MULTIPLE SPEAKER TRACKING WITH THE FACTORIAL VON MISES-FISHER FILTER

Johannes Traa[♭] *Paris Smaragdis*[♯]

[♭] Department of Electrical and Computer Engineering, UIUC

[♯] Departments of Electrical and Computer Engineering and Computer Science, UIUC; Adobe Systems, Inc.

ABSTRACT

Multiple-target tracking with a microphone array is often addressed via the Bayesian filtering framework. For compact arrays, each source is represented by its direction-of-arrival (DOA), which evolves on the unit sphere. The unique topology of this space leads to analytical intractabilities that are often resolved via costly particle-based methods.

In this paper, we derive a novel, deterministic inference algorithm called the von Mises-Fisher Filter (vMFF) for a dynamical system model defined on the sphere, and extend it to the multi-source scenario in the Factorial vMFF (FvMFF). We apply sensor fusion and probabilistic data association techniques to handle clutter and data association ambiguities in the observation set. We show that the vMFF combines the computational efficiency of a Kalman filter with the tracking accuracy of a particle filter to perform well across all noise levels. Finally, we apply the FvMFF to track multiple speakers in a reverberant environment.

Index Terms— von Mises-Fisher, speaker tracking, bayesian filtering

1. INTRODUCTION

Tracking one or more sound sources in a reverberant environment is a challenging task that finds applications in many areas [1]. It is often useful for source separation and speech enhancement algorithms that require on-line directionality information. Tracking algorithms are often based on the Kalman Filter (KF) [2] or the non-linear/non-Gaussian variants of it such as the extended KF [3], the unscented KF [4], and the particle filter [5]. For small arrays with inter-element distances of 1-5 cm, it is more meaningful to track the direction-of-arrival (DOA) of each source rather than its 3D position. This is a quantity that lies on the surface of the unit circle or sphere, depending on the array configuration. Thus, it is beneficial to develop methods that are tailored to the unique statistics of such spaces [6, 7, 8, 9, 10].

We address the problem of tracking the DOAs of multiple sources with a compact, 4-microphone array in a noisy, reverberant environment. We will approach the problem from a generative model perspective in which uncertainty in the DOA is expressed with the von Mises-Fisher (vMF) [11] distribution. Mixtures of vMFs were studied in [7, 12, 13] for clustering high-dimensional directional datasets. A vMF-based particle filter was proposed for tracking white matter fibers in [14] and was adapted for tracking speakers in [15]. However, we would like to avoid using a particle-based representation because it (1) can be computationally demanding and (2) complicates the inference procedure in the case of multiple sources and measurements.

We introduce a spherical dynamical system (SDS) model based on that of [14] that describes the evolution of a source DOA over time. Uncertainty in the source position and the observation is modeled with the vMF distribution and uncertainty in the source's rotation velocity is modeled with a Normal distribution. We derive an efficient, particle-free inference procedure for the SDS called the vMF Filter (vMFF). This is extended to the setting where K sources and $M \gg K$ observations (many of which may be clutter) are present. This is the case when using interchannel time delay (ITD) [16] features extracted from the recorded signals. Sensor fusion [17] and probabilistic data association (PDA) [18, 19] techniques are applied to derive a Factorial vMFF (FvMFF) for tracking multiple sources in the presence of multiple observations.

2. BACKGROUND

2.1. von Mises-Fisher Distribution

The unit sphere is the manifold containing all unit vectors:

$$\mathbb{S}^2 = \left\{ \mathbf{x} : \mathbf{x} \in \mathbb{R}^3, \|\mathbf{x}\|_2 = 1 \right\} \quad . \tag{1}$$

Several probability distributions exist for modeling random vectors on $\mathbb{S}^2$ such as the Bingham [20], Fisher-Bingham (aka Kent) [21], and von Mises-Fisher (vMF) [11] distributions. We will use the vMF for its simplicity and analytical tractability in deriving the vMFF. It is parameterized by mean $\boldsymbol{\mu}$ and concentration κ and has density function:

$$p\left(\mathbf{x}\,;\,\boldsymbol{\mu},\kappa\right) = \frac{\kappa}{4\pi \sinh\left(\kappa\right)} e^{\kappa \mathbf{x}^\top \boldsymbol{\mu}} \quad . \tag{2}$$

The vMF is derived by conditioning a Gaussian random variable $\mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}\right), \|\boldsymbol{\mu}\|_2 = 1$ on the unit sphere. The transformation is such that $\kappa = 1/\sigma^2$, and so κ behaves like an inverse variance. We can visualize the vMF on $\mathbb{S}^2$ as shown in Fig. 1. We will use it to model DOA information.

2.2. Rotations on the unit sphere

Rotations on the unit sphere can be described with quaternions [20] or, equivalently, an angle-axis representation. We will model the velocity of a source directly on $\mathbb{S}^2$ with a Gaussian random variable $\mathbf{r} \in \mathbb{R}^3$ such that $\|\mathbf{r}\|_2$ defines the amount of rotation in radians and $\mathbf{r}/\|\mathbf{r}\|_2 \in \mathbb{S}^2$ defines the axis of rotation. One can rotate a vector by ν radians about an axis $\mathbf{a} \in \mathbb{S}^2$ by pre-multiplying it with:

$$\mathbf{R}(\mathbf{a},\nu) = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix} \sin(\nu) + (\mathbf{I} - \mathbf{a}\mathbf{a}^\top)\cos(\nu) + \mathbf{a}\mathbf{a}^\top. \tag{3}$$

For a rotation vector $\mathbf{r}$, we write $\mathbf{R}\left(\mathbf{r}\right)$ as a shorthand for $\mathbf{R}\left(\mathbf{r}/\|\mathbf{r}\|_2, \|\mathbf{r}\|_2\right)$. With this representation, we can track the source's rotation $\mathbf{r}$ via the Kalman filter.

978-1-4799-3694-6/14/$31.00

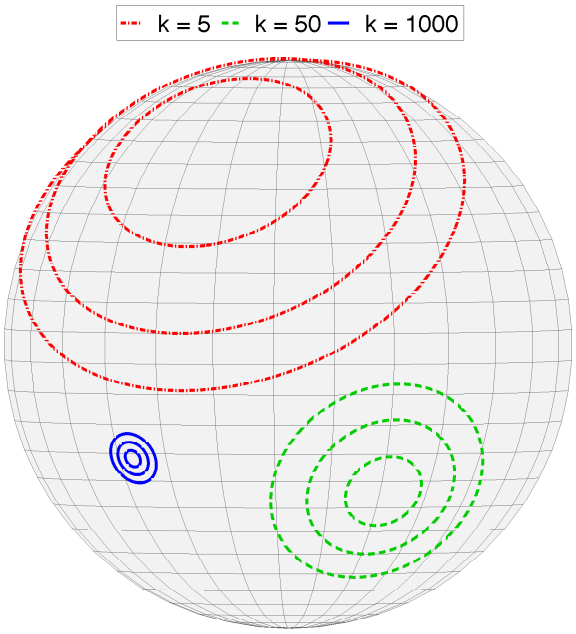

Fig. 1. Contours of the von Mises-Fisher distribution on the unit sphere for three values of the concentration parameter κ.

3. SPHERICAL DYNAMICAL SYSTEM (SDS)

We define a generative model for the SDS in analogy with the standard linear dynamical system (LDS). It can be written as:

$$\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{r}_{t-1} \sim vMF\left(\mathbf{R}\left(\mathbf{r}_{t-1}\right)\mathbf{x}_{t-1},\kappa_{\mathbf{x}}\right) \quad , \tag{4}$$

$$\mathbf{r}_t|\mathbf{r}_{t-1} \sim \mathcal{N}\left(\mathbf{A}\mathbf{r}_{t-1},\mathbf{\Sigma}_{\mathbf{r}}\right) \quad , \tag{5}$$

$$\mathbf{y}_t|\mathbf{x}_t \sim vMF\left(\mathbf{x}_t,\kappa_{\mathbf{y}}\right) \quad , \tag{6}$$

where $\mathbf{x}_t \in \mathbb{S}^2$, $\mathbf{r}_t \in \mathbb{R}^3$, and $\mathbf{y}_t \in \mathbb{S}^2$ denote the source position, source rotation, and observation. Equations (4)-(5) describe the evolution of a speaker's position and rotation directly on the unit sphere. We denote the complete state vector as $\mathbf{s}_t = \begin{bmatrix}\mathbf{x}_t^\top & \mathbf{r}_t^\top\end{bmatrix}^\top$.

One approach to handling the statistics of the SDS uses a particle-based representation [14, 15]. To avoid computational burden, we will make approximations in the Bayesian filtering equations to derive a simple and efficient tracking algorithm called the von Mises-Fisher Filter (vMFF).[1] We extend this to the multi-source, multi-observation case, yielding the Factorial vMFF (FvMFF).

4. APPROXIMATE INFERENCE FOR THE SPHERICAL DYNAMICAL SYSTEM

We derive a deterministic inference algorithm for the SDS by approximating the corresponding Bayesian filtering equations. We will assume that the filtered state distribution at time $t-1$ factors as:

$$\begin{aligned} p\left(\mathbf{s}_{t-1}|\mathbf{y}_{1:t-1}\right) &= p\left(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}\right)p\left(\mathbf{r}_{t-1}|\mathbf{y}_{1:t-1}\right) \\ &= vMF\left(\mathbf{x}_{t-1}\,;\,\widehat{\boldsymbol{\mu}}_{t-1},\widehat{\kappa}_{t-1}\right)\mathcal{N}\left(\mathbf{r}_{t-1}\,;\,\widehat{\boldsymbol{\gamma}}_{t-1},\widehat{\mathbf{\Sigma}}_{t-1}\right) . \end{aligned} \tag{7}$$

[1]The approximations mirror those used in [22] to derive a tracking algorithm on the unit circle that uses von Mises (vM) [11] distributions.

This mirrors the statistical representation of the state in the SDS (see (4)-(5)). To propagate it to the next time step t, we must solve the Bayesian filtering equations:

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t-1}\right) = \int_{\mathbb{S}^2\times\mathbb{R}^3} p\left(\mathbf{s}_t|\mathbf{s}_{t-1}\right)p\left(\mathbf{s}_{t-1}|\mathbf{y}_{1:t-1}\right)d\,\mathbf{s}_{t-1} \quad , \tag{8}$$

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t}\right) \propto p\left(\mathbf{y}_t|\mathbf{s}_t\right)p\left(\mathbf{s}_t|\mathbf{y}_{1:t-1}\right) \quad . \tag{9}$$

4.1. Predict Step

We can write (8) as:

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t-1}\right) = p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right)p\left(\mathbf{r}_t|\mathbf{y}_{1:t-1}\right) \quad , \tag{10}$$

where:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) = \int_{\mathbb{S}^2} p\left(\mathbf{x}_t|\mathbf{s}_{t-1}\right)p\left(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}\right)d\,\mathbf{x}_{t-1} \quad , \tag{11}$$

$$p\left(\mathbf{r}_t|\mathbf{y}_{1:t-1}\right) = \int_{\mathbb{R}^3} p\left(\mathbf{r}_t|\mathbf{r}_{t-1}\right)p\left(\mathbf{r}_{t-1}|\mathbf{y}_{1:t-1}\right)d\,\mathbf{r}_{t-1} \quad . \tag{12}$$

The first term (11) represents a rotation followed by vMF sampling (see (4)). Although the rotation is performed deterministically conditioned on $\mathbf{s}_{t-1}$, there is still statistical coupling between the position $\mathbf{x}_t$ and rotation $\mathbf{r}_{t-1}$.[2] We found empirically that when $\mathbf{\Sigma}_{\mathbf{r}} = \sigma_{\mathbf{r}}^2\mathbf{I}$, the noise due to the rotation is well-approximated as vMF-distributed with concentration $1/\sigma_{\mathbf{r}}^2$. So, we write (11) as:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) \approx \int_{\mathbb{S}^2} \bar{p}\left(\mathbf{x}_t|\tilde{\mathbf{x}}_{t-1}\right)p\left(\tilde{\mathbf{x}}_{t-1}|\mathbf{y}_{1:t-1}\right)d\,\tilde{\mathbf{x}}_{t-1} \quad , \tag{13}$$

where $\tilde{\mathbf{x}}_{t-1} = \mathbf{R}\left(\widehat{\boldsymbol{\gamma}}_{t-1}\right)\mathbf{x}_{t-1}$ and $\bar{p}\left(\mathbf{x}_t|\tilde{\mathbf{x}}_{t-1}\right)$ denotes the convolution of the position and rotation noise vMFs. We recognize (13) as the convolution of two vMF distributions and approximate it by convolving two wrapped Normal distributions that are matched to the vMFs [11, 23]. This amounts to computing:

$$\kappa' = A^{-1}\left(A\left(\kappa_1\right)A\left(\kappa_2\right)\right) \quad , \tag{14}$$

where:

$$A\left(\kappa\right) = \frac{1}{\tanh\left(\kappa\right)} - \frac{1}{\kappa} \quad . \tag{15}$$

We can do this easily by finding the root of:

$$A\left(\kappa'\right) - A\left(\kappa_1\right)A\left(\kappa_2\right) = 0 \quad , \tag{16}$$

with Newton's method, initialized at $\kappa' = \frac{1}{2}min\left(\kappa_1,\kappa_2\right)$. This converges to within machine precision in at most 10 iterations. A close look at (13)-(14) reveals that we can approximate the convolutions due to both position and rotation noise by finding the root of:

$$A\left(\widehat{\kappa}_t^-\right) - A\left(\widehat{\kappa}_{t-1}\right)A\left(\widehat{\kappa}_{\mathbf{x}}\right)A\left(1/\sigma_{\mathbf{r}}^2\right) = 0 \quad , \tag{17}$$

initialized at $\widehat{\kappa}_t^- = \frac{1}{3}min\left(\widehat{\kappa}_{t-1},\widehat{\kappa}_{\mathbf{x}},1/\sigma_{\mathbf{r}}^2\right)$.

The second term (12) can be evaluated in closed form as in the Kalman filter. Thus, we have that the predicted density is:

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t-1}\right) \approx vMF\left(\mathbf{x}_t\,;\,\widehat{\boldsymbol{\mu}}_t^-,\widehat{\kappa}_t^-\right)\mathcal{N}\left(\mathbf{r}_t\,;\,\widehat{\boldsymbol{\gamma}}_t^-,\widehat{\mathbf{\Sigma}}_t^-\right) \quad , \tag{18}$$

where:

[2]This is analogous to the coupling between position and velocity components in the LDS.

$$\widehat{\boldsymbol{\mu}}_t^- = \mathbf{R}\left(\widehat{\boldsymbol{\gamma}}_{t-1}\right)\widehat{\boldsymbol{\mu}}_{t-1} \quad , \tag{19}$$

$$\widehat{\kappa}_t^- = A^{-1}\left(A\left(\widehat{\kappa}_{t-1}\right)A\left(\widehat{\kappa}_{\mathbf{x}}\right)A\left(1/\sigma_{\mathbf{r}}^2\right)\right) \quad , \tag{20}$$

$$\widehat{\boldsymbol{\gamma}}_t^- = \mathbf{A}\widehat{\boldsymbol{\gamma}}_{t-1} \quad , \tag{21}$$

$$\widehat{\boldsymbol{\Sigma}}_t^- = \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{t-1}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{r}} \quad . \tag{22}$$

4.2. Correct Step

We next update the predicted density (18) by taking the observation $\mathbf{y}_t$ into account. This is an application of Bayes' rule. We can write (9) as:

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t}\right) = p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right)p\left(\mathbf{r}_t|\mathbf{y}_{1:t}\right) \quad , \tag{23}$$

where:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t}\right) \propto p\left(\mathbf{y}_t|\mathbf{x}_t\right)p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) \quad , \tag{24}$$

$$p\left(\mathbf{r}_t|\mathbf{y}_{1:t}\right) \propto p\left(\mathbf{y}_t|\mathbf{r}_t\right)p\left(\mathbf{r}_t|\mathbf{y}_{1:t-1}\right) \quad . \tag{25}$$

The first term (24) takes the form of a vMF distribution. The second term (25) is problematic in that the emission density $p\left(\mathbf{y}_t|\mathbf{r}_t\right)$ is not defined in the SDS (only a source position measurement is available).[3] We will handle this by generating an auxiliary observation $\mathbf{y}_t^{\mathbf{r}}$ in $\mathbf{r}$-space (from $\mathbf{y}_t$) and applying the regular Kalman filter update scheme. We choose $\mathbf{y}_t^{\mathbf{r}}$ to be the rotation vector required to move from the previous position estimate $\widehat{\boldsymbol{\mu}}_{t-1}$ to the observation $\mathbf{y}_t$:

$$\mathbf{y}_t^{\mathbf{r}} = \cos^{-1}\left(\widehat{\boldsymbol{\mu}}_{t-1}\cdot\mathbf{y}_t\right)\frac{\widehat{\boldsymbol{\mu}}_{t-1}\times\mathbf{y}_t}{\|\widehat{\boldsymbol{\mu}}_{t-1}\times\mathbf{y}_t\|_2} \quad , \tag{26}$$

where $\cdot$ and $\times$ denote dot and cross products, respectively. The first term provides the amount of rotation and the second term provides the axis of rotation as described in Section 2.2.

Altogether, we approximate the filtered density at time t as:

$$p\left(\mathbf{s}_t|\mathbf{y}_{1:t}\right) \approx vMF\left(\mathbf{x}_t\,;\,\widehat{\boldsymbol{\mu}}_t,\widehat{\kappa}_t\right)\mathcal{N}\left(\mathbf{r}_t\,;\,\widehat{\boldsymbol{\gamma}}_t,\widehat{\boldsymbol{\Sigma}}_t\right) \quad , \tag{27}$$

where:

$$\widehat{\boldsymbol{\mu}}_t = \frac{\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t}{\|\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t\|_2} \quad , \tag{28}$$

$$\widehat{\kappa}_t = \|\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t\|_2 \quad , \tag{29}$$

$$\widehat{\boldsymbol{\gamma}}_t = \widehat{\boldsymbol{\gamma}}_t^- - \mathbf{K}_t\left(\mathbf{y}_t^{\mathbf{r}} - \widehat{\boldsymbol{\gamma}}_t^-\right) \quad , \tag{30}$$

$$\widehat{\boldsymbol{\Sigma}}_t = \left(\mathbf{I} - \mathbf{K}_t\right)\widehat{\boldsymbol{\Sigma}}_t^- \quad , \tag{31}$$

and the Kalman gain is $\mathbf{K}_t = \widehat{\boldsymbol{\Sigma}}_t^-\left(\widehat{\boldsymbol{\Sigma}}_t^- + \boldsymbol{\Sigma}_{\mathbf{y}^{\mathbf{r}}}\right)^{-1}$. The noise covariance $\boldsymbol{\Sigma}_{\mathbf{y}^{\mathbf{r}}}$ controls how sensitive the rotation vector update is to the auxiliary observation.

The overall filtering procedure, called the von Mises-Fisher Filter, recursively computes (18) and (27) over time and is summarized in Algorithm 1.

Algorithm 1 von Mises-Fisher Filter

Predict

Position (vMF)

$\widehat{\boldsymbol{\mu}}_t^- = \mathbf{R}\left(\widehat{\boldsymbol{\gamma}}_{t-1}\right)\widehat{\boldsymbol{\mu}}_{t-1}$

$\widehat{\kappa}_t^- = A^{-1}\left(A\left(\widehat{\kappa}_{t-1}\right)A\left(\widehat{\kappa}_{\mathbf{x}}\right)A\left(1/\sigma_{\mathbf{r}}^2\right)\right)$

Rotation (Normal)

$\widehat{\boldsymbol{\gamma}}_t^- = \mathbf{A}\widehat{\boldsymbol{\gamma}}_{t-1}$

$\widehat{\boldsymbol{\Sigma}}_t^- = \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{t-1}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{r}}$

Correct

Position (vMF)

$\widehat{\boldsymbol{\mu}}_t = \frac{\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t}{\|\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t\|_2}$

$\widehat{\kappa}_t = \|\widehat{\kappa}_t^-\widehat{\boldsymbol{\mu}}_t^- + \kappa_{\mathbf{y}}\mathbf{y}_t\|_2$

Rotation (Normal)

$\mathbf{y}_t^{\mathbf{r}} = \cos^{-1}\left(\widehat{\boldsymbol{\mu}}_{t-1}\cdot\mathbf{y}_t\right)\frac{\widehat{\boldsymbol{\mu}}_{t-1}\times\mathbf{y}_t}{\|\widehat{\boldsymbol{\mu}}_{t-1}\times\mathbf{y}_t\|_2}$

$\mathbf{K}_t = \widehat{\boldsymbol{\Sigma}}_t^-\left(\widehat{\boldsymbol{\Sigma}}_t^- + \boldsymbol{\Sigma}_{\mathbf{y}^{\mathbf{r}}}\right)^{-1}$

$\widehat{\boldsymbol{\gamma}}_t = \widehat{\boldsymbol{\gamma}}_t^- - \mathbf{K}_t\left(\mathbf{y}_t^{\mathbf{r}} - \widehat{\boldsymbol{\gamma}}_t^-\right)$

$\widehat{\boldsymbol{\Sigma}}_t = \left(\mathbf{I} - \mathbf{K}_t\right)\widehat{\boldsymbol{\Sigma}}_t^-$

5. MULTI-SOURCE TRACKING IN CLUTTER

When multiple sources and observations are present, we must decide how to group observations with sources. We will use a probabilistic data association (PDA) strategy [15, 18] as it provides a simple way to extend the vMFF to the multi-source case. The underlying states $\mathbf{s}_{t,j}$ of the sources are assumed to evolve independently over time. Thus, the generative model is factorial in nature.

To apply PDA to this setting, we re-work the emission model (6) so that M observations are drawn i.i.d. from a mixture of $K+1$ distributions. The first K are vMFs (one for each source) and the last is a uniform vMF that accounts for outliers. Thus, the observation set $\mathbf{y}_{t,1},\ldots,\mathbf{y}_{t,M}$ is drawn according to:

$$\mathbf{y}_{t,m}|\mathbf{x}_{t,1:K} \sim \frac{\beta}{K}\sum_{j=1}^{K} vMF\left(\mathbf{x}_{t,j},\kappa_{\mathbf{y}}\right) + \left(1-\beta\right)vMF\left(\mathbf{u},0\right) \quad , \tag{32}$$

where β is the proportion of inliers and $\mathbf{u}$ is any vector in $\mathbb{S}^2$. We can associate observations to source vMFs with the posterior probabilities:

$$\eta_{t,jm} = \frac{\frac{\beta}{K}vMF\left(\mathbf{y}_{t,m}\,;\,\widehat{\boldsymbol{\mu}}_{t,j}^-,\kappa_{\mathbf{y}}\right)}{\frac{\beta}{K}\sum_{j=1}^{K} vMF\left(\mathbf{y}_{t,m}\,;\,\widehat{\boldsymbol{\mu}}_{t,j}^-,\kappa_{\mathbf{y}}\right) + \left(1-\beta\right)vMF\left(\mathbf{y}_{t,m}\,;\,\mathbf{u},0\right)} \quad . \tag{33}$$

We incorporate these weights in the filter by computing the parameters in (27) for each target using composite observations:

$$\bar{\mathbf{y}}_{t,j} = \sum_{m=1}^{M}\eta_{t,jm}\,\mathbf{y}_{t,m} \quad , \quad \mathbf{y}_{t,j}^{\mathbf{r}} = \frac{\sum_{m=1}^{M}\eta_{t,jm}\,\mathbf{y}_{t,m}^{\mathbf{r}}}{\sum_{m=1}^{M}\eta_{t,jm}} \quad . \tag{34}$$

When there is only one source and no outliers ($\beta = 1$), these expressions reduce to more familiar ones for posterior inference with M measurements [24]. When multiple sources are present, PDA splits the "weight" of the observations among the K source tracks. And when $\beta < 1$, outliers are "soaked up" by the uniform vMF, preventing the target vMFs from diverging. The predict step updates are performed as in the one-source case and independently for each target. This yields an inference algorithm very similar to the vMFF that we call the Factorial vMFF (FvMFF), summarized in Algorithm 2.

[3] This complication does not arise in a Kalman filter because the coupling between the position and velocity components is captured in off-diagonal blocks of a state covariance $\boldsymbol{\Sigma}_{\mathbf{s}}$.

Algorithm 2 Factorial von Mises-Fisher Filter

Predict

Position (vMF)

$$\widehat{\boldsymbol{\mu}}_{t,j}^{-} = \mathbf{R}\left(\widehat{\boldsymbol{\gamma}}_{t-1,j}\right)\widehat{\boldsymbol{\mu}}_{t-1,j}$$
$$\widehat{\kappa}_{t,j}^{-} = A^{-1}\left(A\left(\widehat{\kappa}_{t-1,j}\right)A\left(\widehat{\kappa}_{\mathbf{x}}\right)A\left(1/\sigma_{\mathbf{r}}^{2}\right)\right)$$

Rotation (Normal)

$$\widehat{\boldsymbol{\gamma}}_{t,j}^{-} = \mathbf{A}\widehat{\boldsymbol{\gamma}}_{t-1,j}$$
$$\widehat{\boldsymbol{\Sigma}}_{t}^{-} = \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{t-1}\mathbf{A}^{\top} + \boldsymbol{\Sigma}_{\mathbf{r}}$$

Data Association

$$\eta_{t,jm} = \frac{\frac{\beta}{K}vMF\left(\mathbf{y}_{t,m};\widehat{\boldsymbol{\mu}}_{t,j}^{-},\kappa_{\mathbf{y}}\right)}{\frac{\beta}{K}\sum_{j=1}^{K}vMF\left(\mathbf{y}_{t,m};\widehat{\boldsymbol{\mu}}_{t,j}^{-},\kappa_{\mathbf{y}}\right)+(1-\beta)vMF\left(\mathbf{y}_{t,m};\mathbf{u},0\right)}$$

Correct

Position (vMF)

$$\bar{\mathbf{y}}_{t,j} = \sum_{m=1}^{M}\eta_{t,jm}\,\mathbf{y}_{t,m}$$
$$\widehat{\boldsymbol{\mu}}_{t,j} = \frac{\widehat{\kappa}_{t,j}^{-}\widehat{\boldsymbol{\mu}}_{t,j}^{-}+\kappa_{\mathbf{y}}\bar{\mathbf{y}}_{t,j}}{\|\widehat{\kappa}_{t,j}^{-}\widehat{\boldsymbol{\mu}}_{t,j}^{-}+\kappa_{\mathbf{y}}\bar{\mathbf{y}}_{t,j}\|_2}$$
$$\widehat{\kappa}_{t,j} = \|\widehat{\kappa}_{t,j}^{-}\widehat{\boldsymbol{\mu}}_{t,j}^{-} + \kappa_{\mathbf{y}}\bar{\mathbf{y}}_{t,j}\|_2$$

Rotation (Normal)

$$\mathbf{y}_{t,m}^{\mathbf{r}} = \cos^{-1}\left(\widehat{\boldsymbol{\mu}}_{t-1,j}\cdot\mathbf{y}_{t,m}\right)\frac{\widehat{\boldsymbol{\mu}}_{t-1,j}\times\mathbf{y}_{t,m}}{\|\widehat{\boldsymbol{\mu}}_{t-1,j}\times\mathbf{y}_{t,m}\|_2}$$
$$\mathbf{K}_t = \widehat{\boldsymbol{\Sigma}}_{t}^{-}\left(\widehat{\boldsymbol{\Sigma}}_{t}^{-} + \boldsymbol{\Sigma}_{\mathbf{y}^{\mathbf{r}}}\right)^{-1}$$
$$\mathbf{y}_{t,j}^{\mathbf{r}} = \frac{\sum_{m=1}^{M}\eta_{t,jm}\,\mathbf{y}_{t,m}^{\mathbf{r}}}{\sum_{m=1}^{M}\eta_{t,jm}}$$
$$\widehat{\boldsymbol{\gamma}}_{t,j} = \widehat{\boldsymbol{\gamma}}_{t,j}^{-} - \mathbf{K}_t\left(\mathbf{y}_{t,j}^{\mathbf{r}} - \widehat{\boldsymbol{\gamma}}_{t,j}^{-}\right)$$
$$\widehat{\boldsymbol{\Sigma}}_{t} = \left(\mathbf{I} - \mathbf{K}_t\right)\widehat{\boldsymbol{\Sigma}}_{t}^{-}$$

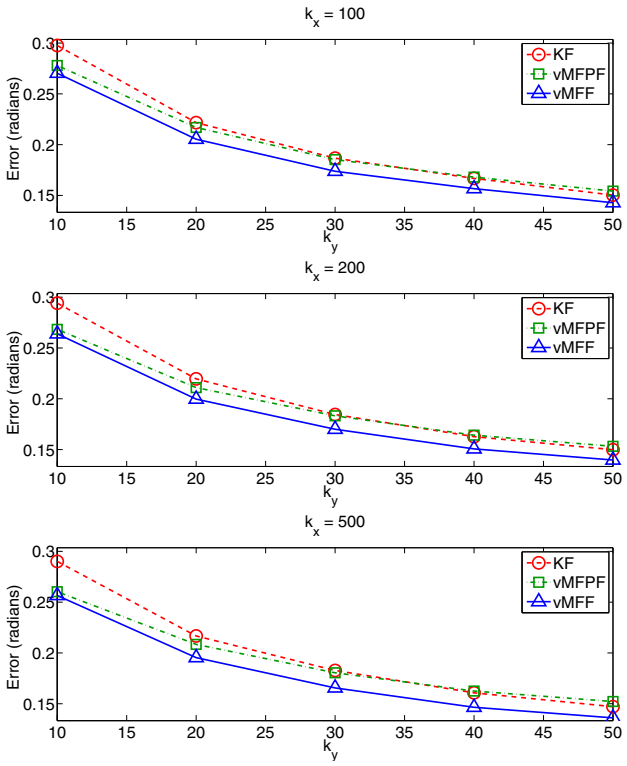

Fig. 2. Accuracy results of synthetic trials with data generated from the spherical dynamical system.

6. EXPERIMENTS

6.1. Synthetic Experiments

We compared the vMFF with (1) a Kalman Filter (KF) for tracking a constant-velocity LDS in $\mathbb{R}^3$ and (2) a particle filter based on vMF sampling (vMFPF) [14]. The vMFPF uses the state transition density as the proposal and multinomial resampling at every step. We used 50 particles and included a rotation vector component in the vMFPF. For each of 100 trials, we sampled a sequence of length 200 from the SDS ($\mathbf{A} = \mathbf{I}$, $\sigma_{\mathbf{r}}^2 = 0.02$, $\beta = 1$, $K = 1$) using the vMF sampling scheme described in [25]. The auxiliary observation covariance in the vMFF was set by inspection to $\boldsymbol{\Sigma}_{\mathbf{y}^{\mathbf{r}}} = \mathbf{I}$. We evaluated the tracking error using average geodesic distance on the sphere:

$$E = \frac{1}{T}\sum_{t=1}^{T}\cos^{-1}\left(\mathbf{x}_t^{\top}\widehat{\boldsymbol{\mu}}_t\right) \quad . \tag{35}$$

Results from these trials are show in Fig. 2. The KF ignores the topology of the sphere, tracking a 3D location rather than a 2D DOA vector. This leads to a decrease in tracking accuracy. The vMFPF does take the topology of $\mathbb{S}^2$ into account, but requires many samples to accurately represent the state. Fig. 3 shows how its performance depends on the number N_p of particles. For large N_p, it achieves a lower error than the vMFF. This is expected since the vMFPF estimate is asymptotically optimal for the SDS as $N_p \to \infty$, while the deterministic approximations of the vMFF are limited in accuracy. However, the vMFPF comes with a dramatic increase in computation. In our MATLAB implementations, the average computation times per iteration were 0.064, 6.485, and 0.305 ms for the KF, vMFPF ($N_p = 50$), and vMFF. To match the performance of the vMFF, the vMFPF must run 60 times more slowly with $N_p \approx 150$. We conclude that the vMFF performs well across all concentration values and strikes a compromise between the efficiency of the KF and the statistical grounding of the vMFPF.

6.2. Feature Extraction for Speaker Tracking

In order to apply the proposed tracking algorithms to the speaker-tracking problem, we must generate observations on $\mathbb{S}^2$ from the incoming audio streams. This is commonly done via the Generalized Cross-Correlation (GCC) [26] function. In [27], the authors used the GCC with an additional PHAse Transform (PHAT) to localize sound sources on a hemisphere with an array of microphones spaced 15 cm apart. However, for compact arrays, Interchannel Time Delay (ITD) features [28, 29] are more effective.

In our speaker-tracking experiments, we extracted 3-dimensional ITD features on a frame-by-frame basis from a 4-channel recording. An N-point DFT was computed for each microphone. ITDs relative to the first channel were then computed for each set of corresponding DFT coefficients, resulting in $N/2$ features per frame. Formally, let $X_{f,t}^{(i)}$ be the f^{th} DFT coefficient at frame t for the i^{th} channel. The f^{th} feature vector at frame t is:

$$\boldsymbol{\delta}_{f,t} = \frac{N}{2\pi f}\left(\angle X_{f,t}^{(1)} - \angle \mathbf{X}_{f,t}^{(2:4)}\right) \quad . \tag{36}$$

The ITD vectors $\{\boldsymbol{\delta}_{f,t}\}, f = 1,\ldots,N/2$ can be mapped to the hemisphere of DOAs above the array using the well-known least-

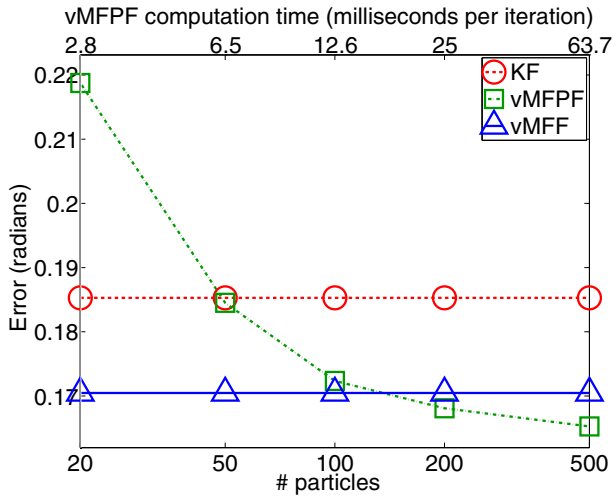

Fig. 3. Impact of the number of particles used in the vMFPF on tracking error and computation time ($\kappa_{\mathbf{x}} = 200$, $\kappa_{\mathbf{y}} = 30$). The horizontal axes are on a logarithmic scale.

squares method described in [30]. It cannot be applied directly since the microphones in our array are coplanar. However, we can adapt the approach by estimating the azimuth and zenith angles in sequence. Without loss of generality, assume that the microphones are located in the x-y plane of a Cartesian coordinate system. Given an ITD vector $\boldsymbol{\delta}$, we first solve for the azimuth θ via:

$$\mathbf{L} = \left(\mathbf{m}_1 \mathbf{1}^\top - \begin{bmatrix}\mathbf{m}_2 & \mathbf{m}_3 & \mathbf{m}_4\end{bmatrix}\right)^\top \quad , \tag{37}$$

$$v = \mathbf{L}^{-1}\,\boldsymbol{\delta} \quad , \tag{38}$$

$$\theta = atan2\,(v_2, v_1) \quad , \tag{39}$$

where $\mathbf{m}_i \in \mathbb{R}^{2\times 1}$ is the location of the i^{th} microphone. We then solve for the zenith ϕ by considering its effect on the x-y component of the wavefront velocity. Thus, we have that:

$$\phi = \sin^{-1}\left(\frac{\delta_*}{\bar{\delta}_*}\right) \quad , \tag{40}$$

where δ_* is any component of $\boldsymbol{\delta}$ and $\bar{\delta}_*$ is the same component of the ITD vector we expect to see for a source with azimuth θ and zenith $\phi = \frac{\pi}{2}$. Converting to Cartesian coordinates gives the measurement:

$$\mathbf{y} = \begin{bmatrix}\cos(\theta)\sin(\phi) & \sin(\theta)\sin(\phi) & \cos(\phi)\end{bmatrix}^\top \quad . \tag{41}$$

In our experiments, we pooled the measurements from a block of 3 frames to use as the observation set. For added robustness, we associated with each observation a weight equal to the magnitude of the corresponding DFT coefficient. To incorporate the weights into Algorithm 2, we multiply them with the posterior probabilities in (33). This ensures that only significant features are considered.

6.3. Speaker Tracking Experiments

We applied the FvMFF to the task of tracking multiple speakers in a simulated reverberant environment. A 4-microphone array positioned in a 2×2-centimeter square was placed in the middle of a room of size $5 \times 5 \times 5$ meters with a T_{60} reverberation time of 100 milliseconds. 2- to 3-second sentences from the TSP corpus [31], down-sampled to 16 kHz, were played as the speakers moved around

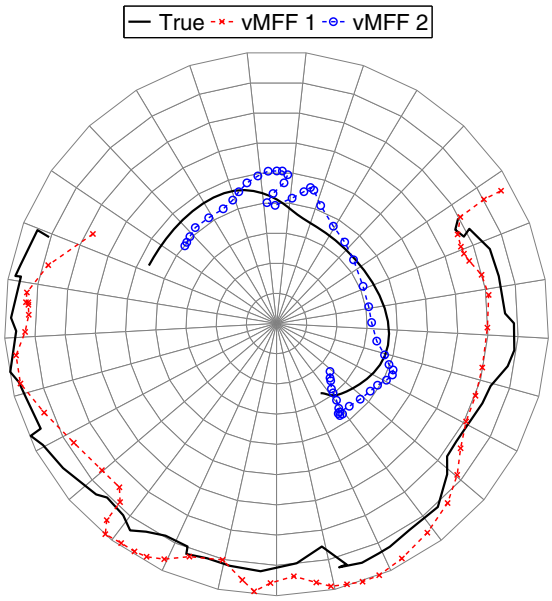

Fig. 4. Example of 2-speaker tracking with the FvMFF on the DOA hemisphere above a 4-channel microphone array. The hemisphere has been flattened such that radial distance from the center of the figure corresponds to zenith angle.

the array to create a reverberant 0-dB mixture. ITD features were extracted from this recording as described in Section 6.2 with a DFT window size of $N = 1024$ and the FvMFF was run on the resulting measurement sequence. Fig. 4 shows an example.

We found empirically that, although the outlier rejection method described in Section 5 behaves well for data generated from a multi-source SDS, a more aggressive approach is necessary for speech data. Thus, we included a gating procedure after the data association step in Algorithm 2 that, for each vMF, zeros-out the posteriors for measurements too far from its mean. This helps the observation set look more like what the FvMFF is expecting.

7. CONCLUSIONS

In this paper, we described the problem of tracking one or more speakers with a compact microphone array. The speaker locations are specified by their directions-of-arrival (DOA), which lie on the surface of the unit sphere, $\mathbb{S}^2$. The Kalman filter is only directly applicable for tracking the source DOAs if we ignore the unique topology of the sphere. To avoid this, we introduced a spherical dynamical system (SDS) model that describes the evolution of a DOA vector directly on $\mathbb{S}^2$ that included a rotation vector representation of the source's velocity along the surface of the sphere. We then applied a series of deterministic approximations to the corresponding Bayesian filtering equations to derive a simple inference algorithm for the SDS: the von Mises-Fisher Filter (vMFF). This was extended to the multi-source case via sensor fusion and probabilistic data association techniques in the Factorial vMFF (FvMFF).

Through synthetic trials, we showed that the vMFF, which maintains the mean and concentration of a von Mises-Fisher distribution over time, is significantly more efficient than a particle filter applied to the SDS. We also showed that the vMFF is generally more accurate than both the particle filter and a 3D Kalman filter. Finally, we

demonstrated that the FvMFF can track multiple speakers in noisy, reverberant conditions using DOA measurements extracted from interchannel time delay (ITD) features.

8. ACKNOWLEDGEMENT

The authors would like to thank Noah Stein and Adam Miller for insightful discussions leading to the development of this work.

9. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Topics in Signal Processing: Microphone Array Signal Processing*, vol. 1, Springer, 2008.

[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[3] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep., University of North Carolina at Chapel Hill, 2006.

[4] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[5] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[6] J. Traa and P. Smaragdis, "A wrapped Kalman filter for azimuthal speaker tracking," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1257–1260, 2013.

[7] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[8] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden Markov models," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 114–117, 2005.

[9] N. Mitianoudis, "A generalized directional Laplacian distribution: Estimation, mixture models and audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2397–2408, 2012.

[10] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angular distributions," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4629–4632, 2012.

[11] K. Mardia and P. Jupp, *Directional Statistics*, Wiley, 1999.

[12] H. Tang, S. M. Chu, and T. S. Huang, "Generative model-based speaker clustering via mixture of von Mises-Fisher distributions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4101–4104, 2009.

[13] G. Siddharth and Y. Yiming, "Von Mises-Fisher clustering models," *Journal of Machine Learning Research*, vol. 32, pp. 154–162, 2014.

[14] F. Zhang, E. R. Hancock, C. Goodlett, and G. Gerig, "Probabilistic white matter fiber tracking using particle filtering and von Mises-Fisher sampling," *Medical Image Analysis*, vol. 13, no. 1, pp. 5–18, 2009.

[15] J. Traa, "Multichannel source separation and tracking with phase differences by random sample consensus," M.S. thesis, University of Illinois at Urbana-Champaign, 2013.

[16] J. Traa and P. Smaragdis, "Blind multi-channel source separation by circular-linear statistical modeling of phase differences," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[17] Q. Gan and C. Harris, "Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 273–279, 2001.

[18] T. Kirubarajan and Y. Bar-Shalom, "Probabilistic data association techniques for target tracking in clutter," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 536–557, 2004.

[19] H. Gauvrit, J.-P. Le Cadre, and C. Jauffret, "A formulation of multitarget tracking as an incomplete data problem," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 4, pp. 1242–1257, 1997.

[20] J. Glover and L. P. Kaelbling, "Tracking 3-D rotations with the quaternion Bingham filter," Tech. Rep., MIT - Computer Science and Artificial Intelligence Laboratory (CSAIL), 2013.

[21] J. T. Kent, "The Fisher-Bingham distribution on the sphere," *Journal of the Royal Statistical Society*, vol. 44, no. 1, pp. 71–80, 1982.

[22] I. Markovic and I. Petrovic, "Bearing-only tracking with a mixture of von Mises distributions," *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 707–712, 2012.

[23] W. Jakob, "Numerically stable sampling of the von Mises-Fisher distribution on S2 (and other tricks)," Tech. Rep., Cornel University, 2012.

[24] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[25] G. Ulrich, "Computer generation of distributions on the m-sphere," *Journal of the Royal Statistical Society, Series C*, vol. 33, no. 2, pp. 158–163, 1984.

[26] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[27] S. T. Birchfield and D. K. Gillmor, "Acoustic source direction by hemisphere sampling," *IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3053–3056, 2001.

[28] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[29] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic source tracking applying Rao-Blackwellized Monte Carlo data association," *IEEE 15th Workshop on Statistical Signal Processing*, pp. 253–256, 2009.

[30] K. M. Varma, "Time delay estimate based direction of arrival estimation for speech in reverberant environments," M.S. thesis, Virginia Polytechnic Institute and State University, 2002.

[31] P. Kabal, "TSP speech database," 2002, Telecommunications and Signal Processing Lab, McGill University.