Tracking Multiple Audio Sources with the von Mises Distribution and Variational EM

Yutong Ban,† Xavier Alameda-Pineda,† Member, IEEE, Christine Evers,‡ Senior Member, IEEE, and Radu Horaud†

Abstract—In this paper we address the problem of simultaneously tracking several audio sources, namely the problem of estimating source trajectories from a sequence of observed features. We propose to use the von Mises distribution to model audio-source directions of arrival (DOAs) with circular random variables. This leads to a multi-target Kalman filter formulation which is intractable because of the combinatorial explosion of associating observations to state variables over time. We propose a variational approximation of the filter’s posterior distribution and we infer a variational expectation maximization (VEM) algorithm which is computationally efficient. We also propose an audio-source birth method that favors smooth source trajectories and which is used both to initialize the number of active sources and to detect new sources. We perform experiments with a recently released dataset comprising several moving sources as well as a moving microphone array.

Index Terms—Multiple target tracking, audio-source tracking, speaker tracking, Bayesian filtering, von Mises distribution, variational approximation, expectation maximization.

I. INTRODUCTION

In this paper we address the problem of simultaneously tracking several audio sources, namely the problem of estimating source trajectories from a sequence of observed features. Audio-source tracking is useful for a number of tasks such as audio-source separation, spatial filtering, speech enhancement and speech recognition, which in turn are essential for robust voice-based home assistants. Audio source tracking is difficult because audio signals are adversely affected by noise, reverberation and interferences between acoustic signals.

Single-source tracking methods are often based on observing the time differences of arrival (TDOAs) between two microphones. Since the mapping between TDOAs and the space of source locations is non-linear, sequential Monte Carlo particle filters are used, e.g. [1]–[3]. Alternatively, microphone arrays can be used to estimate DOAs of audio sources. The problem can then be cast into a linear dynamic model, e.g. the adaptive Kalman filter [4]. In this case source directions should however be modeled as circular random variables, e.g. the wrapped Gaussian distribution [5], or the von Mises distribution [6], [7].

Multiple-source tracking is more challenging since it raises additional difficulties: (i) the number of active audio sources is unknown and it varies over time, (ii) multiple DOAs must be simultaneously estimated, and (iii) DOA-to-source assignments must be computed over time. The problem of tracking an unknown number of sources is specifically addressed in the framework of random finite sets [8]. Since the probability density function (pdf) is computationally intractable, its first order approximation can be propagated in time using the probability hypothesis density (PHD) filter [8], [9]. In [10] the PHD filter was applied to audio recordings to track multiple sources from TDOA estimates. In [11] the wrapped Gaussian distribution is incorporated within a PHD filter. A mixture of von Mises distributions is combined with a PHD filter in [12]. The main drawback of PHD-based filters is that they cannot provide explicit observation-to-source associations without recourse to ad-hoc post-processing. Providing such associations is crucial when several audio sources must be identified and tracked.

Multiple target tracking is also formulated as a variational approximation of Bayesian filtering [13]. Observation-to-target associations are modeled as discrete latent variables and their realizations are estimated within a compact and efficient VEM solver. Moreover, the problem of tracking a varying number of targets is addressed via track-birth and track-death processes. The variational approximation of [13] was recently extended to track multiple audio sources using a mixture of Gaussian distributions [14].

This paper builds on [6], [13] and [14] and proposes to use the von Mises distribution to model DOAs with circular random variables. This leads to a multi-target Kalman filter which is intractable because of the combinatorial explosion of associating observations to state variables over time. We propose a variational approximation of the filter’s posterior distribution and we infer a VEM algorithm which is computationally efficient. We also propose an audio-source birth method that favors smooth source trajectories and which is used both to initialize the number of active sources and to detect new sources. We perform experiments with a recently released dataset comprising several moving sources as well as a moving microphone array.

The paper is organized as follows. Section II describes the probabilistic model and Section III describes a variational approximation of the filtering distribution and the VEM algorithm. Section IV briefly describes the source birth method. Experiments and comparisons with other methods are

†Y. Ban, X. Alameda-Pineda and R. Horaud are with Inria Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: first.last@inria.fr
‡C. Evers is with Dept. Electrical and Electronic Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK. Email: c.evers@imperial.ac.uk
This work was supported by the ERC Advanced Grant VHIA #340113 and the UK EPSRC Fellowship grant no. EP/P001017/1.
described in Section V. Supplemental materials (mathematical derivations, software and videos) can be found at the site.1

II. THE FILTERING DISTRIBUTION

Let N denote the unknown number of audio sources and the state stn ∈ (−π, π] be the DOA of source n ∈ {1, ..., N} at t, and st = (s1, ..., stN). Furthermore, let ytm ∈ (−π, π] denote the m-th DOA observation at time t, where m ∈ {1, ..., Mt} and Mt is the number of observations at t. Each observation is accompanied by its corresponding confidence ωtm ∈ [0, 1]. Let ztm denote the observation-to-source assignment variable, where ztm = n means that ytm is an observation of source n, and ztm = 0 means that ytm is an observation of a dummy source.

Within a Bayesian model, multiple target tracking can be formulated as the estimation of the filtering distribution p(st, zt|y1:t). We assume that the state variables stn follow a first-order Markov model, and that the observations depend only on the current state and on the assignment variables. Under these two hypotheses the posterior pdf is given by:

\[ p(s_t, z_t | y_{1:t}) \propto p(y_t | z_t, s_t)p(z_t)p(s_t | y_{1:t-1}), \]  
(1)

where \( p(y_t | z_t, s_t) \) is the observation likelihood, \( p(z_t) \) is the prior pdf of the assignment variables and \( p(s_t | y_{1:t-1}) \) is the predictive pdf of the state variables.

1) Observation likelihood: Assuming that DOA estimates are independently and identically distributed (i.i.d.), the observation likelihood can be written as:

\[ p(y_t | z_t, s_t) = \prod_{m=1}^{M_t} p(y_{tm} | z_t, s_t). \]  
(2)

The likelihood that a DOA corresponds to a source is modelled by a von Mises distribution [6], whereas the likelihood that a DOA corresponds to a dummy source is modelled by a uniform distribution:

\[ p(y_{tm} | z_t = n, s_t) = \begin{cases} \mathcal{M}(y_{tm}; s_t, \kappa_n \omega_{tm}) & n \neq 0 \\ \mathcal{U}(y_{tm}) & n = 0 \end{cases}, \]  
(3)

where \( \mathcal{M}(y; s, \kappa) = (2\pi I_0(\kappa))^{-1} \exp\{\kappa \cos(y - s)\} \) denotes the von Mises distribution with mean s and concentration \( \kappa, I_0(\cdot) \) denotes the modified Bessel function of the first kind of order \( p, \kappa_y \) denotes the concentration of audio observations, and \( \mathcal{U}(y_{tm}) = (2\pi)^{-1} \) denotes the uniform distribution along the support of the unit circle.

2) Prior pdf of the assignment variables: Assuming that the assignment variables are i.i.d., the prior pdf is given by:

\[ p(z_t) = \prod_{m=1}^{M_t} p(Z_{tm} = n), \]  
(4)

and we denote with \( \pi_n = p(Z_{tm} = n), \sum_{n=0}^{N} \pi_n = 1 \), the prior probability that source n is associated with ytm.

3) Predictive pdf of the state variables: The predictive pdf extrapolates information inferred in the past to the current time step using a dynamical model of the source motion, i.e.,

\[ p(s_t | y_{1:t-1}) = \int p(s_t | s_{t-1})p(s_{t-1} | y_{1:t-1})ds_{t-1}. \]  
(5)

where \( p(s_t | s_{t-1}) \) denotes the prior pdf modelling the source motion and \( p(s_{t-1} | y_{1:t-1}) \) corresponds to the filtering distribution at time \( t-1 \). The source motion model is assumed source-independent and follows a von Mises distribution:

\[ p(s_t | s_{t-1}) = \prod_{n=1}^{N} \mathcal{M}(s_{tn}; s_{t-1,n}, \kappa_d), \]  
(6)

where \( \kappa_d \) is the concentration of the state dynamics. \( \Theta = \{\kappa_y, \kappa_d, \pi_0, ..., \pi_N\} \) denotes the set of model parameters.

As already mentioned in Section I, the filtering distribution corresponds to a mixture model whose number of components grows exponentially along time, therefore solving (1) directly is computationally intractable. Below we infer a variational approximation of (1) which drastically reduces the explosion of the number of mixture components; consequently, it leads to a computationally tractable algorithm.

III. VARIATIONAL APPROXIMATION AND ALGORITHM

Since solving (1) is computationally intractable, we propose to approximate the conditional independence between the states and the assignment variables, more precisely

\[ p(s_t, z_t | y_{1:t}) \approx q(s_t)q(z_t). \]  
(7)

The proposed factorization leads to a VEM algorithm [15], where the posterior distribution of the two variables are found by two variational E-steps:

\[ q(z_t) \propto \exp \left\{ \mathcal{E}_q(s_t) \log p(s_t, z_t | y_{1:t}) \right\}, \]  
(8)

\[ q(s_t) \propto \exp \left\{ \mathcal{E}_q(z_t) \log p(s_t, z_t | y_{1:t}) \right\}. \]  
(9)

The model parameters \( \Theta \) are estimated by maximizing the expected complete-data log-likelihood:

\[ Q(\Theta, \tilde{\Theta}) = \mathcal{E}_q(s_t, z_t) \left\{ \log p(y_t, s_t, z_t | y_{1:t-1}, \Theta) \right\}. \]  
(10)

where \( \tilde{\Theta} \) are the old parameters. To illustrate the impact of the proposed approximation on the filtering distribution, we observe that (i) the posterior pdf of the assignment is observation-independent, and that (ii) the posterior pdf of the state variables is source-independent, i.e.,

\[ q(z_t) = \prod_{m=1}^{M_t} q(z_{tm}), \quad q(s_t) = \prod_{n=1}^{N} q(s_{tn}). \]  
(11)

Therefore, the predictive pdf is also separable:

\[ p(s_{tn} | y_{1:t-1}) = \int p(s_{tn} | s_{t-1,n})p(s_{t-1,n} | y_{1:t-1})ds_{t-1,n}. \]  

Moreover, assuming that the filtering pdf at time \( t-1 \) follows a von Mises distribution, i.e. \( q(s_{t-1,n}) = \]
\[ M(s_{t-1}; \mu_{t-1}, \kappa_{t-1}) \] then the predictive pdf is approximately a von Mises distribution (see [6], [16, (3.5.43)]):
\[ p(s_{tm}|y_{1:t-1}) \approx M(s_{tm}; \mu_{t-1}, \hat{\kappa}_{t-1}) \]
where the predicted concentration parameter, \( \hat{\kappa}_{t-1} \), is:
\[ \hat{\kappa}_{t-1} = A^{-1}(A(\kappa_{t-1})A(\kappa_d)) \]
and where \( A(a) = I_1(a)/I_0(a) \), and \( A^{-1}(a) \approx (2a - a^3)/(1 - a^2) \). Using (8), (9) and (10), the filtering distribution is therefore obtained by iterating through three steps, i.e. the E-S, E-Z and M steps, provided below (detailed mathematical derivations can be found in the appendices).

1) E-S step: Inserting (1) and (12) in (9), \( q(s_{tn}) \) reduces to a von Mises distribution, \( M(s_{tn}; \mu_{tn}, \kappa_{tn}) \). The mean \( \mu_{tn} \) and concentration \( \kappa_{tn} \) are given by:
\[ \mu_{tn} = \tan^{-1}\left(\frac{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn} \omega_{tm} \sin(y_{tm}) + \hat{\kappa}_{t-1,n} \sin(\mu_{t-1,n})}{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn} \omega_{tm} \cos(y_{tm}) + \hat{\kappa}_{t-1,n} \cos(\mu_{t-1,n})}\right) \]
\[ \kappa_{tn} = \left(\kappa_y^2 \sum_{m=1}^{M_t} (\alpha_{tmn} \omega_{tm})^2 + \hat{\kappa}_{t-1,n}^2 \right)
+ 2(\kappa_y^2) \sum_{m=1}^{M_t} \sum_{l=m+1}^{M_t} \alpha_{tmn} \omega_{tm} \alpha_{tln} \omega_{tl} \cos(y_{tm} - y_{tl})
+ 2\kappa_y \hat{\kappa}_{t-1,n} \sum_{m=1}^{M_t} \alpha_{tmn} \omega_{tm} \cos(\mu_{tm} - \mu_{t-1,n}) \right)^{1/2}, \]
where \( \alpha_{tmn} = q(Z_{tm} = n) \) denotes the variational posterior probability of the assignment variable. Therefore, the expressability of the posterior distribution as a mixture of von Mises propagates over time, and only needs to be assumed at \( t = 1 \). Please consult the supplementary materials for more details.

2) E-Z step: By computing the expectation over \( s_t \) in (8), the following expression is obtained:
\[ \alpha_{tmn} = q(z_{tm} = n) = \frac{\pi_n \beta_{tmn}}{\sum_{l=0}^{N} \pi_l \beta_{tml}} \]
where \( \beta_{tmn} \) is given by (please consult the supplementary materials for a detailed derivation):
\[ \beta_{tmn} = \frac{\omega_{tm} \kappa_y A(\omega_{tm} \kappa_y) \cos(\mu_{tm} - \mu_{tn})}{2\pi} \]
\[ \beta_{tnl} = \frac{\omega_{tm} \kappa_y A(\omega_{tm} \kappa_y) \cos(\mu_{tm} - \mu_{tn})}{2\pi} \]

3) M step: The parameter set \( \Theta \) are evaluated by maximizing (10). The priors (4) are obtained using the conventional update rule [15]: \( \pi_n \propto \sum_{m=1}^{M_t} \alpha_{tmn} \). The concentration parameters, \( \kappa_y \) and \( \kappa_d \), are evaluated using gradient descent, namely (please consult the supplementary materials):
\[ \frac{\partial Q}{\partial \kappa_y} = \sum_{m=1}^{M_t} \alpha_{tmn} \omega_{tm} \left( A(\kappa_{tn}) \cos(\mu_{tm} - y_{tm}) - A(\kappa_y \omega_{tm}) \right) \]
\[ \frac{\partial Q}{\partial \kappa_d} = \sum_{n=1}^{N} \left( A(\kappa_{tn}) \cos(\mu_{tn} - \mu_{t-1,n}) - A(\hat{\kappa}_{t-1,n}) \right) B(\kappa_d). \]

where \( B(\kappa_d) = \frac{\partial \kappa_{t-1,n}}{\partial \kappa_d} \).

Based on the E-S, E-Z and M step formulas above, the proposed VEM algorithm iterates until convergence at each time step, in order to estimate the posterior distributions and to update the values of the model parameters.

IV. AUDIO-SOURCE BIRTH PROCESS

We now describe in detail the proposed birth process which is essential to initialize the number of audio sources as well as to detect new sources at any time. The birth process gathers all the DOAs that were not assigned to a source, i.e. assigned to \( n = 0 \), at current frame \( t \) as well over the \( L \) previous frames (\( L = 2 \) in all our experiments). From this set of DOAs we build DOA/observation sequences (one observation per frame) and let \( \hat{y}_{t-L:L} \) be such a sequence of DOAs, where \( j \) is the sequence index. We consider the marginal likelihood:
\[ \tau_j = p(\hat{y}_{t-L:L}^j) = \int p(\hat{y}_{t-L:L}^j; s_{t-L:L}) ds_{t-L:L}. \]

Using (12) and the harmonic sum theorem, the integral (17) becomes (please consult the supplementary materials):
\[ \tau_j = \prod_{l=0}^{L} \frac{I_0(\sigma_{t-l}^j)}{2\pi I_0(\kappa_y \hat{\omega}_{t-l}^j) I_0(\hat{\kappa}_{t-l}^j)}, \]
where \( \hat{\omega}_t \) is the confidence associated with \( \hat{y}_t \). The concentration parameters, \( \pi_{t-l}^j \) and \( \hat{\kappa}_{t-l}^j \), depend on the observations and are recursively computed for each sequence \( j \):
\[ \pi_{t-l}^j = \sqrt{\left(\hat{\kappa}_{t-l}^j\right)^2 + (\kappa_y \hat{\omega}_{t-l}^j)^2 + \hat{\kappa}_{t-l}^j \kappa_y \hat{\omega}_{t-l}^j \cos(\hat{y}_{t-l}^j - \hat{\mu}_{t-l}^j)}, \]
\[ \hat{\mu}_{t-l+1} = \tan^{-1}\left(\frac{\hat{\kappa}_{t-l}^j \sin(\hat{\mu}_{t-l}^j) + \kappa_y \hat{\omega}_{t-l}^j \sin(\hat{y}_{t-l}^j)}{\hat{\kappa}_{t-l}^j \cos(\hat{\mu}_{t-l}^j) + \kappa_y \hat{\omega}_{t-l}^j \cos(\hat{y}_{t-l}^j)}\right), \]
\[ \hat{\kappa}_{t-l+1} = A^{-1}(A(\hat{\kappa}_{t-l}^j)A(\kappa_d)). \]
The sequence \( j^* \) with the maximal marginal likelihood (18), namely \( j^* = \arg\max_j(\tau_j) \), is supposed to be generated from a not yet known audio source only if \( \tau_{j^*} \) is larger than a threshold: a new source \( \hat{n} \) is created in this case and \( q(s_{t\hat{n}}) = M(s_{t\hat{n}}; \hat{\mu}_{tj^*}, \hat{\kappa}_{tj^*}) \).

V. EXPERIMENTAL EVALUATION

The proposed method was evaluated using the audio recordings from Task 6 of the IEEE-AASP LOCATA challenge development dataset [17], which involve multiple moving sound sources and moving microphone arrays. The LOCATA dataset consists of real-life recordings with ground-truth source locations provided by an optical tracking system. The size of the recording room is 7.1 × 9.8 × 3 m, with \( T_{st} \approx 0.55 \) s. Task 6 contains three sequences of of a total duration of 188.4 s and two moving speakers. In our experiments we used microphones 5, 8, 11, and 12, embedded in a robot head and located on a horizontal plane. The online sound-source localization method [14] was used to provide DOA estimates at each STFT frame. The STFT uses the Hamming window with
TABLE I: Method evaluation with the LOCATA dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>MD (%)</th>
<th>FA (%)</th>
<th>MAE (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vM-PHD [12]</td>
<td>33.4</td>
<td>9.5</td>
<td>4.5</td>
</tr>
<tr>
<td>GM-ZO [14]</td>
<td>27.0</td>
<td>10.8</td>
<td>4.7</td>
</tr>
<tr>
<td>GM-FO [14]</td>
<td>22.3</td>
<td>6.3</td>
<td>3.2</td>
</tr>
<tr>
<td>vM-VEM (proposed)</td>
<td>23.9</td>
<td>5.9</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Method MD (%) FA (%) MAE (°)

a length of 16 ms and 8 ms shifts. Peak detection is applied with a threshold of 0.3. Both DOA estimates and associated confidence measures, $w_{tm}$, are provided at each frame.

To evaluate the method quantitatively, the estimated source trajectories are compared with the ground-truth trajectories over audio-active frames. Ground-truth audio-active frames are obtained using the voice activity detection (VAD) of [18]. The permutation problem between the detected trajectories and the ground-truth trajectories is solved by means of a greedy gating algorithm: the error between all possible pairs of estimated and ground-truth trajectories is evaluated. Minimum-error pairs are selected for further comparison. A DOA estimate that is 15° away from the ground-truth is treated as false detection. Sources that are not associated with a trajectory correspond to missed detections (MDs). For performance evaluation, the percentage of MDs and false alarms (FAs) are evaluated over voice-active frames. The mean absolute error (MAE) denotes the average absolute error of all the correct associations between all the sources and all the voice-active frames.

The observation-to-source assignment posteriors and the DOAs confidence weights are used to estimate voice-active frames:

$$
\sum_{t'=t-D}^{t} \sum_{m=1}^{M_t} a_{t'm} w_{tm} \begin{cases} \text{active} & \frac{\delta_{t'm}}{\text{silent}} \\ \end{cases} \delta
$$

(19)

where $D = 2$ and $\delta = 0.025$ is a VAD threshold. Once an active source is detected, we output its trajectory.

The MAEs, MDs and FAs values, averaged over all recordings, are summarized in Table I. We compared the proposed von Mises VEM algorithm (vM-VEM) with three multi-speaker trackers: the von Mises PHD filter (vM-PHD) [12] and two versions the multiple speaker tracker of [14] based on Gaussians models (GM). [14] uses a first-order dynamic model whose effect is to smooth the estimated trajectories. We compared with both first-order (GM-FO) and zero-order (GM-ZO) dynamics. The proposed vM-VEM tracker yields the lowest false alarm (FA) rate of 5.9% and MAE of 3.2, and the second lowest MD rate of 23.9%. The GM-FO variant of [14] yields an MD rate of 22.3% since it uses velocity information to smooth the trajectories. This illustrates the advantage of the von-Mises distribution to model directional data (DOA). The proposed von-Mises model uses a zero-order dynamics; nevertheless it achieves performance comparable with the Gaussian model that uses first-order dynamics.

The results for recordings #1 and #2 in Task 6 are shown in Fig. 1. Note that the PHD-based filter method [12] has two caveats. First, observation-to-source assignments cannot be estimated (unless a post-processing step is performed), and second, the estimated source trajectories are not smooth. This stays in contrast with the proposed method which explicitly represents assignments with discrete latent variables and estimates them iteratively with VEM. Moreover, the proposed method yields smooth trajectories similar with those estimated by [14] and quite close to the ground truth.

VI. CONCLUSION

We proposed a multiple audio-source tracking method using the von Mises distribution and we inferred a tractable solver based on a variational approximation of the posterior filtering distribution. Unlike the Gaussian distribution, the von Mises distribution explicitly models the circular variables associated with audio-source localization and tracking based on source DOAs. Using the recently released LOCATA dataset, we empirically showed that the proposed method compares favorably with two recent methods.

APPENDIX

Please consult the supplemental materials.
REFERENCES


