Real-Time Tracking of Multiple Acoustical Sources Utilising Rao-Blackwellised Particle Filtering

Leo McCormack[1], Archontis Politis[2], Simo Särkkä[3] and Ville Pulkki[1]
[1]*Department of Signal Processing and Acoustics, Aalto University*, Espoo, Finland
[2]*Department of Information Technology and Communication Sciences, Tampere University*, Finland
[3]*Department of Electrical Engineering and Automation, Aalto University*, Espoo, Finland
leo.mccormack@aalto.fi

***Abstract*—This paper proposes a system for localising and tracking multiple simultaneous acoustical sound sources in the spherical harmonic domain, intended as a precursor for developing parametric sound-field editors and spatial audio effects. The real-time system comprises a novel combination of a direct-path dominance test, grid-less subspace localisation, and multiple target tracking based on Rao–Blackwellised particle filtering. It has robust multi-source performance and can adapt to sources that vary in number and direction over time. The proposed system was evaluated by using the framework established under the LOCATA 2018 localisation and tracking challenge, and comparing the results with the original submissions to the challenge. The results demonstrate that the proposed system yields the lowest angular error on the horizontal plane for both moving source(s) tasks, and provides results close to the winning submission for the static source(s) tasks. The proposed system also fares well in the other LOCATA evaluation metrics, and, more importantly, does so as a real-time system; i.e. with no requirement for offline post-processing of the tracker data and operating within reasonable computational constraints.**

***Index Terms*—multi-target tracking, source localisation**

I. Introduction

Accompanying direction of arrival (DoA) estimators with data association methods (i.e. trackers) has many applications in the fields of: speech enhancement [1], source separation [2], [3], and acoustic scene analysis. Whereas, in the field of spatial audio processing, solutions have traditionally relied on signal-independent linear combinations of the input microphone array signals to generate the target loudspeaker, binaural, or spatially manipulated output signals. Ambisonics [4] is an example of such a linear framework, based on first encoding microphone recordings into spherical harmonic (SH) signals, allowing for spatial manipulations and effects [5], followed by reproduction to headphones or arbitrary loudspeaker setups [6]. More recently, however, signal-dependent parametric alternatives to sound-field reproduction [7]–[10] have been shown to outperform their linear counterparts in perceptual tests. Many of them are also formulated in the spherical harmonic domain (SHD), but more importantly: all of them rely on DoA estimators applied over time and frequency. However, associating these DoA estimates with their respective sound sources has received minimal attention in this spatial audio reproduction context, despite there being certain applications where source tracking may be useful; such as offering further opportunities for spatial audio effects and sound-field manipulations that go beyond the current state-of-the-art, or as a way to stabilise DoA estimates utilised by existing parametric reproduction methods.

The focus of this work, therefore, was to develop a robust acoustic source tracking solution, which may serve as a precursor for undertaking these aforementioned avenues. The task of multi-source tracking does, however, pose a number of challenges, as a practical system should ideally operate with minimal prior information on the complexity of the sound scene, or the types of sound sources and their number. Additionally, the number of active sound sources may change over time in a musical setting, and some performances may involve moving sources. Methods for tackling the problem of data association include: multiple hypothesis tracking (MHT) [11], joint probabilistic data association (JPDA) [12], [13], and sequential Monte Carlo (SMC) based particle filtering methods [14]–[18] and their Rao-Blackwellised variants [3], [19]–[21]. Applications of these data association methods to acoustic tracking include the systems [22]–[26] that were submitted to the LOCATA 2018 localisation and tracking challenge [27]. In particular, the tracking by a real-time ambisonic-based particle filter (TRAMP) system [24] shared similar scope to this present work, but was limited to first-order SH input and fared less favourably regarding some LOCATA 2018 evaluation metrics compared to the other submissions.

The proposed system looks to overcome these drawbacks by combining the direct-path test of [28] with the high-resolution Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) method from [29]. The DoA estimates are then fed to a particle-filter based on the Rao–Blackwellised Monte Carlo data association (RBMCDA) method [19], [20], which was modified to better suit real-time operation and the intended future applications of the system. It is demonstrated that the proposed system outperforms the TRAMP system [24] with regard to a number of the LOCATA 2018 evaluation metrics [30], and is also not constrained to first-order SH input. To promote reproducible research, the ESPRIT and modified RBMCDA implementations have also been open-sourced[1]. Additionally, a real-time version of the system was developed as a VST audio plugin[2], which may serve as a template for integrating it into other real-time systems.

[1]https://github.com/leomccormack/Spatial_Audio_Framework
[2]http://research.spa.aalto.fi/projects/sparta_vsts/

II. Source detection and Localisation

The estimation of the number of sources, a problem also referred to as *detection* in sensor array processing literature, is commonly based on the analysis of the eigenvalues or eigenvectors of the spatial covariance matrix, or information theoretic criteria; for a comparison of such approaches see e.g. [31]. The approaches detailed in [31] generally err on the side of overestimating the source number. Alternatively, direct-path dominance testing (DPD-T) [28] may be employed, which is less permissive, but potentially more effective at isolating dominant sound sources under challenging conditions; such as high noise and reverberation. Once the number of sources is known, subspace-based localisation may be employed, such as SHD Multiple-Signal Classification or ESPRIT [29], which generally offer higher resolution than their steered-response power/intensity-based counterparts; with the penalty of increased implementation complexity. ESPRIT is also a grid-less approach that directly provides the DoA estimates, which makes it especially suited to real-time operation.

III. Target Tracking

In this work, a tracking framework based on the RBMCDA method was employed [19], [20], which formulates the tracking and data associations as a Bayesian estimation problem. Here, the inference is conducted with SMC methods (i.e. particle filtering), and the accuracy and efficiency of the method is improved with Rao-Blackwellisation. The tracking of the target position and velocity is conducted by feeding DoA estimates as unit-length Cartesian vectors. Note that the purpose of this section is to provide a summary of the framework, and detail where it differs from the original toolbox described in [32]; for a more comprehensive description of the RBMCDA method, the reader is referred to [19], [20].

A. Multi-target filtering model

In the multiple target tracking model of K targets, it is assumed that the dynamics of each target are given by a Markovian model

$$\mathbf{x}_{t,j} \sim p(\mathbf{x}_{t,j} \mid \mathbf{x}_{t-1,j}), \quad j = 1, \ldots, K, \tag{1}$$

where $\mathbf{x}_{t,j} \in \mathbb{R}^6$ denotes the state of target j at time step t, with its values corresponding to the true target direction and velocity in Cartesian space $(x, y, z, \dot{x}, \dot{y}, \dot{z})$. The single-target dynamics are modelled using the Wiener velocity model [33], which has the form

$$p(\mathbf{x}_{t,j} \mid \mathbf{x}_{t-1,j}) = \mathcal{N}(\mathbf{x}_{t,j} \mid \mathbf{A}\mathbf{x}_{t-1,j}, \mathbf{Q}), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{6\times6}$ and $\mathbf{Q} \in \mathbb{R}^{6\times6}$ are the transition matrix and process noise covariance matrices, respectively.

The measurement model has the form

$$\mathbf{y}_t \sim \begin{cases} p(\mathbf{y}_t \mid c_t = 0), & \text{when } \mathbf{y}_t \text{ is clutter,} \\ p(\mathbf{y}_t \mid \mathbf{x}_{t,j}, c_t = j), & \text{when } \mathbf{y}_t \text{ associated with target } j \end{cases} \tag{3}$$

where $\mathbf{y}_t \in \mathbb{R}^3$ denotes a DoA estimate of (x, y, z) to be fed to the tracker and c_t is the (unknown) data association indicator. Following [20], a uniform clutter model $p(\mathbf{y}_t \mid c_t = 0) = 1/V$ was employed, where V is a constant. The target specific DoA measurements can be modelled with independent linear Gaussian models

$$p(\mathbf{y}_t \mid \mathbf{x}_{t,j}, c_t = j) = \mathcal{N}(\mathbf{y}_t \mid \mathbf{H}\mathbf{x}_{t,j}, \mathbf{R}), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{3\times6}$ is the measurement matrix (i.e. truncated identity matrix in this case, as no target velocity estimates are passed to the tracker); and $\mathbf{R} \in \mathbb{R}^{3\times3}$ is the measurement noise covariance matrix. Note that $\mathbf{A}$, $\mathbf{Q}$, and $\mathbf{R}$ are assumed to be target and time invariant and may be tuned based on a particular sound scene/distribution.

The aim is to estimate the current state of each target, j, given all DoA estimates that have been presented to the tracker thus far, by recursively computing the *posterior distributions*

$$p(\mathbf{x}_{t,j} \mid \mathbf{y}_{1:t}), \quad j = 1, \ldots, K. \tag{5}$$

Denoting the multiple target state as $\mathbf{x}_t = (\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,K}, \lambda_t)$, where λ_t includes the data association and birth/death variables (cf. [20]), the Bayesian filtering solution to the multiple target tracking problem may be initialised with the prior distribution $p(\mathbf{x}_0)$, and the *predictive and filtering distributions* of the state of $\mathbf{x}_t$ can be obtained recursively by the *Chapman–Kolmogorov equation* and Bayesian inference (cf. [32]).

B. Particle structure and Rao-Blackwellised particle filtering

In the following, the operation of the Rao–Blackwellised particle filter algorithm used for target tracking is described. As in [20], the following notation is employed henceforth: $\mathrm{KF}_p(.)$ and $\mathrm{KF}_u(.)$ denote a Kalman filter *prediction step* and *update step*, respectively. $\mathrm{KF}_{lh}(.)$ is then the *marginal measurement likelihood* of $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$. The tracker framework comprises a set of N particles, with each particle i at time step t containing

$$\left\{ c, w, \{\mathbf{m}, \mathbf{P}, T, \mathrm{id}\}_{j=1:K} \right\}_t^{(i)}, \tag{6}$$

where w is the particle importance weight; $\mathbf{m}_j = (\hat{x}, \hat{y}, \hat{z}, \hat{\dot{x}}, \hat{\dot{y}}, \hat{\dot{z}})$ and $\mathbf{P}_j \in \mathbb{R}^{6\times6}$ are the means and covariance matrices for each of the currently tracked targets, respectively; $T_j \in \mathbb{I}$ is a counter indicating how many time steps the target has been alive for; and $\mathrm{id}_j \in \mathbb{I}$ is a unique value assigned to each target. Note that particle filtering (see, e.g., [34], [35]) is intended to approximate complex distributions via a set of discrete samples (particles), and can be very accurate provided there are a sufficient number of them. It is based on the use of *importance sampling* where each of the particles is associated with an importance weight w, which is recursively updated during filtering. If the number of effective particles (i.e. those with high importance weight values relative to the other particles) falls below a specified threshold (e.g., $N/4$), then a re-sampling scheme is employed; whereby particles with lower importance weights are replaced with duplicates of particles with higher importance weights.

Rao–Blackwellised particle filters (see, e.g., [34], [35]) aim to improve the computational efficiency of particle filtering by

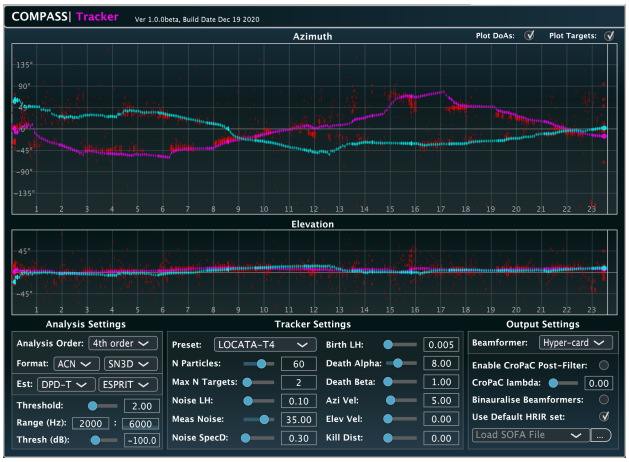

Fig. 1. The interface for the proposed real-time acoustic tracker system. DoA estimates are depicted in red, and the two target trajectories in magenta/cyan.

solving some parts, namely the Gaussian parts, of the filtering equations in closed form. In practice, a Rao–Blackwellised particle filter consists of a bank of Kalman filters, which are used to solve the conditionally Gaussian parts of the system.

C. Tracker update step

For each DoA estimate fed to the tracker, all of the particles consider the following three possible event hypotheses: **1)** Associating the estimate as clutter with an event likelihood related to the prior probability of noise parameter c_d, which is tuned based on the assumed performance of the localiser. **2)** Associating the estimate with a new target that has the following prior Gaussian distribution

$$p(\mathbf{x}_{t,j}) = \mathcal{N}(\mathbf{x}_{0,j} \mid \mathbf{m}_{0,j}, \mathbf{P}_{0,j}), \tag{7}$$

where $\mathbf{m}_{0,j} = [1, 0, 0, 0, 0, 0]$, i.e. assuming that the target is directly in-front and not moving; although, in practice, $\mathbf{P}_{0,j}$ is often configured to have high positional variance so that no specific target direction is favoured. The velocity variance priors are tuned based on whether the input scene comprises static or moving sound sources. A prior probability of target birth parameter, $p_b \in [0, 1]$, then influences the likelihood of a target birth. In this work, the original RBMCDA framework was modified so that should the number of targets exceed a specified maximum number K_{max} then $p_b = 0$. Otherwise, the likelihood of this event is calculated with $\mathrm{KF}_{lh}(\mathbf{y}_t, \mathbf{m}_{0,j}, \mathbf{P}_{0,j}, \mathbf{H}, \mathbf{R})$. **3)** Associating the estimate with an already established target, with an event likelihood of $\mathrm{KF}_{lh}(\mathbf{y}_t, \mathbf{m}^-_{t,j}, \mathbf{P}^-_{t,j}, \mathbf{H}, \mathbf{R})$, which (if chosen) would have its target state updated with

$$[\mathbf{m}_{t,j}, \mathbf{P}_{t,j}] = \mathrm{KF}_u(\mathbf{y}_t, \mathbf{m}^-_{t,j}, \mathbf{P}^-_{t,j}, \mathbf{H}, \mathbf{R}), \tag{8}$$

where $\mathbf{m}^-_{t,j}, \mathbf{P}^-_{t,j}$ are the predicted means and covariances.

The event hypothesis is then chosen based on drawing a sample from the optimal importance distribution for each particle independently; the importance weights are then updated based on their previous values and the selected event likelihood, and re-normalised. It is assumed that each association is independent of previous associations. Note that particles that select unlikely event hypotheses for many consecutive time steps are penalised with a reduced importance weight value (and eventually replaced due to the employed re-sampling scheme), whereas particles that consistently select likely events are gradually weighted higher. Real-time tracking is then based on the hypothesis held by the most dominant particle.

D. Tracker prediction step

For each time step, the probability of death $p_d \in [0, 1]$ is computed for all active targets. This calculation considers how long the target has been alive (T_j multiplied by time-step delta), and is modelled based on a Gamma distribution; thus allowing the user to influence how likely a target death can occur to better suit a specific sound scene or application. Additionally, the RBMCDA framework was modified so that if a target comes too close to another target, then $p_d = 1$ may be forced upon the younger of the two targets, which can improve performance for static source scenarios in challenging acoustical conditions. This may also mitigate beamformer instabilities, should null constraints be imposed upon them during practical use cases of the tracker. The modified framework features an additional novel contribution of permitting the possibility of more than one target death to occur during one time step, as was suggested as future work in [32]. After the target death checks have been conducted, the states of all targets left alive are subjected to the following Kalman filter prediction step

$$[\mathbf{m}^-_{t+1,j}, \mathbf{P}^-_{t+1,j}] = \mathrm{KF}_p(\mathbf{m}_{t,j}, \mathbf{P}_{t,j}, \mathbf{A}, \mathbf{Q}). \tag{9}$$

IV. Evaluation

The evaluation of the proposed system was based on the framework established under the LOCATA 2018 challenge [27], which comprises a corpus of recordings and a number of evaluation metrics intended for assessing the performance of localisation and tracking algorithms. The challenge results have since been published in [30], along with the MATLAB evaluation scripts[3], thus allowing new systems to be compared against the original submissions. Tasks 1–4 were selected for this study, as they comprised the following respective scenarios of interest: a single static source, multiple static sources, a single moving source, and multiple moving sources. The task recordings were captured using four different microphone arrays including an Eigenmike (a commercially available 32-capsule spherical array); the signals of which were first converted into fourth-order SH signals using [36] before being passed to the proposed system. All LOCATA 2018 recordings comprised speech stimuli and were captured in a reverberant room. The recordings were divided into two subsets: the first for the purpose of algorithm development (*dev*), and the second (*eval*) used for evaluating the submissions. The ground-truth positional data was provided for the development data set at the beginning of the challenge, whereas the evaluation ground-truth was released after the challenge ended. However,

[3]https://github.com/cevers/sap_locata_eval

TABLE I

AVERAGE AZIMUTH ERRORS AVERAGED OVER ALL RECORDINGS. THE HIGHLIGHTED ALGORITHMS INDICATE RESULTS TAKEN DIRECTLY FROM TABLE III IN [30]. STANDARD DEVIATION σ VALUES AND THE ORIGINAL SUBMISSION IDS ARE PROVIDED WHERE AVAILABLE.

Average Azimuth Error (degrees)					
Algorithm	ID	Task 1 (single-static)	Task 2 (multi-static)	Task 3 (single-moving)	Task 4 (multi-moving)
4th-order ESPRIT & RBMCDA	-	**2.1** ($\sigma = 1.0$)	**2.4** ($\sigma = 1.5$)	**5.9** ($\sigma = 4.2$)	**6.3** ($\sigma = 5.0$)
1st-order ESPRIT & RBMCDA	-	2.5 ($\sigma = 1.4$)	6.1 ($\sigma = 5.3$)	7.9 ($\sigma = 6.1$)	8.3 ($\sigma = 6.3$)
PIV & RBMCDA	-	4.0 ($\sigma = 2.3$)	5.9 ($\sigma = 6.5$)	9.4 ($\sigma = 6.9$)	7.9 ($\sigma = 6.4$)
MUSIC & PHD [22]	2	-	-	-	12.8
SRP-PHAT [23]	6	6.4	-	**8.1**	-
TRAMP (PIV & SMC) [24]	10	8.9	7.3	11.5	**9.0**
DPD-T & MUSIC-based [25]	12	**1.1**	**1.4**	-	-
Subspace-PIV [26]	15	8.1	7.1	-	-

TABLE II

RESULTS FOR THE OTHER LOCATA EVALUATION METRICS FOR THE PROPOSED SYSTEM (USING 4TH-ORDER ESPRIT), WHICH CAN BE COMPARED VISUALLY WITH THE PLOTS FOUND IN THE LOCATA RESULTS PUBLICATION[4] (CF. [30]). STANDARD DEVIATION VALUES ARE PROVIDED IN BRACKETS.

Metric	Task 1 (single-static)	Task 2 (multi-static)	Task 3 (single-moving)	Task 4 (multi-moving)
Average Elevation Error (degrees)	3.5 ($\sigma = 1.1$)	3.3 ($\sigma = 2.1$)	3.6 ($\sigma = 2.8$)	4.8 ($\sigma = 8.3$)
Track Latency (ms)	34.3 ($\sigma = 3.9$)	46.1 ($\sigma = 5.1$)	237.2 ($\sigma = 5.3$)	37.3 ($\sigma = 1.2$)
Probability of detection (%)	99.8 ($\sigma = 0.9$)	66.8 ($\sigma = 4.6$)	95.9 ($\sigma = 1.7$)	91.7 ($\sigma = 2.4$)
Fragmentation rate (frags/second)	0.057 ($\sigma = 0.111$)	0.156 ($\sigma = 0.225$)	0.036 ($\sigma = 0.081$)	0.111 ($\sigma = 0.090$)

in keeping with the spirit of the challenge, only the development recordings and ground-truth data were used to tune a parameter preset for each task. These presets were then employed on the *blind* evaluation recordings, and the results passed through the provided evaluation scripts.

A number of the LOCATA evaluation metrics are applicable to the intended future applications of the proposed tracking system. The metrics considered were: the direction estimation accuracy, probability of detection (PD), Track Latency (TL), and Track Fragmentation Rate (TFR). Estimation accuracy refers to the angular errors evaluated separately for azimuth and elevation (defined as the angle between the ground-truth and the estimated target direction). PD refers to the percentage of time stamps during which the source is associated with a valid track. TL is a measure of timeliness, evaluating the delay between the onset and the first detection of a valid sound source; and TFR is a measure of continuity, indicating the number of track fragmentations per second, which include instances of tracks *swapping* ids, or tracks *breaking*. Note that all of these metrics were computed during periods of ground-truth voice-activity.

V. RESULTS

The azimuth angle error values for the proposed system and the five original submissions [22]–[26] (where the Eigenmike was employed), for Tasks 1, 2, 3 and/or 4, are presented in Table I. Note that: in [22] (ID2), MUSIC was used for DoA estimation, followed by a Probability Hypothesis Density (PHD) [16] filter; [23] (ID6) applied the steered response power using the phase transform (SRP-PHAT) algorithm; the TRAMP system [24] (ID10) combined pseudo-intensity vector (PIV) localisation (using first-order SH input), followed by particle filtering; [25] (ID12) employed the DPD-T and a MUSIC-like measure described in [1], followed by k-means clustering; and [26] (ID15) employed the subspace-PIV approach [37]. It can be observed that the proposed system, (using both first- and fourth-order ESPRIT), yielded lower error values than all original submissions to the challenge for the moving source(s) tasks, and values that are close to the winning submission for the static source(s) tasks; although, it will be noted that no offline processing that exploits knowledge of static conditions was performed for the proposed system, as was the case for the winning submission (ID12). Furthermore, there appears to be minimal performance penalty going from fourth-order ESPRIT to first-order ESPRIT or PIV (as used by the closely related TRAMP system), for Tasks 1, 3 and 4. However, the ability to employ the full input resolution for the source localisation may be beneficial for other more complex scenarios, and is shown to yield some additional benefit for Task 2, where the sources were in particularly close proximity to each other for many of the recordings.

The results of the other LOCATA metrics of interest, when using fourth-order ESPRIT, are presented in Table II. Visually comparing the result values with the graphs in the LOCATA results publication[4] (cf. [30]), the PD for the proposed system is similar to ID6 and ID12; which were all slightly higher than ID10. The TL metric is similar to all original submissions for Task 1, and is lower than ID10 for Task 3. Finally, the TFR results are either lower than or comparable to ID10, ID12 and ID15 for Tasks 2 and 4.

VI. CONCLUSION

This paper has proposed an acoustic source tracking system based on direct-path dominance testing [28] and high-resolution subspace-based localisation [29]. The direction estimates are tracked using a modified version of the Rao-Blackwellised Monte Carlo data association (RBMCDA)

[4]The final scores for the original submissions were not made publicly available, outside of the tables and graphs presented in [30], or made available to the authors upon request; hence only a visual comparison is possible.

framework first laid out in [19], [20]. These modifications permit the system to operate better in real-time, and tailor it towards forthcoming work regarding flexible parametric sound-field manipulation and spatial audio effects. The system was evaluated by employing the LOCATA 2018 challenge Eigenmike recordings, and using the same evaluation scripts as used to evaluate the original submissions and presented in [30]. It is demonstrated that the proposed system yielded similar or better performance metric values to the winning original submissions to the challenge. However, importantly, the proposed system achieves these results also as a practical, real-time and open-source implementation. Furthermore, the tracking performance when using PIV or first-order ESPRIT is not significantly reduced compared to the fourth-order ESPRIT results for many of the tested scenarios; suggesting that the modified RBMCDA tracking framework may still be robust even when lower resolution, and computationally less complex, localisers are employed.

REFERENCES

[1] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE J. Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 131–142, 2018.

[2] M. Taseska and E. A. Habets, "Blind Source Separation of Moving Sources Using Sparsity-Based Source Detection and Tracking," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 3, pp. 657–670, 2018.

[3] J. Nikunen, A. Diment, and T. Virtanen, "Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Tracking," *IEEE/ACM Trans. Audio Speech and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

[4] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[5] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of Ambisonic recordings," in *Proc. 2nd Int. Conf. on Spatial Audio, Erlangen*, 2014.

[6] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[7] V. Pulkki, A. Politis, M.-V. Laitinen, J. Vilkamo, and J. Ahonen, "First-order directional audio coding (DirAC)," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. John Wiley & Sons, 2017, pp. 89–138.

[8] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. of the 2nd Int. Symp. on Ambisonics and Spherical Acoustics*, 2010, pp. 6–7.

[9] A. Politis, S. Tervo, and V. Pulkki, "COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6802–6806.

[10] L. McCormack and S. Delikaris-Manias, "Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm," in *EAA Spatial Audio Signal Processing Symposium*, 2019.

[11] S. Blackman and R. Popoli, "Design and analysis of modern tracking systems," *Norwood, MA: Artech House, 1999.*, 1999.

[12] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, 1995.

[13] J. Traa and P. Smaragdis, "Multiple speaker tracking with the Factorial von Mises-Fisher Filter," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[14] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[15] J. M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[16] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[17] S. Challa, M. R. Morelande, D. Mušicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[18] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

[19] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized Monte Carlo data association for multiple target tracking," in *Proc. 7th Int. Conf. Information Fusion*, vol. 1. I, 2004, pp. 583–590.

[20] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-Blackwellized particle filter for multiple target tracking," *Information Fusion Journal*, vol. 8, no. 1, pp. 2–15, 2007.

[21] X. Zhong and J. R. Hopgood, "A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE Trans. Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2356–2370, 2015.

[22] Y. Liu, W. Wang, and V. Kilic, "Intensity particle flow SMC-PHD filter for audio speaker tracking," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[23] R. Lebarbenchon, E. Camberlein, D. Di Carlo, C. Gaultier, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[24] S. Kitić and A. Guérin, "TRAMP: Tracking by a realtime ambisonic-based particle filter," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[25] L. Madmoni, H. Beit-On, H. Morgenstern, and B. Rafaely, "Description of algorithms for Ben-Gurion University submission to the LOCATA challenge," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[26] A. H. Moore, "Multiple source direction of arrival estimation using subspace pseudointensity vectors," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, 2018.

[27] 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 10th Sensor Array and Multichannel Signal Processing Workshop*, 2018, pp. 410–414.

[28] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[29] B. Jo and J.-W. Choi, "Parametric direction-of-arrival estimation with three recurrence relations of spherical harmonics," *J. Acoustical Society of America*, vol. 145, no. 1, pp. 480–488, 2019.

[30] C. Evers, H. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[31] K. Han and A. Nehorai, "Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing," *IEEE Trans. Signal Processing*, vol. 61, no. 23, pp. 6118–6128, 2013.

[32] J. Hartikainen and S. Särkkä, "RBMCDAbox - Matlab toolbox of Rao-Blackwellized data association particle filters," *Documentation of RBMCDA Toolbox for Matlab V*, 2008.

[33] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. Wiley, 2001.

[34] A. Doucet, N. De Freitas, and N. Gordon, "An introduction to sequential Monte Carlo methods," in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 3–14.

[35] S. Särkkä, *Bayesian Filtering and Smoothing*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013, vol. 3.

[36] L. McCormack, S. Delikaris-Manias, A. Farina, D. Pinardi, and V. Pulkki, "Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localized sub-band sound-field analysis," in *AES Convention 144*, 2018.

[37] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 178–192, 2016.