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

Yutong Ban,[1] Xavier Alameda-Pineda,[1] *Member, IEEE*,
Christine Evers,[2] *Senior Member, IEEE*, and Radu Horaud[1]

***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].

[1]Y. Ban, X. Alameda-Pineda and R. Horaud are with Inria Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: first.last@inria.fr

[2]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.

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

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 $s_{tn} \in (-\pi, \pi]$ be the DOA of source $n \in \{1, \dots, N\}$ at t, and $\boldsymbol{s}_t = (s_{t1}, \dots, s_{tN})$. Furthermore, let $y_{tm} \in (-\pi, \pi]$ denote the m-th DOA observation at time t, where $m \in \{1, \dots, M_t\}$ and M_t is the number of observations at t. Each observation is accompanied by its corresponding confidence $\omega_{tm} \in [0, 1]$. Let z_{tm} denote the observation-to-source assignment variable, where $z_{tm} = n$ means that y_{tm} is an observation of source n, and $z_{tm} = 0$ means that y_{tm} is an observation of a dummy source.

Within a Bayesian model, multiple target tracking can be formulated as the estimation of the filtering distribution $p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t})$. We assume that the state variables s_{tn} 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:

$$p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t}) \propto p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t) p(\boldsymbol{z}_t) p(\boldsymbol{s}_t | \boldsymbol{y}_{1:t-1}), \quad (1)$$

where $p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t)$ is the observation likelihood, $p(\boldsymbol{z}_t)$ is the prior pdf of the assignment variables and $p(\boldsymbol{s}_t | \boldsymbol{y}_{1:t-1})$ 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:

$$p(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{s}_t) = \prod_{m=1}^{M_t} p(y_{tm} | \boldsymbol{z}_t, \boldsymbol{s}_t). \quad (2)$$

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:

$$p(y_{tm} | z_{tm} = n, s_{tn}) = \begin{cases} \mathcal{M}(y_{tm}; s_{tn}, \kappa_y \omega_{tm}) & n \neq 0 \\ \mathcal{U}(y_{tm}) & n = 0 \end{cases}, \quad (3)$$

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 κ, $I_p(\cdot)$ denotes the modified Bessel function of the first kind of order p, κ_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:

$$p(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} p(Z_{tm} = n), \quad (4)$$

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 y_{tm}.

[1] https://team.inria.fr/perception/research/audiotrack-vonm/

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.,

$$p(\boldsymbol{s}_t | \boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{s}_t | \boldsymbol{s}_{t-1}) p(\boldsymbol{s}_{t-1} | \boldsymbol{y}_{1:t-1}) d\boldsymbol{s}_{t-1}. \quad (5)$$

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:

$$p(\boldsymbol{s}_t | \boldsymbol{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{M}(s_{tn}; s_{t-1,n}, \kappa_d), \quad (6)$$

where κ_d is the concentration of the state dynamics. $\Theta = \{\kappa_y, \kappa_d, \pi_0, \dots, \pi_N\}$ 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

$$p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t}) \approx q(\boldsymbol{s}_t) q(\boldsymbol{z}_t). \quad (7)$$

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

$$q(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t})\Big), \quad (8)$$

$$q(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t})\Big). \quad (9)$$

The model parameters Θ are estimated by maximizing the expected complete-data log-likelihood:

$$Q(\Theta, \tilde{\Theta}) = \mathrm{E}_{q(\boldsymbol{s}_t) q(\boldsymbol{z}_t)} \Big\{ \log p(\boldsymbol{y}_t, \boldsymbol{s}_t, \boldsymbol{z}_t | \boldsymbol{y}_{1:t-1}, \Theta) \Big\}. \quad (10)$$

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.,

$$q(\boldsymbol{z}_t) = \prod_{m=1}^{M_t} q(z_{tm}), \quad q(\boldsymbol{s}_t) = \prod_{n=1}^{N} q(s_{tn}). \quad (11)$$

Therefore, the predictive pdf is also separable:

$$p(s_{tn} | \boldsymbol{y}_{1:t-1}) = \int p(s_{tn} | s_{t-1,n}) p(s_{t-1,n} | \boldsymbol{y}_{1:t-1}) \mathrm{d}s_{t-1,n}.$$

Moreover, assuming that the filtering pdf at time $t-1$ follows a von Mises distribution, i.e. $q(s_{t-1,n}) =$

$\mathcal{M}(s_{t-1,n};\mu_{t-1,n},\kappa_{t-1,n})$, then the predictive pdf is approximately a von Mises distribution (see [6], [16, (3.5.43)]):

$$p(s_{tn}|\boldsymbol{y}_{1:t-1}) \approx \mathcal{M}(s_{tn};\mu_{t-1,n},\tilde{\kappa}_{t-1,n}), \tag{12}$$

where the predicted concentration parameter, $\tilde{\kappa}_{t-1,n}$, is:

$$\tilde{\kappa}_{t-1,n} = A^{-1}(A(\kappa_{t-1,n})A(\kappa_d)), \tag{13}$$

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, $\mathcal{M}(s_{tn};\mu_{tn},\kappa_{tn})$. The mean μ_{tn} and concentration κ_{tn} are given by:

$$\mu_{tn} = \tan^{-1}\left(\frac{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn}\omega_{tm}\sin(y_{tm}) + \tilde{\kappa}_{t-1,n}\sin(\mu_{t-1,n})}{\kappa_y \sum_{m=1}^{M_t} \alpha_{tmn}\omega_{tm}\cos(y_{tm}) + \tilde{\kappa}_{t-1,n}\cos(\mu_{t-1,n})}\right), \tag{14}$$

$$\kappa_{tn} = \Bigg((\kappa_y)^2 \sum_{m=1}^{M_t}(\alpha_{tmn}\omega_{tm})^2 + \tilde{\kappa}_{t-1,n}^2 + 2(\kappa_y)^2 \sum_{m=1}^{M_t}\sum_{l=m+1}^{M_t} \alpha_{tmn}\omega_{tm}\alpha_{tln}w_{tl}\cos(y_{tm}-y_{tl}) + 2\kappa_y\tilde{\kappa}_{t-1,n}\sum_{m=1}^{M_t}(\alpha_{tmn}\omega_{tm}\cos(y_{tm}-\mu_{t-1,n}))\Bigg)^{1/2}, \tag{15}$$

where $\alpha_{tmn} = q(Z_{tm} = n)$ denotes the variational posterior probability of the assignment variable. Therefore, the expressibility of the posterior distribution as a mixture of von Misses propagates over time, and only needs to be assumed at $t = 1$. Please consult Appendix A for more details.

2) E-Z step: By computing the expectation over $\boldsymbol{s}_t$ in (8), the following expression is obtained:

$$\alpha_{tmn} = q(z_{tm} = n) = \frac{\pi_n \beta_{tmn}}{\sum_{l=0}^{N} \pi_l \beta_{tml}} \tag{16}$$

where β_{tmn} is given by (please consult Appendix B for a detailed derivation):

$$\beta_{tmn} = \begin{cases} \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y)\cos(y_{tm}-\mu_{tn}) & n \neq 0 \\ 1/2\pi & n = 0, \end{cases}$$

3) M step: The parameter set Θ are evaluated by maximizing (10). The priors (4) are obtained using the conventional update rule [15]: $\pi_n \propto \sum_{m=1}^{M_t}\alpha_{tnm}$. The concentration parameters, κ_y and κ_d, are evaluated using gradient descent, namely (please consult Appendix C):

$$\frac{\partial Q}{\partial \kappa_y} = \sum_{m,n=1}^{M_t,N} \alpha_{tnm}\omega_{tm}\Big(A(\kappa_{tn})\cos(\mu_{tn}-y_{tm}) - A(\kappa_y\omega_{tm})\Big)$$

$$\frac{\partial Q}{\partial \kappa_d} = \sum_{n=1}^{N}\Big(A(\kappa_{tn})\cos(\mu_{tn}-\mu_{t-1,n}) - A(\tilde{\kappa}_{t-1,n})\Big)B(\kappa_d).$$

where $B(\kappa_d) = \frac{\partial \tilde{\kappa}_{t-1,n}}{\partial \kappa_d}$.

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 $\hat{y}^j_{t-L:t}$ be such a sequence of DOAs, where j is the sequence index. We consider the marginal likelihood:

$$\tau_j = p(\hat{y}^j_{t-L:t}) = \int p(\hat{y}^j_{t-L:t}, s_{t-L:t})\mathrm{d}s_{t-L:t}. \tag{17}$$

Using (12) and the harmonic sum theorem, the integral (17) becomes (please consult Appendix D):

$$\tau_j = \prod_{l=0}^{L} \frac{I_0(\overline{\kappa}^j_{t-l})}{2\pi I_0(\kappa_y\hat{\omega}^j_{t-l})I_0(\hat{\kappa}^j_{t-l})}, \tag{18}$$

where $\hat{\omega}_t$ is the confidence associated with $\hat{y}_t$. The concentration parameters, $\overline{\kappa}^j_{t-l}$ and $\hat{\kappa}^j_{t-l+1}$, depend on the observations and are recursively computed for each sequence j:

$$\overline{\kappa}^j_{t-l} = \sqrt{(\hat{\kappa}^j_{t-l})^2 + (\kappa_y\hat{\omega}^j_{t-l})^2 + \hat{\kappa}^j_{t-l}\kappa_y\hat{\omega}^j_{t-l}\cos(\hat{y}^j_{t-l} - \hat{\mu}^j_{t-l})},$$

$$\hat{\mu}^j_{t-l+1} = \tan^{-1}\left(\frac{\hat{\kappa}^j_{t-l}\sin(\hat{\mu}^j_{t-l}) + \kappa_y\hat{\omega}^j_{t-l}\sin(\hat{y}^j_{t-l})}{\hat{\kappa}^j_{t-l}\cos(\hat{\mu}^j_{t-l}) + \kappa_y\hat{\omega}^j_{t-l}\cos(\hat{y}^j_{t-l})}\right),$$

$$\hat{\kappa}^j_{t-l+1} = A^{-1}(A(\tilde{\kappa}^j_{t-l})A(\kappa_d)).$$

The sequence j^* with the maximal marginal likelihood (18), namely $j^* = \mathrm{argmax}_j(\tau_j)$, is supposed to be generated from a not yet known audio source only if τ_{j^*} is larger than a threshold: a new source $\tilde{n}$ is created in this case and $q(s_{t\tilde{n}}) = \mathcal{M}(s_{t\tilde{n}};\hat{\mu}_{tj^*},\hat{\kappa}_{tj^*})$.

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 $7.1 \times 9.8 \times 3$ m, with $T_{60} \approx 0.55$ 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

Method	MD (%)	FA (%)	MAE (°)
vM-PHD [12]	33.4	9.5	4.5
GM-ZO [14]	27.0	10.8	4.7
GM-FO [14]	**22.3**	6.3	3.2
vM-VEM (proposed)	23.9	**5.9**	**2.6**

TABLE I: Method evaluation with the LOCATA dataset.

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:

$$\sum_{t'=t-D}^{t} \sum_{m=1}^{M_t} \alpha_{t'mn}\omega_{t'm} \underset{silent}{\overset{active}{\gtrless}} \delta \tag{19}$$

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

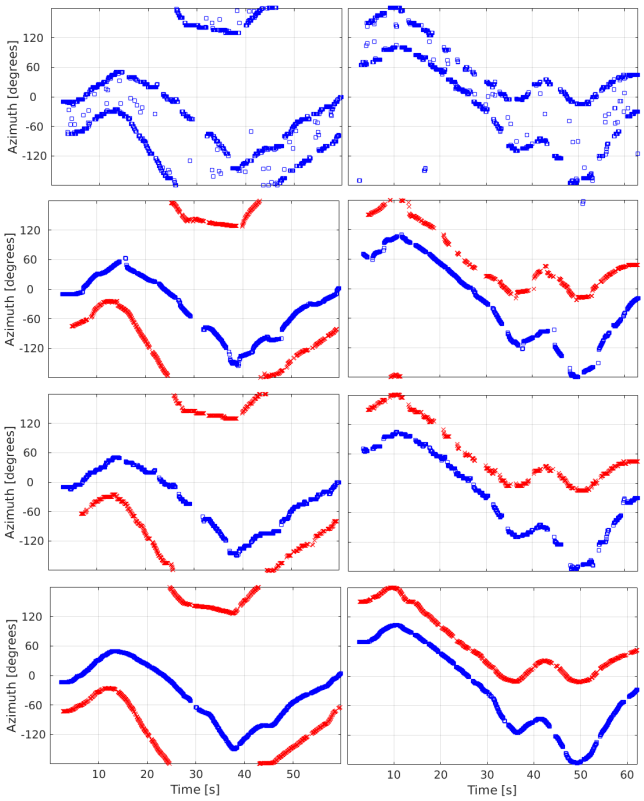

Fig. 1: Results obtained with recordings #1 (left) and #2 (right) from Task 6 of the LOCATA dataset. Top-to-down: vM-PHD [12], GM-FO [14], vM-VEM (proposed) and ground-truth trajectories. Different colors represent different audio sources. Note that vM-PHD is unable to associate sources with trajectories.

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 A
DERIVATION OF THE E-S STEP

In order to obtain the formulae for the E-S step, we start from its definition in (9):

$$q(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t})\Big). \tag{20}$$

We now use the decomposition in (1) to write:

$$q(\boldsymbol{s}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t)\Big) p(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1}). \tag{21}$$

Let us know develop the expectation:

$$\begin{aligned}
&\mathrm{E}_{q(\boldsymbol{z}_t)} \log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t) \\
&= \mathrm{E}_{q(\boldsymbol{z}_t)} \sum_{m=1}^{M_t} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \mathrm{E}_{q(z_{tm})} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \mathrm{E}_{q(z_{tm})} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} q(z_{tm}=n) \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}=n) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm} \log p(y_{tm}|s_{tn}, z_{tm}=n) \\
&= \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm} \log \mathcal{M}(y_{tm}; s_{tn}, \omega_{tm}\kappa_y) \\
&\overset{\boldsymbol{s}_t}{=} \sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}),
\end{aligned}$$

where $\overset{\boldsymbol{s}_t}{=}$ denotes the equality up to an additive constant that does *not* depend on $\boldsymbol{s}_t$. Such a constant would become a multiplicative constant after the exponentiation in (21), and therefore can be ignored.

By replacing the developed expectation together with (12) we obtain:

$$\begin{aligned}
q(\boldsymbol{s}_t) \propto &\exp\Big(\sum_{m=1}^{M_t} \sum_{n=0}^{N} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn})\Big) \\
&\prod_{n=0}^{N} \mathcal{M}(\boldsymbol{s}_{tn}; \mu_{t-1,n}, \tilde{\kappa}_{t-1,n}),
\end{aligned}$$

which can be easily rewritten as:

$$\begin{aligned}
q(\boldsymbol{s}_t) \propto \prod_{n=0}^{N} \exp\Big(&\sum_{m=1}^{M_t} \alpha_{tnm}\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}) \\
&+ \tilde{\kappa}_{t-1,n} \cos(s_{tn} - \mu_{t-1,n})\Big).
\end{aligned}$$

This is important since it demonstrates that the a posteriori distribution on $\boldsymbol{s}_t$ is separable on n and therefore independent for each speaker. In addition, it allows us to rewrite the a posteriori distribution for each speaker, i.e. on s_{tn} as a von Mises distribution by using the harmonic addition theorem, thus obtaining

$$q(\boldsymbol{s}_t) = \prod_{n=0}^{N} q(s_{tn}) = \prod_{n=0}^{N} \mathcal{M}(s_{tn}; \mu_{tn}, \kappa_{tn}), \tag{22}$$

with μ_{tn} and κ_{tn} defined as in (14) and (15).

APPENDIX B
DERIVATION OF THE E-Z STEP

Similarly to the previous section, and in order to obtain the closed-form solution of the E-Z step, we start from its definition in (8):

$$q(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{s}_t, \boldsymbol{z}_t|\boldsymbol{y}_{1:t})\Big), \tag{23}$$

and we use the decomposition in (1) to write:

$$q(\boldsymbol{z}_t) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(\boldsymbol{y}_t|\boldsymbol{s}_t, \boldsymbol{z}_t)\Big) p(\boldsymbol{z}_t). \tag{24}$$

Since both the observation likelihood and the prior distribution are separable on z_{tm}, we can write:

$$q(\boldsymbol{z}_t) \propto \prod_{m=1}^{M_t} \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm})\Big) p(z_{tm}), \tag{25}$$

proving that the a posteriori distribution is also separable on m.

We can thus analyze the posterior of each z_{tm} separately, by computing $q(z_{tm}=n)$:

$$q(z_{tm}=n) \propto \exp\Big(\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}=n)\Big) p(z_{tm}=n)$$

Let us first compute the expectation for $n \neq 0$:

$$\begin{aligned}
&\mathrm{E}_{q(\boldsymbol{s}_t)} \log p(y_{tm}|\boldsymbol{s}_t, z_{tm}=n) \\
&= \mathrm{E}_{q(s_{tn})} \log p(y_{tm}|s_{tn}, z_{tm}=n) \\
&= \mathrm{E}_{q(s_{tn})} \log \mathcal{M}(y_{tm}; s_{tn}, \omega_{tm}\kappa_y) \\
&\overset{z_{tm}}{=} \int_0^{2\pi} q(s_{tn})\omega_{tm}\kappa_y \cos(y_{tm} - s_{tn}) \mathrm{d}s_{tn} \\
&= \frac{\omega_{tm}\kappa_y}{2\pi I_0(\omega_{tm}\kappa_y)} \int_0^{2\pi} \exp\Big(\cos(s_{tn} - \mu_{tn})\Big) \cos(s_{tn} - y_{tm}) \mathrm{d}s_{tn} \\
&= \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y) \cos(y_{tm} - \mu_{tn}),
\end{aligned}$$

where for the last line we used the following variable change $\bar{s} = s_{tn} - \mu_{tn}$ and the definition of I_1 and A.

The case $n = 0$ is even easier since the observation distribution is a uniform: $\mathrm{E}_{q(s_{tn})} \log p(y_{tm}|s_{tn}, z_{tm} = n) = \mathrm{E}_{q(s_{tn})} - \log 2\pi = -\log(2\pi)$.

By using the fact that the prior distribution on z_{tm} is denoted by $p(z_{tm} = n) = \pi_n$, we can now write the a posteriori distribution as $q(z_{tm} = n) \propto \pi_n \beta_{tmn}$ with:

$$\beta_{tmn} = \begin{cases} \omega_{tm}\kappa_y A(\omega_{tm}\kappa_y) \cos(y_{tm} - \mu_{tn}) & n \neq 0 \\ 1/2\pi & n = 0 \end{cases},$$

and finally obtaining the results in (16) and (3).

APPENDIX C
DERIVATION OF THE M STEP

In order to derive the M step, we need first to compute the Q function in (10),

$$\begin{aligned} Q(\Theta,\tilde{\Theta}) &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log p(\boldsymbol{y}_t,\boldsymbol{s}_t,\boldsymbol{z}_t|\boldsymbol{y}_{1:t-1},\Theta)\Big\} \\ &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\underbrace{\log p(\boldsymbol{y}_t|\boldsymbol{s}_t,\boldsymbol{z}_t,\Theta)}_{\kappa_y}+ \\ &= +\underbrace{\log p(\boldsymbol{z}_t|\Theta)}_{\pi_n' s}+\underbrace{\log p(\boldsymbol{s}_t|\boldsymbol{y}_{1:t-1},\Theta)}_{\kappa_d}\Big\}, \end{aligned}$$

where each parameter is show below the corresponding term of the Q function. Let us develop each term separately.

A. Optimizing κ_y

$$\begin{aligned} Q_{\kappa_y} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log\prod_{m=1}^{M_t}p(y_{tm}|\boldsymbol{s}_t,z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t}\mathrm{E}_{q(\boldsymbol{s}_t)q(z_{tm})}\Big\{\log p(y_{tm}|\boldsymbol{s}_t,z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t}\mathrm{E}_{q(\boldsymbol{s}_t)}\sum_{n=0}^{N}\alpha_{tmn}\Big\{\log p(y_{tm}|\boldsymbol{s}_t,z_{tm}=n)\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\mathrm{E}_{q(s_{tn})}\Big\{\log\mathcal{M}(y_{tm};s_{tn},\omega_{tm}\kappa_y)\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\int_0^{2\pi}q(s_{tn})(\omega_{tm}\kappa_y\cos(y_{tm}-s_{tn})-\log(I_0(\omega_{tm}\kappa_y))\mathrm{d}s_{tn} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\Big(\omega_{tm}\kappa_y\cos(y_{tm}-\mu_{tn})A(\kappa_{tn})-\log(I_0(\omega_{tm}\kappa_y))\Big), \end{aligned}$$

and by taking the derivative with respect to κ_y we obtain:

$$\frac{\partial Q}{\partial\kappa_y}=\sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\omega_{tm}\Big(\cos(y_{tm}-\mu_{tn})A(\kappa_{tn})-A(\omega_{tm}\kappa_y)\Big),$$

which corresponds to what was announced in the manuscript.

B. Optimizing π_n's

$$\begin{aligned} Q_{\pi_n} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log\prod_{m=1}^{M_t}p(z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t}\mathrm{E}_{q(z_{tm})}\Big\{\log p(z_{tm})\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\Big\{\log p(z_{tm}=n)\Big\} \\ &= \sum_{m=1}^{M_t}\sum_{n=0}^{N}\alpha_{tmn}\Big\{\log\pi_n\Big\} \end{aligned}$$

This is the very same formulae that for any mixture model, and therefore the solution is standard and corresponds to the one reported in the manuscript.

C. Optimizing κ_d

$$\begin{aligned} Q_{\kappa_d} &= \mathrm{E}_{q(\boldsymbol{s}_t)q(\boldsymbol{z}_t)}\Big\{\log\prod_{n=1}^{N}p(s_{tn}|\boldsymbol{y}_{1:t-1})\Big\} \\ &= \sum_{n=1}^{N}\mathrm{E}_{q(s_{tn})}\Big\{\log\mathcal{M}(s_{tn};\mu_{t-1,n},\tilde{\kappa}_{t-1,n})\Big\} \\ &= \sum_{n=1}^{N}\mathrm{E}_{q(s_{tn})}\Big\{-\log I_0(\tilde{\kappa}_{t-1,n})+\tilde{\kappa}_{t-1,n}\cos(s_{tn}-\mu_{t-1,n})\Big\} \\ &= \sum_{n=1}^{N}-\log I_0(\tilde{\kappa}_{t-1,n})+\tilde{\kappa}_{t-1,n}\cos(\mu_{tn}-\mu_{t-1,n})A(\kappa_{tn}), \end{aligned}$$

where the dependency on κ_d is implicit in $\tilde{\kappa}_{t-1,n} = A^{-1}(A(\kappa_{t-1,n})A(\kappa_d))$.

By taking the derivative with respect to κ_d we obtain:

$$\frac{\partial Q}{\partial\kappa_d}=\sum_{n=1}^{N}\Big(A(\kappa_{tn})\cos(\mu_{tn}-\mu_{t-1,n})-A(\tilde{\kappa}_{t-1,n})\Big)\frac{\partial\tilde{\kappa}_{t-1,n}}{\partial\kappa_d}$$

with

$$\frac{\partial\tilde{\kappa}_{t-1,n}}{\partial\kappa_d}=\tilde{A}(A(\kappa_{t-1,n})A(\kappa_d))A(\kappa_{t-1,n})\frac{I_2(\kappa_d)I_0(\kappa_d)-I_1^2(\kappa_d)}{I_0^2(\kappa_d)},$$

where $\tilde{A}(a)=\mathrm{d}A^{-1}(a)/\mathrm{d}a=(2-a^2+a^4)/(1-a^2)^2$.

By denoting the previous derivative as $B(\kappa_d)=\frac{\partial\tilde{\kappa}_{t-1,n}}{\partial\kappa_d}$, we obtain the expression in the manuscript.

APPENDIX D
DERIVATION OF THE BIRTH PROBABILITY

In this section we derive the expression for τ_j by computing the integral (17). Using the probabilistic model defined, we can write (the index j is omitted):

$$\begin{aligned} &\int p(\hat{y}_{t-L:t},s_{t-L:t})\mathrm{d}s_{t-L:t} \\ &\quad=\int\prod_{\tau=-L}^{0}p(\hat{y}_{t+\tau}|s_{t+\tau})\prod_{\tau=-L+1}^{0}p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L})\mathrm{d}s_{t-L:t} \end{aligned}$$

We will first marginalize s_{t-L}. To do that, we notice that if $p(s_{t-L})$ follows a von Mises with mean $\hat{\mu}_{t-L}$ and concentration $\hat{\kappa}_{t-L}$, then we can write:

$$\begin{aligned} &p(\hat{y}_{t-L}|s_{t-L})p(s_{t-L}) \\ &\quad=\mathcal{M}(\hat{y}_{t-L};s_{t-L},\hat{\omega}_{t-L}\kappa_y)\mathcal{M}(s_{t-L};\hat{\mu}_{t-L},\hat{\kappa}_{t-L}) \\ &\quad=\mathcal{M}(s_{t-L};\bar{\mu}_{t-L},\bar{\kappa}_{t-L})\frac{I_0(\bar{\kappa}_{t-L})}{2\pi I_0(\hat{\omega}_{t-L}\kappa_y)I_0(\hat{\kappa}_{t-L})} \end{aligned}$$

with

$$\begin{aligned} \bar{\mu}_{t-L} &= \tan^{-1}\left(\frac{\hat{\omega}_{t-L}\kappa_y\sin\hat{y}_{t-L}+\hat{\kappa}_{t-L}\sin\hat{\mu}_{t-L}}{\hat{\omega}_{t-L}\kappa_y\cos\hat{y}_{t-L}+\hat{\kappa}_{t-L}\cos\hat{\mu}_{t-L}}\right), \\ \bar{\kappa}_{t-L}^2 &= (\hat{\omega}_{t-L}\kappa_y)^2+\hat{\kappa}_{t-L}^2+2\hat{\omega}_{t-L}\kappa_y\hat{\kappa}_{t-L}\cos(\hat{y}_{t-L}-\hat{\mu}_{t-L}), \end{aligned}$$

where we used the harmonic addition theorem.

Now we can effectively compute the marginalization. The two terms involving s_{t-L} are:

$$\begin{aligned}&\int \mathcal{M}(s_{t-L+1}; s_{t-L}, \kappa_d)\mathcal{M}(s_{t-L}; \bar{\mu}_{t-L}, \bar{\kappa}_{t-L})\\ &\approx \mathcal{M}(s_{t-L+1}; \hat{\mu}_{t-L+1}, \hat{\kappa}_{t-L+1})\end{aligned}$$

with

$$\begin{aligned}\hat{\mu}_{t-L+1} &= \bar{\mu}_{t-L},\\ \hat{\kappa}_{t-L+1} &= A^{-1}(A(\bar{\kappa}_{t-L})A(\kappa_d)).\end{aligned}$$

Therefore, the marginalization with respect to s_{t-L} yields the following result:

$$\begin{aligned}&\int p(\hat{y}_{t-L:t}, s_{t-L:t})\mathrm{d}s_{t-L:t}\\ &= \int \prod_{\tau=-L}^{0} p(\hat{y}_{t+\tau}|s_{t+\tau}) \prod_{\tau=-L+1}^{0} p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L})\mathrm{d}s_{t-L:t}\\ &= \frac{I_0(\bar{\kappa}_{t-L})}{2\pi I_0(\hat{\omega}_{t-L}\kappa_y)I_0(\hat{\kappa}_{t-L})} \int \prod_{\tau=-L+1}^{0} p(\hat{y}_{t+\tau}|s_{t+\tau})\times\\ &\qquad \prod_{\tau=-L+2}^{0} p(s_{t+\tau}|s_{t+\tau-1})p(s_{t-L+1})\mathrm{d}s_{t-L+1:t}.\end{aligned}$$

Since we have already seen that $p(s_{t-L+1})$ is also a von Mises distribution, we can use the same reasoning to marginalize with respecto to s_{t-L+1}. This strategy yields to the recursion presented in the main text.

REFERENCES

[1] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2001, pp. 3021–3024.

[2] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 826–836, 2003.

[3] X. Zhong and J. R. Hopgood, "Particle filtering for TDOA based acoustic source tracking: Nonconcurrent multiple talkers," *Signal Processing*, vol. 96, pp. 382–394, 2014.

[4] D. Bechler, M. Grimm, and K. Kroschel, "Speaker tracking with a microphone array using Kalman filtering," *Advances in Radio Science*, vol. 1, no. B. 3, pp. 113–117, 2003.

[5] 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.

[6] C. Evers, E. A. Habets, S. Gannot, and P. A. Naylor, "DoA reliability for distributed acoustic tracking," *IEEE Signal Processing Letters*, 2018.

[7] I. Marković and I. Petrović, "Bearing-only tracking with a mixture of von Mises distributions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 707–712.

[8] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[9] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[10] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–18, 2013.

[11] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[12] I. Marković, J. Ćesić, and I. Petrović, "Von Mises mixture PHD filter," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2229–2233, 2015.

[13] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[14] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environment," *ArXiV preprint*, 2018.

[15] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[16] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009, vol. 494.

[17] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July 2018.

[18] X. Li, R. Horaud, L. Girin, and S. Gannot, "Voice activity detection based on statistical likelihood ratio with adaptive thresholding," in *IEEE International Workshop on Acoustic Signal Enhancement*, 2016, pp. 1–5.