Active Speaker Localization with Circular Likelihoods and Bootstrap Filtering*

Ivan Marković[1], Alban Portello[2], Patrick Danès[3], Ivan Petrović[4], Sylvain Argentieri[5]

***Abstract*— This paper deals with speaker localization in two dimensions from a mobile binaural head. A bootstrap particle filtering scheme is used to perform active localization, i.e. to infer source location by fusing the binaural perception with the sensor motor commands. It relies on an original pseudo-likelihood of the source azimuth which captures both the interaural level and phase differences. Since the pseudo-likelihood is discrete, it is fitted with a mixture of circular distributions in order to enhance its resolution. For the fitting task two mixtures are compared and evalutated, namely the mixture of von Mises and wrapped Cauchy distributions. Furthermore, a solution is presented for calculating the von Mises curvefitting with low uncertainty, since the direct implementation can quickly surpass double precision floating number representation. The performance of the filter is compared using both the raw and fitted pseudo-likelihoods on experiments recorded in an acoustically prepared room with ground-truth obtained from a motion capture system. The results show that the proposed algorithm successfully localizes the speaker with an advantage in the direction of the fitted von Mises mixture likelihood.**

I. INTRODUCTION

In the field of robotics, the subject of sound source localization has been approached and studied from aspects of many different fields, namely speech processing, estimation theory, and sensor fusion to name but a few. From the aspect of sensors, researchers have been using microphone arrays featuring two to more than a hundred of microphones, placing them on wheeled mobile robots, humanoid walking robots, and even autonomous aerial vehicles. Furthermore, when moving sensors are used, the seamless fusion of their motor commands with the binaural perception—active localization—has been acknowledged to overcome ambiguities inherent to the use of static sensors

When considering tracking with bearing-only values, the pertinent problem was tackled foremostly in naval warfare. In [?] it was shown that tracking in modified polar coordinates with an extended Kalman filter (EKF) provided better and more stable results that when tracking in Cartesian coordinates. This brought higher complexity in the motion model, but made the observation model linear and separated observable and unobservable entries in the state vector. This model was further developed in [?] where the tracking was performed with a bank of range-parameterized EKFs in modified polar coordinates. Although this problem has been studied for few decades, it still receives attention due to emerging new filtering methods. In [?] three different filters were compared for the task, while in [?] various methods for tracking and decentralized sensor fusion were studied, including bearing-only scenarios. In [?], relative localization is performed from a pair of moving microphones, based on a multiple-hypothesis square-root unscented Kalman filter. The filtering scheme uses time delays estimated from the sensed audio signals, together with information on the sensor's velocities to perform a consistent source localization. Results show that the strategy, together with a suitable sensor motion, allows to break front-back ambiguity and get accurate range information.

In the context of speaker localization, the bootstrap particle filter has been utilized in [?] for multiple speaker bearing and elevation estimation with an 8-channel microphone array mounted on a mobile robot. In [?], the authors adress the problem of localizing multiple sound sources in an outdoor environment from a microphone array mounted on an arial vehicle. An extension of the MUSIC algorithm is used, that uses adaptive estimation of the—dynamically changing—environment noise correlation matrix. The proposed method is tested with a Parrot AR.Drone and a Kinect device. In [?] the authors used a 4-channel array to localize narrow-band emergency signals from a micro air vehicle, where the sensor model was based on the cross-correlation and doppler shift in frequency due to the motion of the vehicle. In [?] the particle filter was used to estimate the bearing of a speaker from a von Mises (VM) mixture with a 4-channel array mounted on a mobile robot. In a non-robotic related context, in [?] particle filtering was utilized to estimate a position of a speaker in a room environment with 4 microphone pairs placed on the room walls, where the generalized cross-correlation and beamformer output power were used as pseudo-likelihood functions. In [?] the authors analyzed strategies for sensor motion in the context of speaker localization with PF in both range and bearing and performed evaluations in a simulated acoustic environment with single sources under both anechoic and reverberant conditions. In [?] the authors used a combination of direction-of-arrival estimates with speaker's fundamental frequencies (pitch) and gammatone

*This work has been supported by European Community's Seventh Framework Programme under grant agreement no. 285939 (ACROSS) and the BINAAHR (BINaural Active Audition for Humanoid Robots) project funded by ANR (France) and JST (Japan) under Contract nANR-09-BLAN-0370-02.

[1,4]I. Marković and I. Petrović are with University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Control and Computer Engineering, HR-10000 Zagreb, Croatia.

[2,3]A. Portello and P. Danès are with CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France, Univ de Toulouse, UPS, LAAS; F-31400 Toulouse, France.

[5]S. Argentieri is with UPMC Univ. Paris 06, 4 place Jussieu, F-75005, Paris, France and ISIR - CNRS UMR 7222, F-75005, Paris, France. (ivan.markovic, ivan.petrovic) at fer.hr, (alban.portello, patrick.danes) at laas.fr, sylvain.argentieri at upmc.fr

prefiltering to form a pseudo-likelihood function for a 24-channel circular array in order to estimate the bearings of multiple speakers.

In the present paper, active speaker localization is performed with two microphones mounted on a spherical head by bootstrap particle filtering [?]. The underlying state space equation describing the evolution of the source position in the head frame is defined in both cartesian and polar coordinates. We propose a pseudo-likelihood function of the source bearing (azimuth) as the measurement model, which captures both the interaural phase difference (IPD) and interaural level difference (ILD) between the binaural signals. Since the pseudo-likelihood has no analytic expression and is only given for a discrete set of candidate bearings, the fitting of circular distributions to the discrete pseudo-likelihood is discussed in order to enhance its resolution for the purpose of estimation. Incidentally, this can give further ground for possible analytical filtering schemes. Two distributions are presented and compared for the task: namely the VM distribution, for which we also present a method for evaluation with a large concentration parameter, and the wrapped Cauchy (WC) distribution. Furthermore, we compare two bootstrap particle filtering schemes on experimental data—one using the raw discrete pseudo-likelihood, and the other based on the fitted circular distribution. As aforementioned, both fuse the known head velocities with binaural data in order to infer the speaker location.

The paper is organized as follows. First, the problem is stated in § II, while § III presents and compares the proposed fitting with the VM and WC distributions. In § IV the proposed speaker localization with the bootstrap algorithm is presented, § IV-B presents the experimental evaluation, and in the end § V concludes the paper.

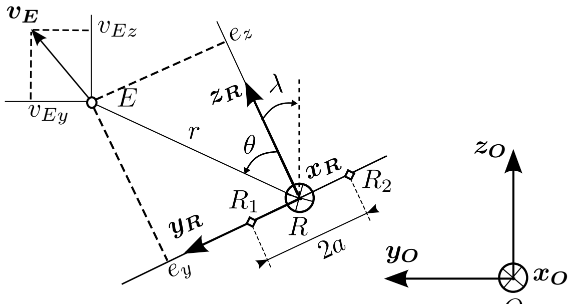

Fig. 1: The considered localization problem.

II. Problem Statement

A. Kinematics and state space equation

A pointwise sound emitter E and a binaural sensor move independently on a common plane parallel to the ground. The two receivers equipping the sensor are denoted by R_1 and R_2. A frame $\mathcal{F}_R : (R, \boldsymbol{x_R}, \boldsymbol{y_R}, \boldsymbol{z_R})$ is rigidly linked to the sensor, with R the midpoint of the line segment $[R_1R_2]$, $\boldsymbol{y_R}$ the vector $\frac{\boldsymbol{RR_1}}{|\boldsymbol{RR_1}|}$ and $\boldsymbol{x_R}$ the downward vertical vector. The frame $\mathcal{F}_E : (E, \boldsymbol{x_O}, \boldsymbol{y_O}, \boldsymbol{z_O})$ attached to the source is parallel to the world reference frame $\mathcal{F}_O : (O, \boldsymbol{x_O}, \boldsymbol{y_O}, \boldsymbol{z_O})$, with $\boldsymbol{x_O} = \boldsymbol{x_R}$ (see Fig. 1). The source undergoes a translational motion (velocities v_{Ey}, v_{Ez} of $\mathcal{F}_E$ w.r.t. $\mathcal{F}_O$ expressed along axes $\boldsymbol{y_O}, \boldsymbol{z_O}$), while the sensor is endowed with two translational and one rotational degrees-of-freedom (velocities v_{Ry}, v_{Rz} of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_O$ expressed along axes $\boldsymbol{y_R}, \boldsymbol{z_R}$; rotation velocity ω of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_O$ around $\boldsymbol{x_O} = \boldsymbol{x_R}$). Assuming v_{Ry}, v_{Rz}, ω are known, the aim is to localize the emitter ($\mathcal{F}_E$) w.r.t. the binaural sensor ($\mathcal{F}_R$) on the basis of the sensed data at R_1, R_2. Importantly, the audio sensor is not localized w.r.t. $\mathcal{F}_O$. The relative attitude of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_E$ can be described, when $v_{Ry}, v_{Rz}, \omega, v_{Ey}, v_{Ez}$ are zero-order held at the sampling period T_s, by the discrete-time deterministic state space equation

$$\boldsymbol{x}_{t+1} = F\boldsymbol{x}_t + G_1\boldsymbol{u_1}_t + G_2(\boldsymbol{x}_t)\boldsymbol{u_2}_t, \text{ with}$$

$$F{=}\begin{bmatrix} \cos(\omega_t T_s) & \sin(\omega_t T_s) & 0 \\ -\sin(\omega_t T_s) & \cos(\omega_t T_s) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; G_1{=}-\begin{bmatrix} \frac{\sin(\omega_t T_s)}{\omega_t} & \frac{1-\cos(\omega_t T_s)}{\omega_t} & 0 \\ \frac{\cos(\omega_t T_s)-1}{\omega_t} & \frac{\sin(\omega_t T_s)}{\omega_t} & 0 \\ 0 & 0 & T_s \end{bmatrix},$$

$$G_2(\boldsymbol{x}_t){=}T_s\begin{bmatrix} \cos(\lambda_t{-}\omega_t T_s) & -\sin(\lambda_t{-}\omega_t T_s) \\ \sin(\lambda_t{-}\omega_t T_s) & \cos(\lambda_t{-}\omega_t T_s) \\ 0 & 0 \end{bmatrix}. \tag{1}$$

Therein, $\boldsymbol{x} \triangleq [e_y, e_z, \lambda]'$, the state vector, gathers the entries $e_y \triangleq \boldsymbol{RE}.\boldsymbol{y_R}$ and $e_z \triangleq \boldsymbol{RE}.\boldsymbol{z_R}$ of $\boldsymbol{RE}$ in $\mathcal{F}_R$, and the angle $\lambda \triangleq \widehat{(\boldsymbol{z_R}, \boldsymbol{z_O})}_{\boldsymbol{x_O}}$. The sensor velocities constituting $\boldsymbol{u_1} \triangleq [v_{Ry}, v_{Rz}, \omega]'$ are supposed known. In the case of a static source—as it is the case in the scope of this paper—$\boldsymbol{u_2} \triangleq [v_{Ey}, v_{Ez}]'$ is simply set to zero. When parametrizing the problem in terms of polar coordinates rather than cartesian, *i.e.* when using the variables $\theta \triangleq \operatorname{atan2}(e_y, e_z)$, $r \triangleq \sqrt{e_y^2 + e_z^2}$, the state space equation comes as

$$r_{t+1}{=}\sqrt{r_t^2{+}\boldsymbol{u}_t'G'G\boldsymbol{u}_t{+}2r_t[\sin\theta_t,\cos\theta_t]G'\boldsymbol{u}_t} \tag{2}$$
$$\theta_{t+1}{=}\operatorname{atan2}(r_t\sin(\theta_t{+}\omega_t T_s){+}\boldsymbol{g_1}\boldsymbol{u}_t, r_t\cos(\theta_t{+}\omega_t T_s){+}\boldsymbol{g_2}\boldsymbol{u}_t)$$
$$\lambda_{t+1}{=}\lambda_t{-}\omega_t T_s,$$

with $\boldsymbol{u} \triangleq [v_{R_y}, v_{R_z}]'$, G the square matrix made up with the first two rows and columns of G_1, $\boldsymbol{g_1}$ (resp. $\boldsymbol{g_2}$) the first (resp. second) row of G. To model uncertainty in the relative motion, a random white Gaussian noise of known statistics is added to (2).

B. Acoustic model, measurement vector, pseudo-likelihood

Consider first a static world where the sensor is motionless. We assume that the source lies in the farfield (*i.e.* the source range $r = |\boldsymbol{RE}|$ is sufficiently high compared to the microphones interspace $2a$ so that the source wavefronts can be considered as planar in the vicinity of the microphone pair). We model the signals y_1, y_2 monitored at R_1, R_2 in the presence of additive noise as follows

$$\begin{cases} y_1(\tau) = s(\tau) + n_1(\tau) \\ y_2(\tau) = (s * h_\theta)(\tau) + n_2(\tau), \end{cases} \tag{3}$$

where the signal s (*i.e.* the contribution of the emitter at R_1) and the noises n_1, n_2 are real, band-limited, individually and jointly stationary random processes, and $*$ denotes convolution. The—deterministic—impulse response h_θ between R_1, R_2, is parameterized by θ, and captures free-field propagation of the emitted signal as well as head scattering.

H_θ, the Fourier transform of h_θ, is supposed known for every θ within a discrete set of values (say, it has been learnt from calibration, or is known theoretically). The process $\boldsymbol{y}(\tau) \triangleq [y_1(\tau), y_2(\tau)]'$ is observed over N adjacent non-overlapping rectangular T/N-width time windows. Denote $\boldsymbol{y}_n$ the observation of $\boldsymbol{y}$ over the n^{th} window. A data vector $\boldsymbol{Z}$ is made up by stacking the values of

$$\boldsymbol{Y}_n[k] = \sqrt{\frac{N}{T}} \int_{\mathbb{R}} \boldsymbol{y}_n(\tau) e^{-2i\pi k \frac{N}{T} \tau} d\tau, \; n = 1, ..., N \tag{4}$$

at $k = k_1, ..., k_B$, the B frequency indexes within the bandwidth of s. $\boldsymbol{Z}$ is hence defined as $\boldsymbol{Z} \triangleq [\boldsymbol{Y}[k_1]', ..., \boldsymbol{Y}[k_B]']'$, with $\boldsymbol{Y}[k] \triangleq [\boldsymbol{Y}_1[k]', ..., \boldsymbol{Y}_N[k]']'$. Assume now that s, n_1, n_2 are zero-mean jointly Gaussian and that n_1, n_2 are identically distributed, uncorrelated with each other and with s. Then, under general mild conditions on the power spectra of s, n_1, n_2 and on H_θ, the maximum likelihood estimate of θ can be obtained, given a sample $\boldsymbol{z}$ of $\boldsymbol{Z}$, by maximizing the following criterion [**?**], hereafter referred to as the "pseudo log-likelihood function"

$$J(\boldsymbol{z}|\theta) = c_2 - N \sum_{k=k_1}^{k_B} \Big(\ln |P_{\boldsymbol{\theta}}[k] \hat{C}[k] P_{\boldsymbol{\theta}}[k] + \hat{\sigma}^2_{\boldsymbol{\theta}}[k] P^{\perp}_{\boldsymbol{\theta}}[k]| \Big), \tag{5}$$

with $c_2 \triangleq -2NB(\ln(\pi)+1), \hat{C}[k] \triangleq \frac{1}{N} \sum_n \boldsymbol{y}_n[k] \boldsymbol{y}_n[k]^\dagger$,
$P_{\boldsymbol{\theta}}[k] \triangleq \boldsymbol{V}_{\boldsymbol{\theta}}[k] (\boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger \boldsymbol{V}_{\boldsymbol{\theta}}[k])^{-1} \boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger, P^{\perp}_{\boldsymbol{\theta}}[k] \triangleq \mathbb{I}_2 - P_{\boldsymbol{\theta}}[k]$,
$\boldsymbol{V}_{\boldsymbol{\theta}}[k] \triangleq [1, H_{\boldsymbol{\theta}}[k]]', \hat{\sigma}^2_{\boldsymbol{\theta}}[k] \triangleq \operatorname{tr}(P^{\perp}_{\boldsymbol{\theta}}[k] \hat{C}[k])$.

Therein, †, $|.|$, $\operatorname{tr}(.)$ stand respectively for the Hermitian transpose, determinant and trace operators, $\boldsymbol{y}_n[k]$ denotes a sample of $\boldsymbol{Y}_n[k]$, and the sample covariance matrix $\hat{C}$ is assumed full rank.

Consider now a real world where the sensor moves and where the source signal and environment noise are possibly nonstationary. All the fundamental hypotheses leading to (3)–(4)–(5) are consequently violated. Nevertheless, the problem can still be handled if, at each process time index t, the data vector $\boldsymbol{z}_t$ is made up from audio data collected over a time window matched to t, sufficiently short so that, along this window, the sensor motion is negligible and the recorded signals can be regarded as finite-time samples of stationary processes. Hence, at each time index t, the pseudo likelihood of θ_t w.r.t. $\boldsymbol{z}_t$, denoted $p(\boldsymbol{z}_t|\theta_t)$, can be output and will henceforth be used in a Bayesian filtering scheme in §IV. Importantly, $p(\boldsymbol{z}_t|\theta_t)$ has in the general case no analytic expression. Its numerical values are just given for a discrete set of tested azimuths. This precludes the use of Bayesian filtering schemes requiring an analytic form of the pseudo likelihood, *e.g.* Gaussian mixture filters, unless an analytic function is fitted to the discrete values. Alternatively, with particle filters, the pseudo likelihood in its discrete form can be utilized as a sensor model. However, low azimuth resolution can affect the particle filter performance and consistency, and it may be useful to fit some distribution to the discrete pseudo likelihood. §III is thus dedicated to the fitting of Von Mises and wrapped Cauchy mixtures models to the discrete pseudo likelihood.

III. Fitting Circular Distributions to the Pseudo Likelihood Function

A. *Circular distributions*

In this section we present two solutions to fitting the pseudo likelihood function, namely fitting with the vM distribution and with the wC distribution. The motivation behind using circular distributions lies in the fact that they intrinsically take noneucledian properties of the angular data into account. For an example, this property proves useful in the optimization since a circular distribution close to π continues contributing to points larger than $-\pi$. Furthermore, in the present paper we do not require the component weights to sum up to one, since the pseudo likelihood function itself is not a valid probability distribution.

A probability distribution on the unit circle with density function given by [**?**]

$$p(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left\{\kappa \cos(\theta - \mu)\right\}, \tag{6}$$

is called the von Mises distribution, where $0 \leq x \leq 2\pi$, μ is the mean direction, $\kappa \geq 0$ is the concentration parameter, and $I_0(\kappa)$ is the modified Bessel function of the first kind and of order zero. The distribution is unimodal and symmetric around the μ and is often referred to as the circular analogue of the Gaussian distribution. When $\kappa \to 0$ the vM becomes the uniform distribution, while if $\kappa \to \infty$ it becomes concentrated at $\theta = \mu$.

We used the vM distribution in the context of robot audition in [**?**] where the sensor model was represented as a mixture of vM distribution in particle filtering, while in [**?**] we extended this approach to model the entire Bayesian tracking procedure in the analytical domain of the distribution. However, both of the aforementioned works were only concerned with tracking the bearing value of the speaker and not its position in two dimensions which is one of the goals of the present paper.

The second distribution that we analyze for the task is a distribution wrapped on a circle. Given a distribution on the line we can wrap it around the circumference of a circle with unit radius. If a random variable θ is defined on a line, then the corresponding random variable of the wrapped distribution is $\theta_w = \theta(\operatorname{mod} 2\pi)$. Furthermore, if θ has a pdf p, then the corresponding wrapped pdf p_w is defined as [**?**]

$$p_w(\theta) = \sum_{k=-\infty}^{k=\infty} p(\theta + 2k\pi). \tag{7}$$

From (7) we can note practical issues when dealing with the infinite number of terms in the summation. However, it can be shown that the Cauchy distribution on the line has an interesting property that its wrapped counterpart, due to certain geometric series expansion property, reduces to [**?**]

$$p(\theta; \mu, \rho) = \frac{1}{2\pi} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos(\theta - \mu)}, \tag{8}$$

where μ is the mean direction and ρ is called the mean resultant length. When $\rho \to 0$ the wC tends to uniform distribution, while if $\rho \to \infty$ the distributions becomes concentrated at point μ.

Naturally, the pseudo likelihood function will suffer from front-back ambiguity since in the present paper we utilize a binaural setup. Hence, our sensor model will contain at least two distinct modes on the interval 0 to 2π and for this reason we chose to model the likelihood as a mixture of distributions. If we denote with $\mathcal{X}$ a set of distributions parameters, then an N component mixture can be defined as

$$p(\theta; \mathcal{X}) = \sum_{i=1}^{N} \omega_i p(\theta; \mathcal{X}_i), \tag{9}$$

where the set $\mathcal{X}$ consists of $\cup_i \{\mu_i, \kappa_i\}$ for the vM distribution and $\cup_i \{\mu_i, \rho_i\}$ for the wC distribution.

B. Computation of the von Mises distribution with large concentration parameters

The direct form of the vM distribution suffers from numerical issues when working with large concentration parameter κ, i.e. with sharp distributions which may be necessary to fit the pseudo likelihood in the vicinity of its local modes. The main problem is that for large κ both the exponent and the modified Bessel function of the first kind quickly reach the maximum value that can be stored in double precision floating point representation.

To solve this problem, we move the normalizer of the vM distribution in the exponent as follows

$$p(\theta; \mu, \kappa) = \exp\left\{\kappa \cos(\theta - \mu) - \log(2\pi I_0(\kappa))\right\}, \tag{10}$$

and approximate $\log(I_0(\kappa))$ as [**?**]

$$\begin{aligned} \log(I_0(x)) &= \log \sum_{k=0}^{\infty} \frac{\exp\{m(x)\}}{\exp\{m(x)\}} \exp\left\{t_k(x)\right\} \\ &= m(x) + \sum_{k=0}^{\infty} \exp\left\{t_k(x) - m(x)\right\}, \end{aligned} \tag{11}$$

where $t_k(x) = 2k \log \frac{x}{2} - 2\sum_{r=0}^{k} \log r$ and $m(x) = \max\{t_k(x)\}$. The number of the terms in (11) required to have an accurate approximation depends on the κ. For the present application we have found that the maximal practical value of the concentration parameter for fitting the pseudo likelihood is $\kappa = 2000$, for which an accurate approximation was empirically determined to be for $k \leq 1100$. But for smaller parameters, e.g. $\kappa = 1000$, the number of terms $k \leq 600$ was sufficient. We did not notice any increase in the computational time when compared to Matlab's implementation based on [**?**].

C. Evaluation of the fitting performance

The fitting of a mixture of distributions to the pseudo likelihood function $\hat{p}(\theta)$ comes down to solving the following optimization problem

$$\begin{aligned} \underset{\omega, \mathcal{X}}{\text{minimize}} \quad & \left(\sum_{i=1}^{N} \omega_i p(\theta; \mathcal{X}_i) - \hat{p}(\theta) \right)^2 \\ \text{s.t.} \quad & 0 \leq \omega_i \leq 1, \; i = 1, \dots, N \\ & 0 \leq \mathcal{X}_i \leq \mathcal{B}, \; i = 1, \dots, N, \end{aligned}$$

where the upper bound $\mathcal{B}$ depends on the parameter and the distribution. For both distributions the upper bound of the mean directions μ is $\mathcal{B} = 2\pi$, while for the vM distribution the upper bound was $\mathcal{B} = 2000$ for the concentration parameter, and for the wC distribution $\mathcal{B} = 1$ for the mean resultant length.

Fig. 2: Fitting the pseudo likelihood for a single frame with vM and wC mixture

Concerning the number of the mixture components all the results were obtained with $N = 4$. Initial conditions for the mean directions were determined by searching recursively for N most dominant peaks in the vein of [**?**] where the authors searched for the number of active speakers. Once the dominant peak is found, an area around it is removed and the search continues until the predetermined number of modes is found. Since in the pseudo likelihood function we expect two peaks to be dominant we set the initial conditions for the first two dominant peaks to be $\kappa = 1500$ or $\rho = 0.9$, while for the rest we set $\kappa = 10$ or $\rho = 0.1$. The weights are initially set to $\omega_i = 0.5, \forall i$.

In Fig. 2 we can see the result of fitting for a single relatively difficult frame when the speaker was close to the end-fire position of the array and the two dominant modes started overlapping. Empirically we noticed that this is the more difficult case for the wC distribution and that often the two distinct nodes tend to be fitted with a single component in between them. Overall, the whole dataset consisted of four experiments with a talking speaker as the source. The average root-mean-square-error (RMSE) of fitting for the speaker scenario was $1.6 \cdot 10^{-3}$ for the vM mixture and $3.7 \cdot 10^{-3}$ for the wC mixture, respectively. Given that, for the rest of the paper we have chosen to work with the vM mixture since it provided better fitting in terms of the average RMSE.

IV. Speaker Localization in 2D

A. Particle filtering

Particle filtering is a versatile method to recursive Bayesian state estimation. It can handle nonlinear prior dynamics and measurements models, as well as nonGaussian noises. The posterior probability density function (pdf) of the state at any time t conditioned on the sequence of observed measurements up to t is estimated by means of a point-mass probability distribution with stochastic support, or "weighted particle set". Let $\{\boldsymbol{x}^p, w^p\}_{p=1}^{P}$ depict the random measure that characterizes the posterior state pdf $p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t})$, where each particle in the set $\{\boldsymbol{x}^p, \; p = 1, \dots, P\}$ is associated to the respective weight in $\{w^p,\ p=1,\dots,P\}$. The weights satisfy $\sum_p w^p = 1$, so that $p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t})$ can be approximated as [**?**], [**?**]

$$p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t}) \approx \sum_{p=1}^{P} w_t^p \delta(\boldsymbol{x}_t - \boldsymbol{x}_t^p), \tag{12}$$

with $\delta(.)$ the Dirac delta measure. In other words, sampling from $p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t})$ returns to sampling a particle with a probability equal to its associated weight.

The particles are drawn according to a so-called importance function, then weighted so that the consequent random measure constitutes a sound approximation to the posterior pdf. As, for any recursive particle filter, the significant weights tend to concentrate on a limited set of particles after few iterations, a resampling step is inserted, which consists in turning $\{\boldsymbol{x}_t^p, w_t^p\}_{p=1}^P$ into the equivalent evenly weighted set $\{\boldsymbol{x}_t'^p, \frac{1}{P}\}_{p=1}^P$ by independently sampling (with replacement) $\boldsymbol{x}_t'^p$ according to $P(\boldsymbol{x}_t'^p = \boldsymbol{x}_t^p) = w_t^p$.

In the sequential importance resampling (SIR) scheme [**?**], or bootstrap filter, the importance function matches the prior dynamics $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, calculated via (2), i.e. each particle $\boldsymbol{x}_t^p$ at time t is drawn from its predecessor $\boldsymbol{x}_{t-1}^p$ at time $t-1$ according to the proposal density $\boldsymbol{x}_t^p \sim p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}^p)$. Then, its weight is updated by evaluating its likelihood $p(\boldsymbol{z}_t|\boldsymbol{x}_t^p)$ prior to setting

$$w_t^p \propto w_{t-1}^p p(\boldsymbol{z}_t|\boldsymbol{x}_t^p), \tag{13}$$

where $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ represents the sensor model, i.e. the fitted vM mixture:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) = \sum_{i=1}^{N} \omega_i \frac{1}{2\pi I_0(\kappa_i)} \exp\left[\kappa_i \cos(\boldsymbol{x}_t - \boldsymbol{z}_{t,i})\right]. \tag{14}$$

Then, all the particle weights are normalized so that they sum up to unity.

Once the random measure approximating the posterior pdf of the state is computed, the posterior mean and posterior covariance can be estimated via

$$\hat{\boldsymbol{x}}_t = E[\boldsymbol{x}_t|\boldsymbol{z}_{1:t}] \approx \sum_{p=1}^{P} w_t^p \boldsymbol{x}_t^p, \tag{15}$$

and

$$\begin{aligned}\hat{\mathbf{P}}_t &= E[(\boldsymbol{x}_t - E[\boldsymbol{x}_t|\boldsymbol{z}_{1:t}])(\boldsymbol{x}_t - E[\boldsymbol{x}_t|\boldsymbol{z}_{1:t}])^{\mathrm{T}}|\boldsymbol{z}_{1:t}] \\ &\approx \sum_{p=1}^{P} w_t^p (\boldsymbol{x}_t^p - \hat{\boldsymbol{x}}_t)(\boldsymbol{x}_t^p - \hat{\boldsymbol{x}}_t)^{\mathrm{T}}.\end{aligned} \tag{16}$$

To avoid a loss of diversity in the particle cloud, the resampling step is applied only when the number of effective weights $P_{\text{eff}} = 1/\sum_p (w^p)^2$ is less than a given threshold, e.g. 33 % of the total number of particles P.

Consequently, particle filtering can be implemented even if a closed-form measurement model is not available, in that the particle likelihoods just need to be evaluated. In our case, the sensor model comes as the pseudo likelihood digitized with a resolution of 4°. However, we assert that the fitting utilized in the present paper constitutes a form of interpolation which yields better resolution. So, we henceforth compare the performance of the bootstrap particle filter which directly utilizes the discrete pseudo likelihood against the particle filter utilizing the fitted vM mixture. Importantly, fitting with a vM mixture would be a prerequisite if the tracking was performed in the vein of [**?**].

B. Experimental results

Experiments were conducted in an acoustically prepared room, equipped with 3D pyramidal pattern studio foams placed on the roof and on the walls. Two surface microphones were mounted at the antipodes of a 8.9 cm radius plastic rigid sphere, itself place on a tripod. The two microphones outputs were synchronously acquired at 44.1 kHz. The sphere tripod was moved manually with a wheeled cart while the source, a loudspeaker placed at the same height as the microphones, was emitting various types of signals. The true source and sensor positions were acquired at 200 Hz with a motion capture system, providing a less than 1mm position error. For that purpose, small infrared active markers were placed on the sphere and the loudspeaker, and their signals were beamed to three infrared camera units placed at the corners of the room. The experimental setup is depicted in Fig. 3.

For the considered case of a rigid sphere, H_θ is shown to have the following analytic expression [**?**]

$$H_\theta(f) = \frac{\psi_{\frac{\pi}{2}+\theta}(f)}{\psi_{-\frac{\pi}{2}-\theta}(f)}, \text{ with} \tag{17}$$

$$\psi_\alpha(f) \triangleq \frac{1}{\left(\frac{2\pi f a}{c}\right)^2} \sum_{m=1}^{\infty} \frac{(-i)^{m-1}(2m+1)P_m(\cos\alpha)}{h_m'\left(\frac{2\pi f a}{c}\right)}.$$

Therein, ψ_β is the normalized Head Related Transfer Function (HRTF) to the microphone at angle β—with respect to boresight—on the sphere, where α stands for the angle between the source bearing and the direction to the considered microphone, P_m is the Legendre polynomial of degree m, h_m is the mth-order spherical Hankel function and h_m' is its first derivative. This expression was thus used in the pseudo likelihood computation. In practice, the infinite sum

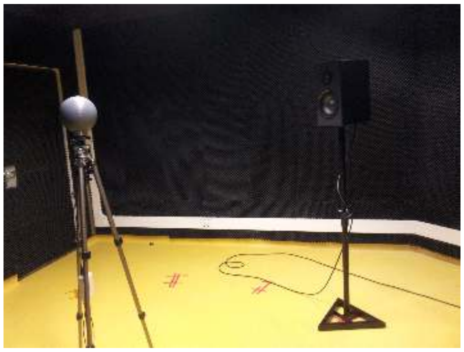

Fig. 3: Experimental setup: plastic sphere and speaker tripods in the acoustic room. Infrared cameras were measuring the ground-true positions.

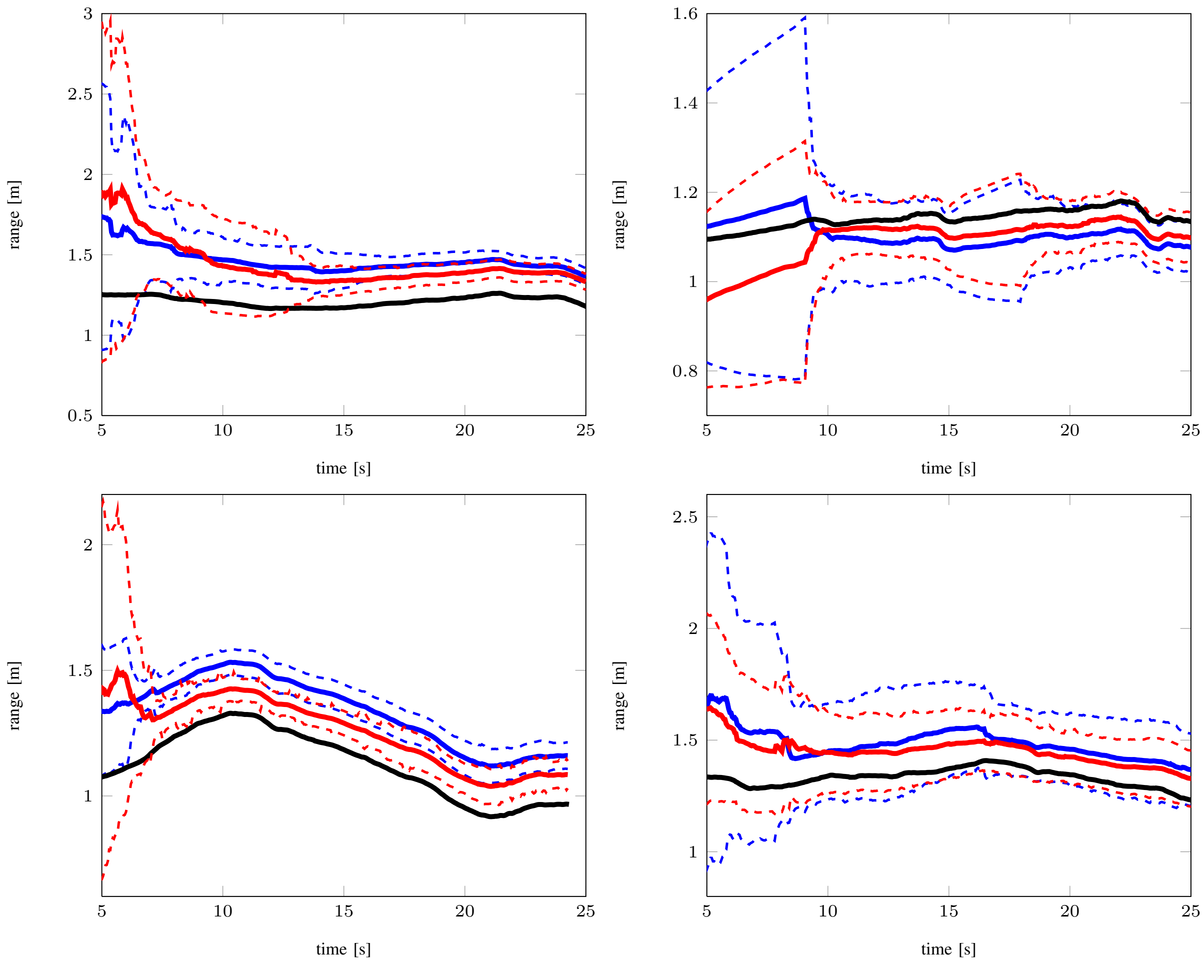

Fig. 4: Mean value of range estimates and pertaining three standard deviations of 50 Monte-Carlo runs of the PF with pseudo likelihood (blue), VM fitted pseudo likelihood (red) and true range (black)

in (18) is approximated by a finite sum, the minimum order required to make the approximation reasonable depending on the maximum frequency considered. To avoid cumbersome computation during localization, H_θ was precomputed and stored offline for a discrete set of bearings.

In order to assess the performance of the PFs, we ran 50 Monte-Carlo runs on the sensed binaural data using either the discrete pseudo likelihood or the VM fitted pseudo likelihood. The runs were performed on four scenarios with different trajectories of the sensor, out of which one scenario included an intermittent sound source. In Fig. 4 we can see the results of range estimation for the four cases, while Fig. 5 shows the estimation of the bearing. By analyzing the results we can see that on average the PF with the VM fitted likelihood gave smaller error in terms of the range estimation although the performance in the bearing was similar for both PFs. The explanation lies in the fact that estimating the range from bearing-only measurements benefited from having an analytical likelihood compared to the $4°$ resolution of the discrete pseudo likelihood.

Then, for each entry of the posterior mean output by the filter, a minimum-width confidence interval was then drawn (from the posterior covariance matrix ouput by the filter) which should enclose the corresponding entry of the genuine hidden state vector with 99% probability. By analyzing the obtained plots concerning the range estimation error, we can see that the present implementation of the PF was not consistent over all the runs, since the true range is outside of the filter's $\pm 3\sigma$ interval calculated from the estimated covariance matrix.

V. CONCLUSION

In the present paper we have studied and proposed a solution for the problem of active speaker localization with a head mounted binaural microphone sensor. The solution was based on calculating a discrete pseudo likelihood function in speaker bearing based on the geometrical properties of the spherical head. The resulting likelihood was fitted with a mixture of circular distributions, namely the VM and wrapped Cauchy distributions, whose comparison showed better results in the case of the VM distribution. A bootstrap algorithm was utilized with the direct and VM fitted pseudo