbed localization is possible when the direct sound signal portions of the measured signals at the ears are evaluated. In order to do so, one can evaluate the linear dependency of the signals received at the ears. The IC is a measure of the linear dependency of the received signals and determined for each frequency bin in the STFT frames. Hence, the IC [5] can be evaluated in order to determine whether a received time frequency (TF) representation is originating from the direct sound. The IC is defined as follows:

$$\Gamma_{l,r}(j\omega,k) = \frac{\Phi_{l,r}(j\omega,k)}{\sqrt{\Phi_{l,l}(j\omega,k) * \Phi_{r,r}(jw,k)}} \tag{3}$$

$\Phi_{l,r}(jw,t)$ represents the cross-power spectral density (CPSD) and $\Phi_{l,l}(jw,t)$ and $\Phi_{r,r}(jw,t)$ the auto-power spectral density (APSD) of the time aligned signals received at the left and the right ear, respectively. In our experiments we discovered that a time alignment is not necessarily required and can be neglected. In the actual implementation a recursive smoothing is applied to the calculation of the CPSD and APSDs, hence:

$$\begin{aligned}\Phi_{l,l}(j\omega,k) =& \beta\Phi_{l,l}(j\omega,k-1) + (1-\beta)|Y_l(j\omega,k)|^2 \\ \Phi_{l,r}(j\omega,k) =& \beta\Phi_{l,r}(j\omega,k-1) \\ &+ (1-\beta)Y_l(j\omega,k) * \overline{Y_r(j\omega,k)}\end{aligned} \tag{4}$$

The smoothed IC $\Phi_{r,r}(jw,k)$ is calculated according to Eq. (4) where β represents the smoothing parameter. In this work a pair of TF units is regarded containing direct sound and therefore valid by thresholding the mean IC of all frequency bins:

$$\Gamma(\hat{k}) = \begin{cases} 1 \text{ if } \frac{\sum \Gamma_{l,r}(j\omega,k)}{N_{bin}} > \alpha \\ 0 \text{ else} \end{cases} \tag{5}$$

This is a valid assumption for wideband signals that are dense with regards to the time frequency representation, which is the case for white noise. Narrowband signals require more elaborate methods of evaluation of the IC which are beyond the scope of this work. In our experiments we evaluated every received pair of TF units received at the ears with Eq. (5) and gated the frames indicated by $\Gamma(\hat{t}) = 1$ into the binaural localization algorithm and then into the BOT PF for further processing. A side effect of this approach is that the computational burden of the binaural localization with the cross channel algorithm can be avoided when the IC is low by discarding the respective TF units.

C. Binaural Bearing only Tracking

Measurements of a binaural sensor consist only of relative bearings to potential sound source locations. The distance to the target is usually unknown. Simple triangulation, however, is infeasible in real applications due to measurement uncertainty. Moreover, binaural measurements contain a high degree of clutter which violates the assumption of a purely Gaussian probability density. The BOT problem can be tackled with a PF in order to achieve robust tracking in the presence of noise and non-linearities like clutter. One drawback of PFs is that they are usually computational demanding due to the large amount of simulated particles required.

1) State space representation: We used a PF in order to track multiple stationary sound sources. In multitarget applications with a PF the data association problem has to be solved in order to identify which target has to be associated with a received measurement. In the literature this is known as the data association problem, where the joint probability data association (JPDA) is estimated with the PF [15], [17]. In our work we avoid the data association problem by approximating the joint multitarget probability density with the PF as presented in [18]. The advantage of this method is that the computational complexity of solving the data association problem can be avoided and the implementation complexity is modest. Moreover, the estimation of the overall system state, like for example the estimated count of targets, is a comparably simple operation. In this work we implemented the JMPD PF without the optimizations for coupled and independent partitions that are presented in [18]. Accordingly the true multitarget state of the system for T targets is defined as:

$$X = [x_1, x_2, ..., x_{T-1}, x_T] \tag{6}$$

Therefore, the JPMD representation in Eq. (7) is related to Eq. (6) so that a particle consists of a variable count of states, hereby called partitions of the multitarget state vector:

$$X_p = [x_{p,1}, x_{p,2}, ..., x_{p,T_p-1}, x_{p,T_p}] \tag{7}$$

where T_p can be an arbitrary positive integer. Every $x_{p,j} = [x,y,z]$ denotes a potential state of cartesian coordinates of a target j relative to the observer in particle p. Let δ_D denote the Dirac delta:

$$\delta(X - X_p) = \begin{cases} 0 & \text{if } T \neq T_P \\ \delta(X - X_p) & \text{otherwise} \end{cases} \tag{8}$$

It follows the weighted JMPD approximation by a set of particles X_p and weights w_p:

$$p(\mathbf{X}, T \mid \mathbf{Z}) \approx \sum_{p=1}^{N_{part}} w_p \delta(\mathbf{x} - \mathbf{x}_p) \tag{9}$$

where $\mathbf{Z}$ is a measurement and $\sum w_p = 1$. N_{part} is the count of particles used in the filter. For details about the theoretical foundations of the JMPD please refer to [18].

2) State Propagation: In order to model the system dynamics the motion of the target and of the observer have to be taken into account. The observer's rotation in the azimuthal plane is assumed to be small, so the motion between two time steps is assumed to be nearly linear. Since the aim of this work is to track stationary targets, a partition's state vector is reduced to a position in Cartesian coordinates. Changes in the state are now only the result of the observer's motion and the influence of the process noise. The state dynamics of a partition are modelled according to the well known constant velocity (CV) model:

$$\mathbf{x}_t(k) = \mathbf{A}\mathbf{x}_t(k-1) - \mathbf{B}\mathbf{u}(k-1) + \mathbf{v}(k-1) \tag{10}$$

Here $\mathbf{A}$ is defined simply as a $3X3$ identity matrix:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{11}$$

We define the control matrix $\mathbf{B}$ as:

$$\mathbf{B} = \begin{pmatrix} \Delta T & 0 & 0 & 0 \\ 0 & \Delta T & 0 & 0 \\ 0 & 0 & \Delta T & 0 \end{pmatrix} \tag{12}$$

and

$$\mathbf{u} = (\dot{x}_{obs}, \dot{y}_{obs}, \dot{z}_{obs}, \phi_{obs})^T \tag{13}$$

models the control input at time step $k-1$ and $\mathbf{v}$ is zero-mean Gaussian process noise with covariance R as in Eq. (19). $\dot{x}_{obs}$, $\dot{y}_{obs}$ and $\dot{z}_{obs}$ are the observer's velocities in Cartesian coordinates and ϕ_{obs} the observer's viewing direction in the azimuthal plane. The possibility for the appearance and disappearance of a target is modeled with a birth/death model taking into account the possibility that a partition is born or dies with probabilities p_b and p_d and the possibility of staying alive or dying $1-p_d$, $1-p_b$, respectively.

At each time step only one partition can be born in a particle. But each partition may die independently from other partitions contained in the particle. A newly generated partition is created following a uniform distribution in a measurement volume around the observer.

3) Measurement Model: The binaural localization algorithm provides one bearing measurement from the sensor to a potential sound source per time instance. In order to solve the BOT problem the non-linear measurement equation takes into account bearing measurements in spherical coordinates: $\mathbf{h}(\mathbf{x}_k, \mathbf{n}_k) = \hat{\mathbf{h}}_k(\mathbf{x}_k) + \mathbf{n}(k)$ where $\mathbf{n}(k)$ is assumed to be zero-mean Gaussian noise with covariance matrix $\mathbf{Q}$, and $\hat{\mathbf{h}}_k(\mathbf{x}_k)$ is defined as:

$$\hat{\mathbf{h}}_k(\mathbf{x}_k) = \begin{pmatrix} arctan(y/x) - \phi \\ arccos(z/\sqrt{x^2+y^2+z^2}) \end{pmatrix} \tag{14}$$

Instead of rotating the particles in the propagation step we subtract ϕ from the azimuth in the measurement equation. In our case the observer is able to rotate around the z axis.

Since each particle represents multiple targets (partitions), we have to consider that at each time-step k the active target is n (out of the T possibilities). Therefore, to obtain the measurement likelihoods for the particles we need to marginalize over n, and consider that the measurement could come from clutter as well (so $n = 0$), with probability p_c:

$$\begin{aligned} p(\mathbf{Z} \mid X_p, n) &= \sum_{j=0}^{T} p(n=j)\ p(\mathbf{Z}|X_p, n=j) && (15) \\ &= p_c + \sum_{j=1}^{T} p(n=j)\ p(\mathbf{Z}|x_{p,j}) && (16) \\ &= p_c + \frac{1}{T}\sum_{j=1}^{T} p(\mathbf{Z}|x_{p,j}) && (17) \end{aligned}$$

where we assume a uniform prior for which target is active $(1/T)$, uniform distribution of clutter in the measurement volume V ($p_c = 1/V$), and $p(\mathbf{Z}|x_{p,j}) = \mathcal{N}(\hat{\mathbf{h}}(\mathbf{x}) - \mathbf{Z}; 0, \mathbf{Q})$.

Until this point we worked in the observer's coordinate system, so the tracked states have to be transformed into global positions, through the known control inputs.

III. EXPERIMENTS

We evaluate the proposed method with simulated and real binaural bearing measurements.

A. Distributions

In order to verify the assumptions about measurement noise, reference measurements were taken in a reverberant room without any acoustical treatment. During the measurement both the observer and the target were stationary. The measured reverberation time (RT60) value is 0.693s on average. There were two computers running inside the room. These computers may result into additional noise sources. The observer stood approximately 1 m away from the target at a height of approximately 1.5 m.

The parameters of the multivariate Gaussian distribution of the measurements and the process noise were empirically estimated. Resulting from this estimation, a multivariate zero-mean Gaussian measurement noise with covariance matrix

$$\mathbf{Q} = \begin{pmatrix} 0.981 & 0 \\ 0 & 2.949 \end{pmatrix} \tag{18}$$

and zero mean is used for the simulations and the evaluation of real data. We used a process noise covariance of

$$\mathbf{R} = \begin{pmatrix} 0.0001 & 0 & 0 \\ 0 & 0.0001 & 0 \\ 0 & 0 & 0.0001 \end{pmatrix} \tag{19}$$

B. Simulations

The method is first evaluated in simulations. The target position x_t is $(0,2,0)^T$. According to the dynamics of the state transition in Eq. (10) the observer moves the first 50 time instances with $u = (0.05, 0, 0, 0)$. The observer moves the next 50 time steps with $u = (0, 0.05, 0, 0)$ and the last 100 steps with $u = (-0.05, 0, 0, 0)$. This results in the trajectory shown in Fig. 5.

The green dots represent the observer trajectory, the red dot is the starting point and the red × marks the position of the target. The birth rate p_b and the death rate p_d were set to 0.3, since it produced the best results in the experiments. The performance of a PF is dependent on a good initial distribution. We distributed 10000 particles on a cubic space of $V = 6m \times 6m \times 6m$ centered around the simulated observer. Fig. 3 shows simulated measurements with Gaussian noise with a covariance according to Eq. (18) and 45% uniformly distributed clutter between azimuth range $[-180°, 180°]$ and elevation range $[0°, 180°]$.

The convergence behavior of the particles' partitions is shown exemplarily for the x coordinate of the partition states in Fig. 4. The RMS error of the estimates which are shown in Fig. 5 is 0.344m. One has to keep in mind that the velocity of the observer can be too small in comparison to the process noise $\mathbf{R}$ of the system. This results in a violation

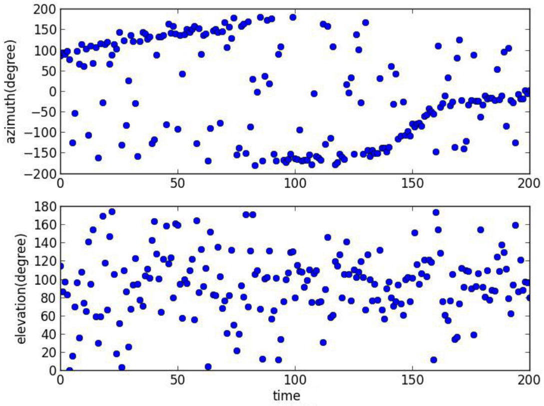

Fig. 3: Simulated measurements according to the trajectory in Fig. 5, with zero-mean Gaussian noise with covariance matrix as in Eq. (18) and 45% uniformly distributed clutter between $[-180°, 180°]$ for azimuth and $[0°, 180°]$ for elevation.

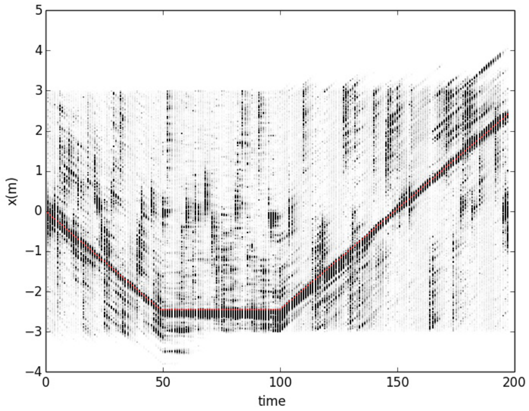

Fig. 4: Visualization of the density of the x coordinate of all particle partitions and new born partitions between ±3m as light grey dots over time resulting from simulation with measurements of Fig. 3. The red line marks the ground truth trajectory.

of the observability criterion for the BOT problem [21]. It causes the process noise to lie in the same magnitude as the movement speed. As a result the PF will not converge. We solve this by applying a threshold in our implementation. The measurement step and the addition of process noise during the propagation step will only take place if the observer moves a certain distance. Another simulation shows that the PF is in principle able to track multiple targets. The first target is at position ${x^{(1)}}_t = (0, 2, 0)^T$ and the second target is at ${x^{(2)}}_t = (2, 3, 0)^T$. The signals from the targets are interleaved, hence every second measurement originates from the same target. In our tests with two targets we simulated bearing measurements that contained no clutter. In the experiments with two targets we found that measurements with clutter require more elaborate measures to solve the permutations that can occur due to the JMPD representation. This will be a topic of future work and here we only show that in theory the JMPD can track two targets in the simulated scenario in Fig. 6.

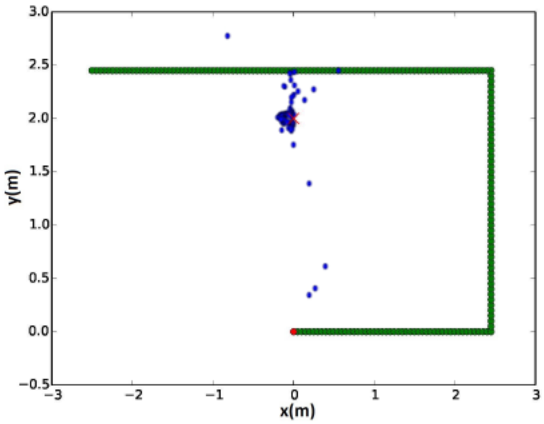

Fig. 5: Estimated positions at each time step for the simulated measurements shown as the projection of the 3D estimates on the x - y plane. The red × marks the real position of the target.

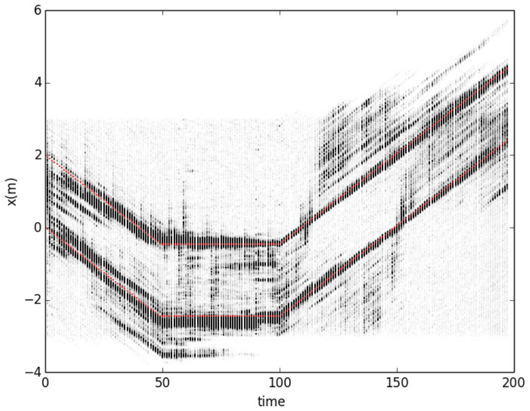

Fig. 6: Visualization of the density on the x coordinate of all particle partitions and new born partitions between ±3m as light grey dots for simulated measurements of two potential targets. The red lines marks the ground truth trajectories.

C. Real Measurements in Reverberant Environment

The observer for the acquisition of real data was a Neumann KU 100 dummy head. The head was moved with a mobile platform. Tracking the position for acquisition of the ground truth data (the position of the observer and of sound sources) was done with a system from Advanced Realtime Tracking GmbH. In the future we plan to use localization data for obtaining the egomotion. We used white noise bursts as a test signal which were played back on coaxial Geithain RL 906 speakers in order to simulate point sound sources in the environment. The sampling rate of the used head related impulse response database was 48 kHz and we chose not to apply resampling to the filters. Consequently, audio was sampled with a rate of 48000 Hz, too. The sample buffer size was 1024 samples with no overlap and the data was windowed with a Hanning window prior to the STFT. The HRTF database [12] we used for binaural localization has a angular resolution of 2°. Measurements were taken in a typical reverberant room. The room was not acoustically treated. The dummy head was fixed to the mobile non-holonomic platform. Fig. 7 shows the trajectory of the observer in the experimental setup.

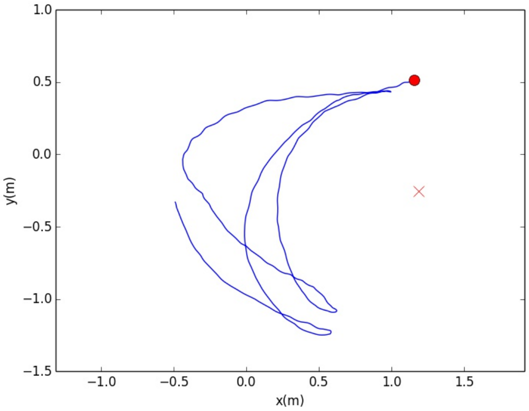

Fig. 7: Motion of the observer and position of the target (red ×) acquired as ground truth with the optical tracking system during the acquisition of real world data.

The difference between the expected measurements based on the ground truth of the positions of observer and target and the actual binaural localization measurements is shown in Fig. 8. As shown in Fig. 9, the particle partitions do not converge well. With a RMS error of 1.764m, the estimates have a large error compared to the distance between target and observer. We reason that the large error is the result of the large amount of reverberations and the influence of clutter.

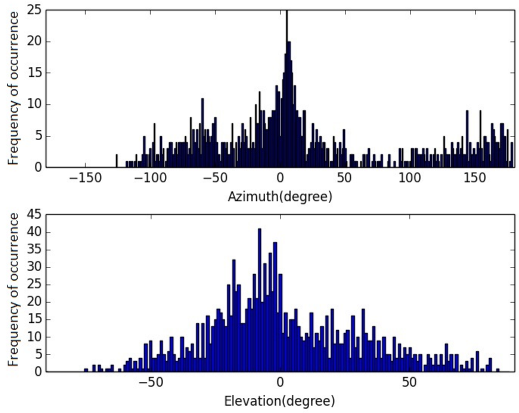

Fig. 8: Deviation of expected measurements based on the positions of observer and target in the ground truth data and the actual bearings resulting from binaural localization with no IC gating applied

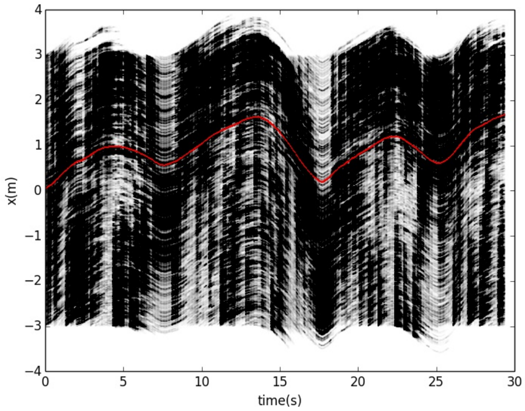

Fig. 9: Visualization of the density of the x coordinate of all particle partitions with real data and observer motion according to Fig. 7 when no IC gating was applied. The red line marks the ground truth.

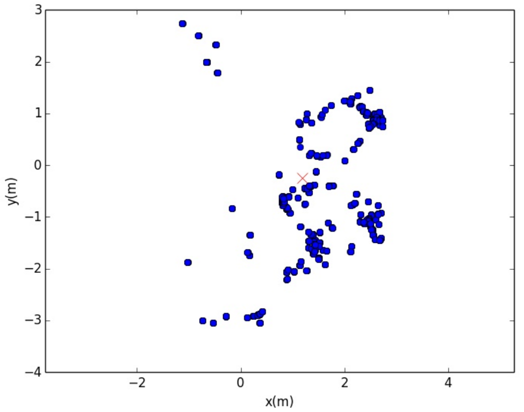

Fig. 10: Estimated positions at each time step for real measurements without IC gating shown as the projection of the 3D estimates on the x - y plane. The red × marks the real position of the target.

In the next experiment we evaluated the proposed method of gating measurements with the IC. The IC threshold for validating a measurement was set to $\alpha = 0.8$ in the localization algorithm. We used a smoothing parameter $\beta = 0.5$. The differences between the angle from the ground truth and the measured angles are shown in Fig. 11. It is clearly evident that the amount of clutter is drastically reduced.

The variance of the measurements is reduced as well. These improvements lead to a good convergence of the PF, as exemplarily shown in Fig. 12. With an RMS error of 0.402m it yields results comparable to the simulations. If only measurements after convergence of the PF are taken into account the RMS error is reduced to approx. 0.157m. The time period necessary for the PF to converge is empirically determined to be approx. 2 seconds. The estimates are shown in Fig. 13. The red dots mark the estimates before convergence.

In order to emphasize the impact of the gating method with the IC, the RMS error of each Cartesian component with and without IC is compared in Table I. It can be seen that the RMS error of the x and y axis for the measurements without IC after convergence is a factor of approx. 15 times larger than that of the measurements with IC. For the RMS

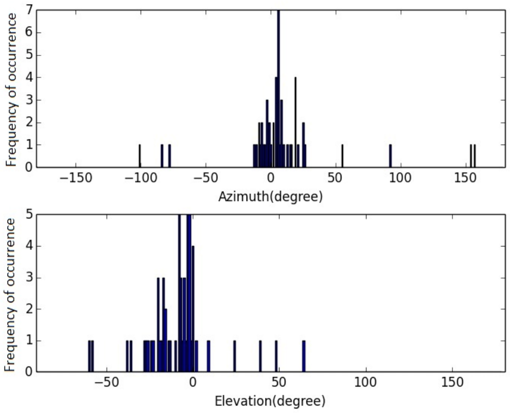

Fig. 11: Deviation of expected measurements based on the positions of observer and target in the ground truth data and the actual bearings resulting from binaural localization with an IC gating threshold of $\alpha = 0.8$.

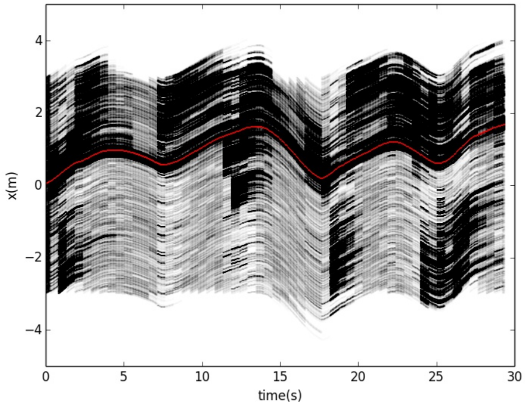

Fig. 12: Visualization of the density of the x coordinate of all particle partitions with real data and observer motion according to Fig. 7 with an IC gating threshold of $\alpha = 0.8$.

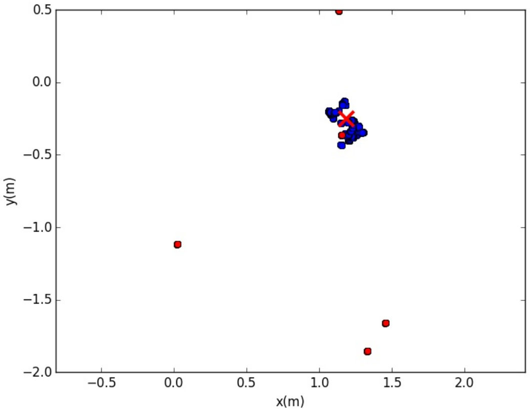

Fig. 13: Estimated positions at each time step for real measurements with an IC gating threshold of $\alpha = 0.8$ shown as the projection of the 3D estimates on the x - y plane. The red × marks the real position of the target. Red: Estimates before convergence, Blue: Estimates after convergence.

	with IC	without IC
$RMSe$ x(m)	**0.171**	1.011
$RMSe$ y(m)	**0.329**	1.372
$RMSe$ z(m)	**0.154**	0.453
$\|RMSe\|$ (m)	**0.402**	1.764

(a) All estimates

	with IC	without IC
$RMSe$ x(m)	**0.064**	1.001
$RMSe$ y(m)	**0.087**	1.329
$RMSe$ z(m)	**0.114**	0.371
$\|RMSe\|$ (m)	**0.157**	1.705

(b) Estimates after convergence

TABLE I: Comparison of the RMSe of the same measurement with and without IC for all estimates (a) and estimates after the convergence (b).

error on the z axis the factor is 3. This can be accounted to the fact that the observer did not move along the z axis. The overall RMS error of estimates after convergence when using the IC results in a 10 times higher accuracy compared to when IC gating is not used.

IV. Conclusions and Outlook

In this work we show first results of applying a PF to the BOT problem in the context of binaural sound localization for robots in real world reverberant environments. We laid out the fundamentals of binaural sound source localization and its application in Sec. II-A and described a gating method based on the precedence effect. We showed the performance of the proposed PF in simulations and evaluated the feasibility of gating real binaural measurements that were acquired in a standard reverberant room with no acoustic treatment with the IC. Additionally, we are able to show that the PF formulation of the JMPD is able to track two targets in simulations.

In real reverberant environment the large amount of clutter and phantom sources due to reflections resulted in estimates with an RMS error of 1.764m. Compared to the distance between target and observer this error is comparably large. Inspired by the precedence effect in human hearing we applied a gating to the localization measurements. With a threshold of the IC of $\alpha = 0.8$ the estimates of real measurements had an RMS error of 0.402m for all estimates. Taking only the estimates after the convergence into account the RMS error was reduced to 0.157m. From these results we conclude that it is possible to binauraly locate a sound source in a reverberant environment with moderate noise taking into account only bearing measurements and knowledge about the observer motion. Compared to the works in [22], our binaural method shows a significant accuracy, specially when gating with the IC is used. When comparing the results we have to take into account that we tracked a stationary sound source with a moving binaural sensor while the authors in [22] show results of tracking moving sources with a stationary and widely spaced microphone array. Therefore, we reason that our results are very promising specially when taking into account the mobility and reduced system complexity.

Since the PF is restricted to stationary targets it could be improved through the use of a more elaborate CV model for the target dynamics in order to track moving targets as well. Furthermore, it should be determined whether the use of modified polar coordinates can yield better results. More tests about the observability should also be conducted in order to improve the convergence time and accuracy of the PF by optimizing the motion of the observer. In future works we will apply the methods taking into account self-localization of a real robot, like for example DLR's Justin, instead of using tracking data for the acquistion of egomotion for the control inputs. Further work needs to be done in order to determine the IC for narrowband signals. Unlike white noise which was used in our experiments natural sounds (e.g. voices) are sparse in the frequency domain. Therefore, not all of the frequency dependent IC bins show a high value for the first wave fronts. This needs to be taken into account in order to gate measurements into the presented PF in arbitrary environments with arbitrary sound sources. Additionally, the IC can also be incorporated in the measurement model of the filter as a measure for the confidence of a particular bearing measurement.

Further investigation will be required for a robust estimation of the locations of multiple sound sources. Due to the possibility of permutations to occur in the partitions' states of the JMPD representation clustering methods are required in order to sort the target states. In our experiments the sound source was in the close vicinity of the binaural dummy head. However, in real scenarios this will not necessarily be the case. Hence, signal processing techniques like for example an auto gain controller need to be applied to the measured raw audio signals in order to amplify or reduce the volume of the signals depending on the distance and volume of the sound sources as well as the conditions in the environment. In summary, we are confident that an acoustic mapping of multiple stationary sound sources is possible. The presented results are promising for a range of applications like for example the localization of victims buried under rubble that are calling for help in disaster recovery scenarios, multimodal data fusion and acoustic mapping for environment modelling in robotics, acoustic sound source detection in industrial scenarios and multimodal interfaces for human robot interaction. We also reason that auditory sound source localization is a promising field of research with a large benefit for perceptional modelling in robotics. There are many unanswered questions for the application of auditory robotic perception outside of laboratory environments, making it an interesting topic for international robotic benchmarks like for example the DARPA challenge. This is specially the case having multimodal environment modeling in disaster recovery scenarios in mind. Our results show that testing under real conditions is crucial for determining the robustness of sound source localization methods and one can not simply rely on results of simulation or on experiments under laboratory conditions.

References

[1] Asano, Futoshi; Suzuki, Yoiti; Sone, Toshio. Role of spectral cues in median plane localization. J. Acoust. Soc. Am., 1990, 88. Jg., Nr. 1, S. 159-168.

[2] Duda, Richard O. Elevation dependence of the interaural transfer function. Binaural and spatial hearing in real and virtual environments, 1997, S. 49-75.

[3] MacDonald, Justin A. A localization algorithm based on head-related transfer functions.J. Acoust. Soc. Am., 2008, 123. Jg., Nr 6, S. 4290-4296.

[4] Wang, DeLiang; Brown, Guy J. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press, 2006.

[5] Faller, Christof; Merimaa, Juha. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence.J. Acoust. Soc. Am., 2004, 116. Jg., Nr. 5, S. 3075-3089.

[6] Valin, Jean-Marc; Michaud, Franois; Rouat, Jean. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robotics and Autonomous Systems, 2007, 55. Jg., Nr. 3, S. 216-228.

[7] Blauert, Jens. Spatial hearing: the psychophysics of human sound localization. MIT press, 1997.

[8] Deleforge, Antoine; Forbes, Florence; Horaud, Radu. Acoustic space learning for sound-source separation and localization on binaural manifolds. International Journal of Neural Systems, 2015, 25. Jg., Nr. 01, S. 1440003.

[9] Deleforge, Antoine; Horaud, Radu. 2D sound-source localization on the binaural manifold. In: Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on. IEEE, 2012. S. 1-6.

[10] Wallach, Hans; Newman, Edwin B.; Rosenzweig, Mark R. A Precedence Effect in Sound Localization. J. Acoust. Soc. Am., 1949, 21. Jg., Nr. 4, S. 468-468.

[11] Litosvky, Ruth Y., et al. The precedence effect. J. Acoust. Soc. Am., 1999, 106. Jg., Nr. 4, S. 1633-1654.

[12] Bernschtz, Benjamin. A spherical far field hrir/hrtf compilation of the neumann ku 100. In: Proc. of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA) Conference on Acoustics. 2013. S. 29.

[13] Rothbucher, Martin, et al. HRTF sound localization. INTECH Open Access Publisher, 2011.

[14] Hue, Carine; Le Cadre, Jean-Pierre; Prez, Patrick. Sequential Monte Carlo methods for multiple target tracking and data fusion. Signal Processing, IEEE Transactions on, 2002, 50. Jg., Nr. 2, S. 309-325.

[15] Kirubarajan, Thiagalingam; Bar-Shalom, Yaakov. Probabilistic data association techniques for target tracking in clutter. Proc. of the IEEE, 2004, 92. Jg., Nr. 3, S. 536-557.

[16] Keyrouz, Fakheredine; Naous, Youssef; Diepold, Klaus. A new method for binaural 3-D localization based on HRTFs. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proc.. 2006 IEEE International Conference on. IEEE, 2006. S. V-V.

[17] Schulz, Dirk, et al. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In: Robotics and Automation, 2001. Proc. 2001 ICRA. IEEE International Conference on. IEEE, 2001. S. 1665-1670.

[18] Kreucher, Chris; Kastella, Keith; Hero III, Alfred O. Multitarget tracking using the joint multitarget probability density. Aerospace and Electronic Systems, IEEE Transactions on, 2005, 41. Jg., Nr. 4, S. 1396-1414.

[19] Smith, Adrian. Sequential Monte Carlo methods in practice. Springer Science and Business Media, 2013.

[20] Evers, Christine et al. Bearing-only Acoustic Tracking of Moving Speakers for Robot Audition. IEEE Intl. Conf. on Digital Signal Processing (DSP), Singapore, July 21-24, 2015 IEEE Intl. Conf. on Digital Signal Processing (DSP), Singapore, July 21-24, 2015

[21] Fogel, Eli; Gavish, Motti. N th-order dynamics target observability from angle measurements. Aerospace and Electronic Systems, IEEE Transactions on, 1988, 24. Jg., Nr. 3, S. 305-308.

[22] Warf, Darren B., et al. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. Speech and Audio Processing, IEEE Transactions on, 2003, 11. Jg., Nr. 6, S. 826-836.

[23] Knapp, Charles H.; Carter, G. Clifford. The generalized correlation method for estimation of time delay. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1976, 24. Jg., Nr. 4, S. 320-327.

[24] Benesty, Jacob. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. J. Acoust. Soc. Am., 2000, 107. Jg., Nr. 1, S. 384-391.