The LOCATA Challenge: Acoustic Source Localization and Tracking

Christine Evers, *Senior Member, IEEE*, Heinrich W. Löllmann, *Senior Member, IEEE*, Heinrich Mellmann, Alexander Schmidt, *Member, IEEE*, Hendrik Barfuss, Patrick A. Naylor, *Fellow, IEEE*, and Walter Kellermann, *Fellow, IEEE*

***Abstract*—The ability to localize and track acoustic events is a fundamental prerequisite for equipping machines with the ability to be aware of and engage with humans in their surrounding environment. However, in realistic scenarios, audio signals are adversely affected by reverberation, noise, interference, and periods of speech inactivity. In dynamic scenarios, where the sources and microphone platforms may be moving, the signals are additionally affected by variations in the source-sensor geometries. In practice, approaches to sound source localization and tracking are often impeded by missing estimates of active sources, estimation errors, as well as false estimates. The aim of the LOCAlization and TrAcking (LOCATA) Challenge is an open-access framework for the objective evaluation and benchmarking of broad classes of algorithms for sound source localization and tracking. This article provides a review of relevant localization and tracking algorithms and, within the context of the existing literature, a detailed evaluation and dissemination of the LOCATA submissions. The evaluation highlights achievements in the field, open challenges, and identifies potential future directions.**

***Index Terms*—Acoustic signal processing, source localization, source tracking, reverberation.**

I. Introduction

THE ABILITY to localize and track acoustic events is a fundamental prerequisite for equipping machines with awareness of their surrounding environment. Source localization provides estimates of positional information, e.g., Directions-of-Arrival (DoAs) or source-sensor distance, of acoustic sources in scenarios that are either permanently static, or static over finite time intervals. Source tracking extends source localization to dynamic scenarios by exploiting 'memory' from information acquired in the past in order to infer the present and predict the future source locations. It is commonly assumed that the sources can be modelled as point sources.

Situational awareness acquired through source localization and tracking benefits applications such as beamforming [1]–[3], signal extraction based on Blind Source Separation (BSS) [4]–[7], automatic speech recognition [8], acoustic Simultaneous Localization and Mapping (SLAM) [9], [10], and motion planning [11], with wide impact on applications in acoustic scene analysis, including robotics and autonomous systems, smart environments, and hearing aids.

In realistic acoustic environments, reverberation, background noise, interference and source inactivity lead to decreased localization accuracy, as well as missed and false detections of acoustic sources. Furthermore, acoustic scenes are often dynamic, involving moving sources, e.g., human talkers, and moving sensors, such as microphone arrays integrated into mobile platforms, such as drones or humanoid robots. Time-varying source-sensor geometries lead to continuous changes in the direct-path contributions of sources, requiring fast updates of localization estimates.

The performance of localization and tracking algorithms is typically evaluated using simulated data generated by means of the image method [12], [13] or its variants [14]. Evaluation by real-world data is a crucial requirement to assess the relevant performance of localization and tracking algorithms. However, open-access datasets recorded in realistic scenarios and suitable for objective benchmarking are available only for scenarios involving static sources, such as loudspeakers, and static microphone array platforms. To provide such data also for a wide range of dynamic scenarios, and thus foster reproducible and comparable research in this area, the LOCAlization and TrAcking (LOCATA) challenge provides a novel framework for evaluation and benchmarking of sound source localization and tracking algorithms, entailing:

1) An open-access dataset [15] of recordings from four microphone arrays in static and dynamic scenarios, completely annotated with the ground-truth positions and orientations for all sources and sensors, hand-labelled voice activity information, and close-talking microphone signals as reference.
2) An open-source software framework [16] of comprehensive evaluation measures for performance evaluation.

Manuscript received August 1, 2019; revised January 31, 2020 and April 3, 2020; accepted April 13, 2020. Date of publication April 27, 2020; date of current version June 5, 2020. This work was supported by the UK EPSRC Fellowship Grant EP/P001017/1, awarded to C. Evers. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Du. *(Corresponding author: Christine Evers.)*

Christine Evers was with the Department Electrical and Electronic Engineering, Imperial College London, London SW7 2BU, U.K. She is now with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: c.evers@soton.ac.uk).

Heinrich W. Löllmann, Alexander Schmidt, Hendrik Barfuss, and Walter Kellermann are with the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg, 91058 Erlangen, Germany (e-mail: heinrich.loellmann@fau.de; alexander.as.schmidt@fau.de; hendrik.barfuss@fau.de; walter.kellermann@fau.de).

Heinrich Mellmann is with the Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany (e-mail: mellmann@informatik.hu-berlin.de).

Patrick A. Naylor is with the Department Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: p.naylor@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2020.2990485

3) Results for all algorithms submitted to the LOCATA challenge for benchmarking of future contributions.

The LOCATA challenge corpus aims at providing a wide range of scenarios encountered in acoustic signal processing, with an emphasis on speech sources in dynamic scenarios. The scenarios represent applications in which machines should be equipped with the awareness of the surrounding acoustic environment and the ability to engage with humans, such that the recordings are focused on human speech sources in the acoustic far-field. All recordings contained in the corpus were made in a realistic, reverberant acoustic environment in the presence of ambient noise from a road in front of the building. The recording equipment was chosen to provide a variety of sensor configurations. The LOCATA corpus therefore provides recordings from arrays with diverse apertures. All arrays integrate omnidirectional microphones in a rigid baffle. The majority of arrays use consumer-type low-cost microphones.

The LOCATA corpus was previously described in [17], [18], and the evaluation measures were detailed in [19]. This paper provides the following additional and substantial contributions:

- A concise, yet comprehensive literature review, providing the background and framing the context of the approaches submitted to the LOCATA challenge.
- A detailed discussion of the benchmark results submitted to the LOCATA challenge, highlighting achievements, open challenges, and potential future directions.

This paper is organized as follows: Section II summarizes the scope of the LOCATA challenge. Sections III and IV summarize the LOCATA challenge tasks and corpus. Section V reviews the literature on acoustic source localization and tracking in the context of the approaches submitted to the LOCATA challenge. Section VI details and discusses the evaluation measures. The benchmarked results are presented in Section VII. Conclusions are drawn and future directions discussed in Section VIII.

II. Scope of the LOCATA Challenge and Corpus

Evaluation of localization and tracking approaches is often performed in a two-stage process. In the first stage, microphone signals are generated using simulated room impulse responses in order to control parameters, such as the reverberation time, signal-to-noise ratio, or source-sensor geometries. The second stage validates the findings based on measured impulse responses using a typically small number of recordings in real acoustic environments.

Since the recording and annotation of data is expensive and time-consuming, available open-access recordings are typically targeted at specific scenarios, e.g., for static sources and arrays [20], or for moving sources [21]. For comparisons of different algorithms across a variety of scenarios, the measurement equipment (notably microphone arrays) should be identical, or at least equivalent in all scenarios. In addition, annotation with ground-truth should be based on the same method, especially for assessing tracking performance.

A. Related Challenges and Corpora

Previous challenges related to LOCATA include, e.g., the CHiME challenges [22] for speech recognition, the ACE challenge [23] for acoustic parameter estimation, and the REVERB challenge [24] for reverberant speech enhancement. These challenges provide datasets of the clean speech signals and microphone recordings across a variety of scenarios, sound sources, and recording devices. In addition to the audio recordings, accurate ground-truth positional information of the sound sources and microphone arrays are required for source localization and tracking in LOCATA.

Available datasets of audio recordings for source localization and tracking are either limited to a single scenario, or are targeted at audio-visual tracking. For example, the SMARD dataset [20] provides audio recordings and the corresponding ground-truth positional information obtained from multiple microphone arrays and loudspeakers in a low-reverberant room ($T_{60} \approx 0.15\,\mathrm{s}$). Only a static single-source scenario is considered, involving microphone arrays and loudspeakers at fixed positions in an acoustically dry enclosure. The DIRHA corpus [25] provides multichannel recordings for various static source-sensor scenarios in three realistic, acoustic enclosures.

For dynamic scenarios, corpora targeted at audio-visual tracking, such as the AV16.3 dataset [21], typically involve multiple moving human talkers. The RAVEL and CAMIL datasets [26], [27] provide camera and microphone recordings from a rotating robot head. Annotation of the ground-truth source positions is typically performed in a semi-automatic manner, where humans label bounding boxes on small video segments. Therefore, ground-truth source positions are available only as 2D pixel positions, specified relative to the local frame of reference of the camera. For evaluation of acoustic source localization and tracking algorithms, the mapping from the pixel positions to DoAs or Cartesian positions is required. In practice, this mapping is typically unknown and depends on the specific camera used for the recordings.

For the CLEAR challenge [28], pixel positions were interpolated between multiple cameras in the environment in order to estimate the Cartesian positions of the sound sources. The CLEAR challenge provided audio-visual recordings from seminars and meetings involving moving talkers. In contrast to LOCATA, which also involves moving microphone arrays, the CLEAR corpus is based on static arrays only.

Infrared tracking systems are used for accurate ground-truth acquisition in [29] and by the DREGON dataset [30]. However, the dataset in [29] provides recordings from only a static, linear microphone array. DREGON is limited to signals emitted by static loudspeakers. Moreover, the microphone array is integrated in a drone, whose self-positions are only known from the motor data and may be affected by drift due to wear of the mechanical parts [31].

III. LOCATA Challenge Tasks

The scenarios contained in the LOCATA challenge corpus are represented by multichannel audio recordings and corresponding positional data. The scenarios were designed to be representative of practical challenges encountered in human-machine interaction, including variation in orientation, position, and speed of the microphone arrays as well as the talkers. Audio signals emitted in enclosed environments are subject to reverberation.

TABLE I
LOCATA CHALLENGE TASKS

Array	Static Loudspeakers		Moving Human Talkers	
	Single	Multiple	Single	Multiple
Fixed	Task 1	Task 2	Task 3	Task 4
Moving	-	-	Task 5	Task 6

Hence, dominant early reflections often cause false detections of source directions, whilst late reverberation, as well as ambient noise, can lead to decreased localization accuracy. Furthermore, temporally sparse or intermittently active sources, e.g., human speakers, result in missing detections during pauses. Meanwhile, interference from competing, concurrent sources requires multi-source localization approaches to ensure that situational awareness can be maintained. In practice, human talkers are directional and highly spatially dynamic, since head and body rotations and translations can lead to significant changes in the talkers positions and orientations within short periods of time. The challenge of localization in dynamic scenarios, involving both source and sensor motion, is to provide accurate estimates for source-sensor geometries that vary significantly over short time frames.

Therefore, machines must be equipped with sound source localization algorithms that prove to be robust against reverberation, noise, interference, and temporal sparsity of sound sources for static as well as time-varying source-sensor geometries. The scenarios covered by the LOCATA corpus are therefore aligned with six increasingly challenging tasks, listed in Table I.

The controlled scenarios of Task 1, involving a single, static sound source, facilitate detailed investigations of the adverse affects of reverberation and noise on source localization. Crucial insights about the robustness against interference and overlapping speech from multiple, simultaneously active sources can be investigated using the static, multi-source scenarios in Task 2. Using the data for Task 3, the impact of source directivity, as well as head and body rotations for human talkers, can be studied. Task 4 provides the recordings necessary to address the ambiguities arising in scenarios involving multiple moving human talkers, such as occlusion and shadowing of crossing talkers, the resolution of individual speakers, and the identification and initialization of new speaker tracks, subject to periods of speech inactivity. The fully dynamic scenarios in Task 5 and Task 6 are designed to bridge the gap between traditional signal processing applications that typically rely on static array platforms, and future directions in signal processing, progressing towards mobile, autonomous systems. Specifically, the data provides the framework required to identify and tackle challenges such as the self-localization of arrays [9], [10] and the integration of acoustic data for motion planning [33].

IV. LOCATA DATA CORPUS

A. Recording Setup

The recordings for the LOCATA data corpus were conducted in the computing laboratory at the Department of Computer Science at the Humboldt Universität zu Berlin, which is equipped with the optical tracking system OptiTrack [34]. The room size is $7.1 \times 9.8 \times 3$ m^3 with a reverberation time of about 0.55 s.

1) Microphone Arrays: The following four microphone arrays were used for the recordings (see [18]):

Robot head: A pseudo-spherical array with 12 microphones integrated into a prototype head for the humanoid robot NAO (see Fig. 1(a)), developed as part of the EU-funded project 'Embodied Audition for Robots (EARS)' [35], [36].

Eigenmike: The Eigenmike by mh acoustics, which is a spherical microphone array equipped with 32 microphones integrated in a rigid baffle of 84 mm diameter [32].

Distant talking Interfaces for Control of Interactive TV (DICIT) array: A planar array providing a horizontal aperture of width 2.24 m, and sampled by 15 microphones, realizing four nested linear uniform sub-arrays (see Fig. 1(b)) with inter-microphone distances of 4, 8, 16 and 32 cm respectively (see also [37]).

Hearing aids: A pair of non-commercial hearing aids (Siemens Signia, type Pure 7 mi) mounted on a head-torso simulator (HMS II of HeadAcoustics). Each hearing aid (see Fig. 1(c)) is equipped with two microphones (Sonion, type 50GC30-MP2) with an inter-microphone distance of 9 mm. The Euclidean distance between the hearing aids at the left and right ear of the head-torso simulator is 157 mm.

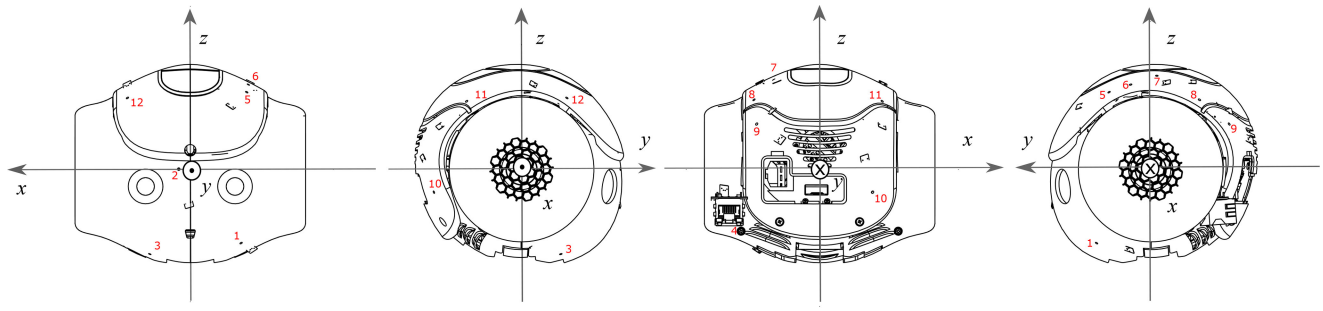
The array geometries were selected to sample the diversity of commonly used arrays in a meaningful and representative way. The multichannel audio recordings were performed with a sampling rate of 48 kHz and synchronized with the ground-truth positional data acquired by the OptiTrack system (see Section IV-C). A detailed description of the array geometries and recording conditions is provided by [18].

B. Speech Material

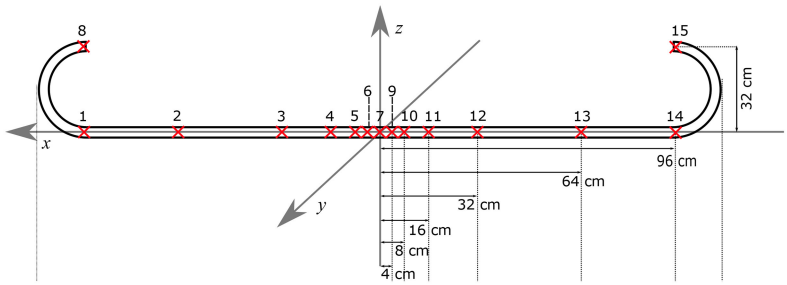
For Tasks 1 and 2, involving static sound sources, anechoic utterances from the Centre for Speech Technology Research (CSTR) Voice Cloning ToolKit (VCTK) dataset [38] were played back at 48 kHz sampling rate using Genelec 1029 A & 8020 C loudspeakers. For Tasks 3 to 6, involving moving sound sources, 5 non-native human talkers read randomly selected sentences from the CSTR VCTK dataset. The talkers were equipped with a DPA d:screet SC4060 microphone near their mouth, such that the close-talking speech signals were recorded. The anechoic and close-talking speech signals were provided to participants as part of the development dataset, but were excluded from the evaluation dataset.

C. Ground-Truth Positional Data

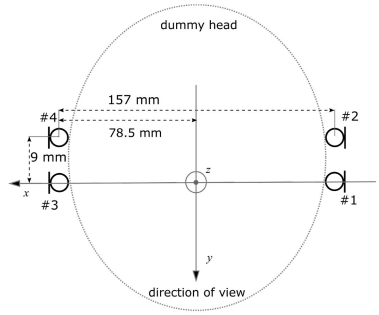
For the recordings, a 4×6 m^2 area was chosen within the $7.1 \times 9.8 \times 3$ m^3 room. Along the perimeter of the recording area, 10 synchronized and calibrated Infra-Red (IR) OptiTrack Flex 13 cameras were installed. Groups of reflective markers, detectable by the IR sensors, were attached to each source (i.e., loudspeaker or human talker) and microphone array. Each group of markers was arranged with a unique, asymmetric geometry,



(a) Robot head



(b) DICIT array



(c) Hearing aids on head-torso simulator

Fig. 1. Schematics of microphone array geometries of (a) the robot head, (b) the DICIT array, (c) the hearing aids used for the LOCATA corpus recordings. Schematics of the Eigenmike can be found in [32].

allowing the OptiTrack system to identify, disambiguate, and determine the orientation of all sources and arrays.

The OptiTrack system provided estimates of each marker position with approximately 1 mm accuracy [34] and at a frame rate of 120 Hz by multilateration using the IR cameras. Isolated outliers of the marker position estimates, caused by visual occlusions and reflections of the IR signals off surfaces, were handled in a post-processing stage that reconstructed missing estimates and interpolated false estimates. Details about the experimental setup are provided in [18].

Audio data was recorded in a block-wise manner and each data block was labeled with a time stamp generated by the global system time of the recording computer. On the the same computer, positional data provided by the OptiTrack system was recorded in parallel. Every position sample was labeled with a time stamp. After each recording was finished, the audio and positional data were synchronized using the time stamps.

For DoA estimation, local reference frames were specified relative to each array centre as detailed in [18]. For convenient transformations of the source coordinates between the global and local reference frames, the corpus provides the translation vectors and rotation matrices for all arrays for each time stamp. Source DoAs are defined within each array's local reference frame.

D. Voice Activity Labels

The Voice-Active Periods (VAPs) for the recordings of the LOCATA datasets were determined manually using the source signals, i.e., the signals emitted by the loudspeakers (Task 1 and 2) and the close-talking microphone signals (Tasks 3 to 6). The VAP labels for the signals recorded at the distant microphone arrays were obtained from the VAP labels for the source signals by accounting for the sound propagation delay between each source and the microphone array as well as the processing delay required to perform the recordings. The propagation delay was determined using the ground-truth positional data. The processing delay was estimated based on the cross-correlation between the source and recorded signals.

The ground-truth VAP labels were provided to the participants of the challenge as part of the development dataset but were excluded from the evaluation dataset.

V. Localization Scenarios, Methods, and Submissions

Localization systems process the microphone signals either as one batch for offline applications and static source-sensor geometries, or using a sliding window of samples for dynamic scenes. For each window, the instantaneous estimates of the source positions are estimated either directly from the signals, or using spatial cues inferred from the data, such as Time Delays of Arrival (TDoAs). To avoid spatial aliasing, nearby microphone pairs or compact arrays are typically used for localization. A few approaches are available to range estimation for acoustic sources, e.g., by exploiting the spatio-temporal diversity of a moving microphone array [10], [52], or by exploiting characteristics of the room acoustics [53], [54]. Nevertheless, in general, it is typically difficult to obtain reliable range estimates using static arrays. As such, the majority of source localization approaches focus on the estimation of the source DoAs, rather than the three-dimensional positions. In the following, the term 'source

TABLE II
SUMMARY OF LOCALIZATION AND TRACKING FRAMEWORKS SUBMITTED TO THE LOCATA CHALLENGE

ID	Details	Tasks	VAD	Localization		Tracking		Arrays
				Algorithm	Section	Algorithm	Section	
1	[39]	1	-	LDA classification	V-B2	-	-	Hearing Aids
2	[40]	4	-	MUSIC	V-B1	Particle PHD filter + Particle Flow	V-C2	Robot Head DICIT Hearing Aids Eigenmike
3	[41]	1,3,5	-	GCC-PHAT	V-A1	Particle filter	V-C1	DICIT
4	[42]	1-6	Variational EM	Direct-path RTF + GMM	V-A1	Variational EM	V-C2	Robot Head
6	[43]	1,3,5	-	SRP-PHAT	V-A3	-	-	Eigenmike Robot Head
7	[44]	1,3,5	CPSD trace	SRP Beamformer	V-A3	Kalman filter	V-C1	DICIT
8	[45]	1,3,5	-	TDE using IPDs	V-A1, V-A2	Wrapped Kalman filter	V-C1	Hearing Aids
9	[46]	1	-	DNN	V-B2	-	-	DICIT
10	[47]	1-4	Noise PSD	PIVs from first-order ambisonics	V-A4	Particle filter	V-C1	Eigenmike
11	[48]	1,2	-	DPD-Test + MUSIC	V-B1	-	-	Robot Head
12	[48]	1,2	-	DPD-Test + MUSIC in SH-domain	V-B1, V-A4	-	-	Eigenmike
13	[49]	1,3	Zero-crossing rate	MUSIC (SVD)	V-B1	Kalman filter	V-C1	DICIT
14	[49]	1,3	Zero-crossing rate	MUSIC (GEVD)	V-B1	Kalman filter	V-C1	DICIT
15	[50]	1	Baseline [51]	Subspace PIV	V-A4, V-B1	-	-	Eigenmike
16	[50]	2	Baseline [51]	Subspace PIV + Peak Picking	V-A4, V-B1	-	-	Eigenmike

localization' will be used synonymously with DoA estimation unless otherwise stated.

Due to reverberation, noise, and non-stationarity of the source signals, the position estimates at the output of the localization system are affected by false, missing and spurious estimates, as well as localization errors. Source tracking approaches incorporate spatial information inferred from past observations by applying spatio-temporal models of the source dynamics to obtain smoothed estimates of the source *trajectories* from the instantaneous DoA estimates presented by the localization system.[1]

This section provides the background and context for the approaches submitted to the LOCATA challenge so that the submissions can be related to each other and the existing literature in the broad area of acoustic source localization (see Table II and Fig. 2). As such, it does not claim the technical depth of surveys like those specifically targeted at sound source localization for robotics, or acoustic sensor networks, e.g., [55]–[57]. The structure of the review is aligned with the LOCATA challenge tasks as detailed in Section III. Details of each submitted approach are provided in the corresponding LOCATA proceedings paper, provided in the references below. Among the 16 submissions to LOCATA, 15 were sufficiently well documented to allow consideration in this paper. 11 were submitted from academic research institutions, 2 from industry, and 2 were collaborations between academia and industry. The global scope of the challenge is reflected by the geographic diversity of the submissions originating from the Asia (3 submissions), Middle East (2 submissions) and Europe (10 submissions).

[1]We note that, within the context of the LOCATA challenge, the following discussion focuses on speech, i.e., non-stationary wideband signals corresponding to energy that is concentrated in the lower acoustic frequency bands.

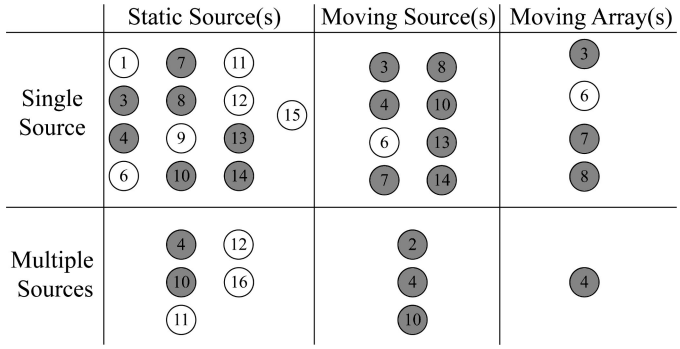


Fig. 2. Submissions to the LOCATA Challenge, ordered by Challenge Task (see Table I). Numbers indicate the submission ID. White shade: approaches incorporating source localization only. Grey shade: Approaches incorporating source localization and tracking.

A. Single-Source Localization

The following provides a review of approaches for localization of a single, static source, such as a loudspeaker.

1) Time Delay Estimation: If sufficient characteristics of a source signal are known *a priori*, the time delay between the received signals obtained at spatially diverse microphone positions can be estimated and exploited to triangulate the position of the emitting sound source. Time Delay Estimation (TDE) effectively maximizes the 'synchrony' [58] between time-shifted microphone outputs in order to identify the source position. A brief summary of TDE techniques is provided in the following. Details and references can be found in, e.g., [3, Chap. 9].

The TDoA, $\tau_{m,\ell}(\mathbf{x}_s)$, of a signal emitted from source position, $\mathbf{x}_s$, between two microphones, m and ℓ, at positions $\mathbf{x}_m$ and $\mathbf{x}_\ell$, respectively, is given by:

$$\tau_{m,\ell}(\mathbf{x}_s) \triangleq \frac{f_s}{c}\left(\|\mathbf{x}_s - \mathbf{x}_m\| - \|\mathbf{x}_s - \mathbf{x}_\ell\|\right), \qquad (1)$$

where f_s is the sampling frequency, c is the speed of sound, and $\|\cdot\|$ denotes the Euclidean norm. If the source signal corresponds to white Gaussian noise and is emitted in an anechoic environment, the TDoA between two microphones can be obtained by identifying the peaks in the cross-correlation between microphone pairs. Since speech signals are often nearly periodic for short intervals, the cross-correlation may exhibit spurious peaks that do not correspond to spatial correlations. The cross-correlation is therefore typically generalized to include a weighting function in the Discrete-Time Fourier Transform (DTFT) domain that causes a phase transform to pre-whiten the correlated speech signals, an approach referred to as Generalized Cross-Correlation (GCC)-PHAse Transform (PHAT). The GCC, $R_{m,\ell}(\tau)$, is defined as:

$$R_{m,\ell}(\tau) \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_{m,\ell}(e^{\jmath\omega})\, S_m(e^{\jmath\omega})\, S_\ell^*(e^{\jmath\omega})\, e^{\jmath\omega\tau} d\omega, \tag{2}$$

where $S_m(e^{\jmath\omega})$ denotes the DTFT of the received signal, s_m, at microphone m, and $*$ denotes the complex conjugate. The PHAT corresponds to a weighting function, $\phi_{m,\ell}(e^{\jmath\omega})$, of the GCC, where

$$\phi_{m,\ell}(e^{\jmath\omega}) \triangleq |S_m(e^{\jmath\omega})\, S_\ell^*(e^{\jmath\omega})|^{-1}. \tag{3}$$

The signal models underpinning the GCC as well as its alternatives rely on a free-field propagation model of the sound waves. Therefore, in reverberant environments, spectral distortions and temporal correlations due to sound reflections often lead to spurious peaks in the GCC function. The presence of multiple, simultaneously active sources can cause severe ambiguities in the distinction of peaks due to the direct path of sources from peaks arising due to reflections.

To explicitly model the reverberant channel, the fact that the Time-of-Arrival (ToA) of the direct-path signal from a source impinging on a microphone corresponds to a dominant peak in the Acoustic Impulse Response (AIR) can be exploited. The EigenValue Decomposition (EVD) [59], realized by, e.g., the gradient-descent constrained Least-Mean-Square (LMS) algorithm, can be applied for estimation of the early part of the relative impulse response. The work in [60] extracts the TDoA as the main peak in the relative impulse response corresponding to the Relative Transfer Function (RTF) [61] for improved robustness against reverberation and stationary noise. The concept of RTFs was also used in [62] for a supervised learning approach for TDoA estimation.

For localization, it is often desirable to estimate the source directions from TDoA estimates, e.g., using multi-dimensional lookup tables [63], by triangulation using Least Squares (LS) optimization if the array geometry is known *a priori* [64], [65], or by triangulation based on the intersection of interhyperboloidal spatial regions formed by the TDoA estimates, e.g., [66], [67].

The following single-source tracking approaches were submitted to the LOCATA challenge:

ID 3 [41] combines TDE for localization with a particle filter (see Section V-C1) for tracking using the DICIT array for the single-source Tasks 1, 3 and 5.

ID 4 [42] combines DoA estimation using the direct-path RTF approach in [62] with a variational Expectation-Maximization (EM) algorithm [68] (see Section V-C2) for multi-source tracking using the robot head for all Tasks.

ID 8 [45] combines TDE (see Section V-A1) with binaural features (see Section V-A2) for localization and applies a wrapped Kalman filter [69] for source tracking using the hearing aids in the single-source Tasks 1, 3 and 5.

2) Binaural Localization: The Head-Related Transfer Functions (HRTFs) [70] at a listener's ears encapsulate spatial cues about the relative source position including Interaural Level Differences (ILDs), Interaural Phase Differences (IPDs), and Interaural Time Differences (ITDs) [71]–[73], equivalent to TDoAs, and are used for source localization in, e.g., [74]–[78].

Sources positioned on the 'cone of confusion' lead to ambiguous binaural cues that cannot distinguish between sources in the frontal and rear hemisphere of the head [79], [80]. Human subjects resolve front-back ambiguities by movements of either their head [81]–[83] or the source controlled by the subject [84], [85]. Changes in ITDs due to head movements are more significant for accurate localization than changes in ILDs [86]. In [87], the head motion is therefore exploited to resolve front-back ambiguity for localization algorithms. In [88], the attenuation effect of an artificial pinna attached to a spherical robot head is exploited in order to identify level differences between signals arriving from the frontal and rear hemisphere of the robot.

The following binaural localization approaches were submitted to the LOCATA challenge:

ID 8 [45] combines TDE (see Section V-A1) with IPDs for localization and apply a wrapped Kalman filter [69] (see Section V-C1) for source tracking using the hearing aids in the single-source Tasks 1, 3 and 5.

3) Beamforming and Spotforming: Beamforming and spotforming techniques can be applied directly to the raw sensor signals in order to 'scan' the acoustic environment for positions corresponding to significant sound intensity [89]–[92]. In [93], a beam is steered in each direction corresponding to a grid, $\mathcal{X}$, of discrete candidate directions. Hence, the Steered Response Power (SRP), $P_{\text{SRP}}(\mathbf{x})$, is:

$$P_{\text{SRP}}(\mathbf{x}) = \sum_{m=1}^{M}\sum_{\ell=1}^{M} R_{m,\ell}(\tau_{m,\ell}(\mathbf{x}_s)), \tag{4a}$$

where M is the number of microphones. An estimate, $\hat{\mathbf{x}}_s$, of the source positions is obtained as:

$$\hat{\mathbf{x}}_s = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\, P_{\text{SRP}}(\mathbf{x}). \tag{5}$$

Similar to GCC, SRP relies on uncorrelated source signals and, hence, may exhibit spurious peaks when evaluated for speech signals. Therefore, SRP-PHAT [94] applies PHAT for pre-whitening of SRP.

The following beamforming approaches were submitted to the LOCATA challenge:

ID 6 [43] applies SRP-PHAT for the single-source Tasks 1, 3, and 5 using the robot head and the Eigenmike.

ID 7 [44] combines diagonal unloading beamforming [95] for localization with a Kalman filter (see Section V-C1) for source tracking using a 7-microphone linear subarray of the DICIT array for the single-source Tasks 1, 3 and 5.

4) Spherical Microphone Arrays: Spherical microphone arrays [96] sample the soundfield in three dimensions using microphones that are distributed on the surface of a spherical and typically rigid baffle. The spherical geometry of the array elements facilitates efficient computation based on an orthonormal wavefield decomposition. The response of a spherical microphone array can be described using spherical harmonics [97]. Equivalent to the Fourier series for circular functions, the spherical harmonics form a set of orthonormal basis functions that can be used to represent functions on the surface of a sphere. The sound pressure impinging from the direction, $\mathbf{\Omega} = [\theta, \phi]^T$, on the surface a spherical baffle with radius, r, from plane wave with unit amplitude and emitted from the source DoA, $\mathbf{\Phi}_s = [\theta_s, \phi_s]^T$, with elevation, θ_s, and azimuth, ϕ_s, is given by [98]:

$$f_{nm}(k, r, \mathbf{\Omega}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} b_n(kr)\ \left(Y_n^m(\mathbf{\Phi})\right)^*\ Y_n^m(\mathbf{\Omega}), \quad (6)$$

where k is the wavenumber, the weights, $b_n(\cdot)$, are available for many array configurations, and $Y_n^m(\cdot)$ denotes the spherical harmonic of order n and degree m.

Therefore, existing approaches to source localization can be extended to the signals in the domain of spherical harmonics. A Minimum Variance Distortionless Response (MVDR) beamformer [2] is applied for near-field localization in the domain of spherical harmonics in [99]. The work in [14], [100] proposes a 'pseudo-intensity vector' approach that steers a dipole beamformer along the three principal axes of the coordinate system in order to approximate the sound intensity using the spherical harmonics coefficients obtained from the signals acquired from a spherical microphone array.

The following approaches, targeted at spherical microphone arrays, were submitted to the LOCATA challenge:

ID 10 [47] combines localization using the first-order ambisonics configuration of the Eigenmike with a particle filter (see Section V-C1) for Tasks 1–4.
ID 12 [48] extends MUltiple SIgnal Classification (MUSIC) (see Section V-B) to processing in the domain of spherical harmonics of the Eigenmike signals for Tasks 1 and 2.
ID 15 [50] applies the subspace pseudo-intensity approach in [101] to the Eigenmike signals in the static-source Task 1.
ID 16 [50] extends the approach of ID 15 for the static multi-source Task 2 by incorporating source counting.

B. Multi-Source Localization

This subsection reviews multi-source localization approaches. Beyond the algorithms submitted to the LOCATA challenge, approaches based on, e.g., blind source separation [102]–[105] can be used for multi-source localization.

1) Subspace Techniques: Since spatial cues inferred from the received signals may not be sufficient to resolve between multiple, simultaneously active sources, subspace-based localization techniques rely on diversity between the different sources. Specifically, assuming that the sources are uncorrelated, subspace-based techniques, such as MUSIC [106] or Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [107]–[109] resolve between temporally overlapping signals by mapping the received signal mixture to a space where the source signals lie on orthogonal manifolds.

MUSIC [106] exploits the subspace linked to the largest eigenvalues of the correlation matrix to estimate the locations of N sources. The fundamental assumption is that the correlation matrix, $\mathbf{R}$, of the received signals can be decomposed, e.g., using Singular Value Decomposition (SVD) [110], into a signal subspace, $\mathbf{U}_s = [\mathbf{U}_s^1 \ldots, \mathbf{U}_s^N]$, consisting of N uncorrelated plane-wave signals, $\mathbf{U}_s^n$ for $n \in \{1, \ldots, N\}$, and an orthogonal noise subspace. The spatial spectrum from direction, $\mathbf{\Omega}$, for plane wave, $n \in \{1, \ldots, N\}$, is:

$$P_{\text{MUSIC}}(\mathbf{\Omega}) = \left(\mathbf{v}^T(\mathbf{\Omega})\left(\mathbf{I} - \mathbf{U}_s^n\left(\mathbf{U}_s^n\right)^H\right)\mathbf{v}^*(\mathbf{\Omega})\right)^{-1}, \quad (7)$$

where H denotes the Hermitian transpose, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{v}$ corresponds to the steering vector. MUSIC extensions to broadband signals, such as speech, can be found in, e.g., [63], [111]. However, the processing of correlated sources remains challenging since highly correlated sources correspond to a rank-deficient correlation matrix, such that the signal and noise space cannot be separated effectively. This is particularly problematic in realistic acoustic environments, since reverberation corresponds to a convolutive process, in contrast to the additive noise model underpinning MUSIC.

For improved robustness in reverberant conditions, [112] introduce a 'direct-path dominance' test. The test retains only the time-frequency bins that exhibit contributions of a single source, i.e., whose spatial correlation matrix corresponds to a rank-1 matrix, hence reducing the effects of temporal smearing and spectral correlation induced by reverberation. For improved computational efficiency, [101] replaces MUSIC with the pseudo-intensity approach in [100].

The following subspace-based localization approaches were submitted to the LOCATA challenge:

ID 2 [40] utilizes DoA estimates from MUSIC as inputs to a Probability Hypothesis Density (PHD) filter [113], [114] (see Section V-C2) for Task 4, evaluated for all four arrays.
ID 11 [48] utilizes the direct-path dominance test [112] and MUSIC in the Short-Time Fourier Transform (STFT) domain for the robot head signals for static-source Tasks 1 and 2.
ID 12 [48] extends the approach of ID 11 to processing in the domain of spherical harmonics (see Section V-A4) of the Eigenmike signals for Tasks 1 and 2.
ID 13 [49] applies MUSIC for localization and a Kalman filter (see Section V-C1) for tracking to single-source Tasks 1 and 3 using the robot head and the Eigenmike.
ID 14 [49] extends the approach of ID 13 to apply the Generalized EVD (GEVD) to MUSIC.

ID 15 and 16 [50] apply the subspace pseudo-intensity approach in [101] (see Section V-A4) to the Eigenmike signals in Tasks 1 and 2, respectively.

2) Supervised Learning and Neural Networks: Data-driven approaches can be used to exploit prior information available from large-scale datasets. The work in [115] assumes that frequency-dependent ILD and IPD values are located on a locally linear manifold. In a supervised learning approach, the mapping between the binaural cues and the source locations is learnt from annotated data using a probabilistic piecewise affine regression model. A semi-supervised approach is proposed in [116] that uses RTF values input features in order to learn the source locations based on manifold regularization.

To avoid the efforts for hand-crafted signal models, neural network-based ('deep') learning approaches can also be applied to sound source localization. Previous approaches use hand-crafted input vectors including established localization parameters such as GCC [117], [118], eigenvectors of the spatial coherence matrix [119], [120] or ILDs and cross-correlation function in [121]. TDoAs were used in, e.g., [122], [123], to reduce the adverse affects of reverberation. End-to-end learning for given acoustic environments uses either the time-domain signals or the STFT-domain signals only as the input for the network. In [124], the DoA of a single desired source from a mixture of the desired source and an interferer is estimated by a Deep Neural Network (DNN) with separate models for the desired source and the interferer. In [125], DoA estimation is considered as a multi-label classification problem, where the range of candidate DoA values is divided into small sectors, each sector representing one class.

The following approaches were submitted to LOCATA:

ID 1 [39] proposes a classifier based on linear discriminant analysis and trained using features based on the amplitude modulation spectrum of the hearing aid signals for Task 1.

ID 9 [46] uses a DNN regression model for localization of the source DoA for Task 1 using four microphone signals of the DICIT array.

C. Tracking of Moving Sources

Source localization approaches provide instantaneous estimates of the source DoAs, independent of information acquired from past observations. The DoA estimates are typically unlabelled and cannot be easily associated with estimates from the past. In order to obtain smoothed source trajectories from the noisy DoA estimates, tracking algorithms apply a two-stage process that a) predicts potential future source locations based on past information, and b) corrects the localized estimates by trading off the uncertainty in the prediction against the estimation error of the localization system.

1) Single-Source Tracking: Tracking algorithms based on Bayesian inference aim to estimate the marginal posterior Probability Density Function (pdf) of the current state of the source, conditional on the full history of observations. In the context of acoustic tracking, the source state often corresponds to either the Cartesian source position, $\mathbf{x}(t)$, or the DoA, $\mathbf{\Phi}(t)$, at time stamp, t. The state may also contain the source velocity and acceleration. The observations correspond to estimates of either the source position, $\mathbf{y}(t)$, TDoAs, $\tau_{m,\ell}(\mathbf{x}(t))$, or DoA, $\boldsymbol{\omega}(t)$ provided by the localization system. Assuming a first-order Markov chain and observations in the form of DoAs, the posterior pdf can be expressed as:

$$p\left(\mathbf{\Phi}(0:t') \mid \boldsymbol{\omega}(1:t')\right) = p\left(\mathbf{\Phi}(0)\right) \prod_{t=1}^{t'} p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(0:t-1), \boldsymbol{\omega}(1:t)\right), \quad (8)$$

where $\mathbf{\Phi}(0:t') \triangleq [\mathbf{\Phi}^T(0), \ldots, \mathbf{\Phi}^T(t')]^T$. Using Bayes's theorem:

$$p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(0:t-1), \boldsymbol{\omega}(1:t)\right) = \frac{p\left(\boldsymbol{\omega}(t) \mid \mathbf{\Phi}(t)\right)\, p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(t-1)\right)}{\int_{\mathcal{P}} p\left(\boldsymbol{\omega}(t) \mid \mathbf{\Phi}(t)\right)\, p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(t-1)\right) d\mathbf{\Phi}(t)}, \quad (9)$$

where $p(\boldsymbol{\omega}(t) \mid \mathbf{\Phi}(t))$ is the likelihood function, $p(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(t-1))$ is the prior pdf, determined using a dynamical model, and $\mathcal{P}$ is the support of $\mathbf{\Phi}(t)$. For online processing, it is often desirable to estimate sequentially the filtering density, $p(\mathbf{\Phi}(t) \mid \boldsymbol{\omega}(1:t))$, instead of (9). For linear Gaussian state spaces [126], where the dynamical model and the likelihood function correspond to normal distributions, the filtering density reduces to a Kalman filter [127], [128].

However, the state space models used for acoustic tracking are typically non-linear and/or non-Gaussian [10], [53]. For example, in [129], [130], the trajectory of Cartesian source positions is estimated from the TDoA estimates. Since the relationship between a source position and the corresponding TDoAs is non-linear, the integral in (9) is analytically intractable. The particle filter is a widely used sequential Monte Carlo method [131] that approximates the intractable posterior pdf by importance sampling of a large number of random variates, $\{\hat{\mathbf{\Phi}}^{(i)}(t)\}_{i=1}^{I}$, - or 'particles' -, from a proposal distribution, $g(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(0:t-1), \boldsymbol{\omega}(1:t))$, i.e.,

$$p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(0:t-1), \boldsymbol{\omega}(1:t)\right) \approx \sum_{i=1}^{I} w^{(i)}(t)\, \delta_{\hat{\mathbf{\Phi}}^{(i)}(t)}(\mathbf{\Phi}(t)), \quad (10)$$

where δ denotes the Dirac measure, and the importance weights, $w^{(i)}(t)$, are given by:

$$w^{(i)}(t) = w^{(i)}(t-1)\, \frac{p\left(\boldsymbol{\omega}(t) \mid \mathbf{\Phi}(t)\right)\, p\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(t-1)\right)}{g\left(\mathbf{\Phi}(t) \mid \mathbf{\Phi}(0:t-1), \boldsymbol{\omega}(1:t)\right)}. \quad (11)$$

The authors of [129], [130] rely on prior importance sampling [132] from the prior pdf. Each resulting particle is assigned a probabilistic weight, evaluated using the likelihood function of the TDoAs estimates. The work in [133] uses the SRP function instead of TDoA estimates as observations. Rao-Blackwellized particle filters [134] are applied in [135], [136] instead of prior

importance sampling. Resampling algorithms [137]–[141] ensure that only stochastically relevant particles are retained and propagated in time.

The tracking accuracy is highly dependent on the specific algorithm used for localization. Moreover, tracking approaches that rely on TDoA estimates are crucially dependent on accurate calibration [142] and synchronization [143]. To relax the dependency on calibration and synchronization, DoA estimates can be used as observations instead of TDoA estimates. To appropriately address the resulting non-Gaussian state-space model, a wrapped Kalman filter is proposed in [69] that approximates the posterior pdf of the source directions by a Gaussian mixture model, where the mixture components account for the various hypotheses that the state at the previous time step, the predicted state at the current time step, or the localized DoA estimate may be wrapped around π. To avoid an exponential explosion of the number of mixture components, mixture reduction techniques [144] are required.

Rather than approximating the angular distribution by a Gaussian mixture, a von Mises filter, based on directional statistics [145], [146], is proposed in [53]. The Coherent-to-Diffuse Ratio (CDR) [147], [148] is used as a measure of reliability of the DoA estimates in order to infer the unmeasured source-to-sensor range.

The following single-source tracking approaches were submitted to the LOCATA challenge:

ID 3 [41] combines TDE (see Section V-A1) for localization with a particle filter for tracking using the DICIT array for the single-source Tasks 1, 3 and 5.

ID 7 [44] combines diagonal unloading beamforming [95] (see Section V-A3) for localization with a Kalman filter for source tracking using a 7-microphone linear subarray of the DICIT array for Tasks 1, 3 and 5.

ID 8 [45] combines TDE (see Section V-A1) with IPDs (see Section V-A2) for localization and apply a wrapped Kalman filter [69] for source tracking using the hearing aids for Tasks 1, 3 and 5.

ID 10 [47] combines localization using the first-order ambisonics configuration (see Section V-A4) of the Eigenmike with a particle filter for Tasks 1-4.

ID 13 and ID 14 [49] apply variants of MUSIC (see Section V-B1) for localization and a Kalman filter for tracking the source DoAs for Tasks 1 and 3 using the robot head and the Eigenmike.

2) Multi-Source Tracking: For multiple sources, not only the source position, but also the number of sources is subject to uncertainty. However, this uncertainty cannot be accounted for within the classical Bayesian framework.

Heuristic data association techniques are often used to associate existing tracks and observations, as well as to initialize new tracks. Data association partitions the observations into track 'gates' [149], or collars, around each predicted track in order to eliminate unlikely observation-to-track pairs. Only observations within the collar are considered when evaluating the track-to-observation correlations. Nearest-neighbour approaches determine a unique assignment between each observation and at most one track by minimizing an overall distance metric. However, in dense, acoustic environments, such as the cocktail party scenario [150], [151], many pairs between tracks and observations may result in similar distance values, and hence a high probability of association errors. For improved robustness, probabilistic data association can be used instead of heuristic gating procedures, e.g., the Probabilistic Data Association Filter (PDAF) [152], [153], or Joint Probabilistic Data Association (JPDA) [154], [155].

To avoid explicit data association, the work in [68] models the observation-to-track associations as discrete latent variables within a variational EM approach for multi-source tracking. Estimates of the latent variables provide the track-to-observation associations. The work in [156] extends the variational EM in [68] to a incorporate a von Mises distribution [53] for robust estimation of the DoA trajectories.

To incorporate track initiation and termination in the presence of false and missing observations, the states of multiple sources can be formulated as realizations of a Random Finite Set (RFS) [114], [157]. In contrast to random vectors, RFSs capture not only the time-varying source states, but also the unknown and time-varying number of sources. Finite set statistics [158], [159] provide the mathematical mechanisms to treat RFSs within the Bayesian paradigm. Since the pdf of RFS realizations is combinatorially intractable, its first-order approximation, the PHD filter [114] provides estimates of the intensity function – as opposed to the pdf – of the number of sources and their states.

The PHD filter was applied in [160], [161] for the tracking of the positions of multiple sources from the TDoA estimates. Due to the non-linear relationship between the Cartesian source positions and TDoA estimates, the prediction and update for each hypothesis within the PHD filter is realized using a particle filter as previously detailed in Section V-C1. A PHD filter for bearing-only tracking from the localized DoA estimates was proposed in [162], incorporating a von Mises mixture filter for the update of the source directions. The work in [9], [10] applies a PHD filter in order to track the source positions from DoA estimates for SLAM.

The following multi-source tracking approaches were submitted to the LOCATA challenge:

ID 2 [40] utilizes DoA estimates from MUSIC (see Section V-B1) as inputs to a PHD filter [113], [114] with intensity particle flow [163] for Task 4, using all four arrays.

ID 4 [42] combines DoA estimation using the direct-path RTF approach in [62] (see Section V-A1) with the variational EM algorithm in [68] for all Tasks using the robot head.

VI. Evaluation Measures

This section provides a discussion of the performance measures used for evaluation of the LOCATA challenge.

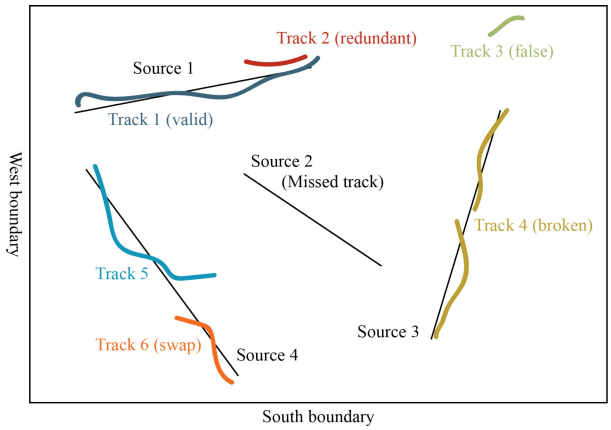


Fig. 3. Tracking ambiguities. Colors indicate unique track IDs.

A. Source Localization and Tracking Challenges

In realistic acoustic scenarios, source localization algorithms are affected by a variety of challenges (see Fig. 3). Fast localization estimates using a small number of time frames often result in estimation errors for signals that are affected by late reverberation and noise. Sources are often missed, e.g., due to periods of voice inactivity, for distant sources corresponding to low signals levels, or for sources oriented away from the sensors. False estimates arise due to, e.g., strong early reflections mistaken as the direct path of a source signal, or reverberation causing temporal smearing of speech energy beyond the endpoint of a talker's utterance, and due to overlapping speech energy in the same spectral bins for multiple, simultaneously active talkers.

Source tracking algorithms typically use localization estimates as observations. To distinguish inconsistent false estimates from consistent observations, tracking approaches often require multiple, consecutive observations of the same source direction or position before a track is initialized. Furthermore, track termination rules are necessary to distinguish between speech endpoints and missing estimates. To avoid premature track deletions due to short-term missing estimates, track termination rules are often based on the lapsed time since the last track update. Uncertainty due to the onsets and endpoints of speech activity may therefore lead to a latency between the onsets and endpoints of speech and the initialization and termination, respectively, of the corresponding source track.

In practice, uncertainty in the source dynamical model and in the observations may lead to divergence of the track from the ground-truth trajectory of an inactive source. In multi-source scenarios, track divergence may also occur by mistakenly updating a source's track with estimates of a different, nearby source. As a consequence, track swaps may occur due to the divergence of a track to the trajectory of a different source. Furthermore, a track may be broken if the track is not assigned to any source for one or more time steps, i.e., the assignment between a source and its estimates is temporarily 'interrupted'.

Measures selected for the objective evaluation are:

Estimation accuracy: The distance between a source position and the corresponding localized or tracked estimate.

Estimation ambiguity: The rate of false estimates directed away from sound sources.

Track completeness: The robustness against missing detections in a track or a sequence of localization estimates.

Track continuity: The robustness against fragmentations due to track divergence or swaps affecting a track or a sequence of localization estimates.

Track timeliness: The delay between the speech onset and either the first estimate in a sequence of localization estimates, or at track initialization.

The evaluation measures detailed in the following subsections are defined based on the following nomenclature. A single recording of duration $\mathcal{T}_{\text{rec}}$, including a maximum number of $N_{\max}$ sources, is considered. Each source $n \in \{1, \ldots, N_{\max}\}$ is associated with $A(n)$ periods of activity of duration $\mathcal{T}(a,n) = T_{\text{end}}(a,n) - T_{\text{srt}}(a,n)$ for $a \in \{1, \ldots, A(n)\}$, where $T_{\text{srt}}(a,n)$ and $T_{\text{end}}(a,n)$, respectively, mark the start and end time of the VAP. The corresponding time step indices are $t_{\text{srt}}(a,n) \geq 0$ and $t_{\text{end}}(a,n) \geq t_{\text{srt}}(a,n)$. Each VAP corresponds to an utterance of speech, which is assumed to include both voiced and unvoiced segments. $\Delta_{\text{valid}}(a,n)$ and $L_{\text{valid}}(a,n)$, respectively, denote the duration and the number of time steps in which source n is assigned to a valid track during VAP a. Participants were required to submit azimuth estimates of each source for a sequence of pre-specified time stamps, t, corresponding to the rate of the optical tracking system used for the recordings. Each azimuth estimate had to be labelled by an integer-valued Identity (ID), $k = 1, \ldots, K_{\max}$, where $K_{\max}$ is the maximum number of source IDs in the corresponding recording. Therefore, each source ID establishes an assignment from each azimuth estimate to one of the active sources.

B. Individual Evaluation Measures

To highlight the various scenarios that need to be accounted for during evaluation, consider, for simplicity and without loss of generality, the case of a single-source scenario, i.e., $N_{\max} = 1$, where $N(t) = 1$ during speech activity and $N(t) = 0$ if the source is inactive. A submission either results in $K(t) = 0$, $K(t) = N(t) = 1$ or $K(t) > N(t)$, where $N(t)$ and $K(t)$, respectively, denote the true and estimated number of sources active at t. If $K(t) = 0$, the source is either inactive, i.e., $N(t) = 0$, or the estimate of an active source is missing, if $N(t) = 1$. For $K(t) = 1$, the following scenarios are possible. a) The source is active, i.e., $N(t) = 1$, and the estimate corresponds to a typically imperfect estimate of the ground-truth source direction. b) The source is active, $N(t) = 1$, but its estimate is missing, whereas a false estimate, e.g., pointing towards the direction of an early reflection, is provided. c) The source is inactive, i.e., $N(t) = 0$, and a false estimate is provided. Evaluation measures are therefore required that quantify, per recording, any missing and false estimates as well as the estimation accuracy of estimates in the direction of the source. Prior to performance evaluation, an assignment of each source to a detection must be established by gating and source-to-estimate association, as detailed in Section VI-B1 and Section VI-B2. The resulting assignment is

for evaluation of the estimation accuracy, completeness, continuity, and timeliness (see Section VI-B3 and Section VI-B4).

1) Gating Between Sources and Estimates: Gating [164] provides a mechanism to distinguish between estimation errors, missing, and false estimates. Gating removes improbable assignments of a source with estimates corresponding to errors exceeding a preset threshold. Any estimate removed by gating is counted as a false estimate. If no detection lies within the gate of a source, the source is counted as missed. The gating threshold needs to be selected carefully: If set too low, estimation errors may lead to unassociated sources where a distorted estimate along an existing track is classified as a false estimate and the source estimate is considered as missing. In contrast, if the gating threshold is set too high, a source may be incorrectly assigned to a false track.

For evaluation of the LOCATA challenge, the gating threshold is selected such that the majority of submissions within the single-source Tasks 1 and 3 is not affected. As will be shown in the evaluation in Section VII, a threshold of 30° applied to the azimuth error allows to identify systematic false estimates.

2) Source-to-Estimate Association: For $K(t) > 1$, source localisation may be affected by false estimates both inside and outside the gate. Data association techniques are used to assign the source to the nearest estimate within the gate. Spurious estimates within the gate are included in the set of false estimates. At every time step, a pair-wise distance matrix corresponding to the angular error between each track and each source is evaluated. The optimum source-to-estimate assignment is established using the Munkres algorithm [165] that identifies the source-to-estimate pairs corresponding to the minimum overall distance. Therefore, each source is assigned to at most one track and *vice versa*.

Source-to-estimate association therefore allows to distinguish estimates corresponding to the highest estimation accuracy from spurious estimates. Similar to data association discussed in Section V, and by extension of the single-source case, gating and association establish a one-to-one mapping of each active source with an estimate within the source gate. Any unassociated estimates are considered false estimates, whereas any unassociated sources correspond to missing estimates.

Based on the assignments between sources and estimates, established by gating and association, the evaluation measures are defined to quantify the estimation errors and ambiguities as a single value per measure, per recording. For each assignment between a source and an estimate, the measures detailed in the following are applied to quantify, as a single measure per recording, the estimation accuracy, ambiguity, track completeness, continuity, and timeliness (see Section VI-A).

For brevity, a 'track' is synonymously used in the following to describe both, the trajectory of estimates obtained from a tracker, as well as a sequence of estimates labelled with the same ID by a localization algorithm. The sequence of ground-truth source azimuth values of a source is referred to as the source's ground-truth azimuth trajectory.

3) Estimation Accuracy: The angular errors are evaluated separately in azimuth and elevation for each assigned source-to-track pair for each time stamp during VAPs. The azimuth and elevation error, $d_\phi(\phi(t), \hat{\phi}(t))$ and $d_\theta(\theta(t), \hat{\theta}(t))$, respectively, are defined as:

$$d_\phi\left(\phi(t), \hat{\phi}(t)\right) = \text{mod}\left(\phi(t) - \hat{\phi}(t) + \pi, 2\pi\right) - \pi, \quad (12a)$$

$$d_\theta\left(\theta(t), \hat{\theta}(t)\right) = \theta(t) - \hat{\theta}(t), \quad (12b)$$

where $\text{mod}(q, r)$ denotes the modulo operator for the dividend, q, and the divisor, r; $\phi(t) \in [-\pi, \pi)$ and $\theta(t) \in [0, \pi]$ are the ground-truth azimuth and elevation, respectively; and $\hat{\phi}(t)$ and $\hat{\theta}(t)$ are the azimuth and elevation estimates, respectively.

4) Ambiguity, Track Completeness, Continuity, and Timeliness: In addition to the angular errors, multiple, complementary performance measures are used to quantify estimation ambiguity, completeness, continuity, and timeliness.

At each time step, the number of valid, false, missing, broken, and swapped tracks are counted. Valid tracks are identified as the tracks assigned to a source, whereas false tracks correspond to the unassociated tracks. The number of missing tracks is established as the number of unassociated sources. Broken tracks are obtained by identifying each source that was assigned to a track at $t-1$, but are unassociated at t, where t and $t-1$ must correspond to time steps within the same voice-activity period. Similar to broken tracks, swapped tracks are counted by identifying each source that was associated to track ID $j \in \{1, \ldots, K_{\max}\}$, and is associated to track ID, $\ell \in \{1, \ldots, K_{\max}\}$, where $j \neq \ell$.

Subsequently, the following measures of estimation ambiguity, completeness, continuity, and timeliness are evaluated:

Probability of detection (p_d) [164]: A measure of completeness, evaluating for each source and voice-activity period the percentage of time stamps during which the source is associated with a valid track.

False Alarm Rate (FAR) [166]: A measure of ambiguity, evaluating the number of false estimates per second. The FAR can be evaluated over the duration of each recording [53], in order to provide a gauge of the effectiveness of any Voice Activity Detector (VAD) algorithms that may have been incorporated in a given submitted localization or tracking framework. In addition, the FAR is evaluated in this paper over the duration of each VAP in order to provide a measure of source counting accuracy of each submission.

Track Latency (TL) [166]: A measure of timeliness, evaluating the delay between the onset and the first detection of source n in VAP a.

Track Fragmentation Rate (TFR) [167]: A measure of continuity, indicating the number of track fragmentations per second. The number of fragmentations corresponds to the number of track swaps plus the number of broken tracks.

The evaluation measures defined above therefore quantify errors and ambiguities by single numerical values per measure, per recording. These individual measures can also be used to quantify, across all recordings in each task, the mean of and standard deviation in the estimation accuracy and ambiguity as well as the track completeness, continuity and timeliness.

C. Combined Evaluation Measure

The Optimal SubPattern Assignment (OSPA) metric [168] and its variants, e.g., [169], correspond to a comprehensive measure that consolidates the cardinality error in the estimated number of sources and the estimation accuracy across all sources into a single distance metric at each time stamp of a recording. The OSPA therefore provides a measure that combines the estimation accuracy, track completeness and timeliness. The OSPA selects, at each time stamp, the optimal assignment of the subpatterns between sources and combines the sum of the corresponding cost matrix with the cardinality error in the estimated number of sources. Since the OSPA is evaluated independently of the IDs assigned to the localization and tracking estimates, the measure is agnostic to uncertainties in the identification of track labels.

The OSPA [113], [170] is defined as:

$$\text{OSPA}(\hat{\mathbf{\Phi}}(t), \mathbf{\Phi}(t)) \triangleq$$

$$\left[\frac{1}{K(t)} \min_{\pi \in \mathbf{\Pi}_{K(t)}} \sum_{n=1}^{N(t)} d_c(\phi_n(t), \hat{\phi}_{\pi(n)}(t))^p + (K(t) - N(t))c^p\right]^{\frac{1}{p}}, \tag{13}$$

for $N(t) \leq K(t)$, where $\hat{\mathbf{\Phi}}(t) \triangleq \{\hat{\phi}_1(t), \ldots, \hat{\phi}_{K(t)}(t)\}$ denotes the set of $K(t)$ track estimates; $\mathbf{\Phi}(t) \triangleq \{\phi_1(t), \ldots, \phi_{N(t)}(t)\}$ denotes the set of $N(t)$ ground-truth sources active at t; $1 \leq p < \infty$ is the order parameter; c is the cutoff parameter; $\mathbf{\Pi}_{K(t)}$ denotes the set of permutations of length $N(t)$ with elements $\{1, \ldots, K(t)\}$ [170]; $d_c(\phi_n(t), \hat{\phi}_{\pi(n)}(t)) \triangleq \min(c, \text{abs}(d_\phi(\phi_n(t), \hat{\phi}_{\pi(n)}(t))))$, where $\text{abs}(\cdot)$ denotes the absolute value; $d_\phi(\cdot)$ is the angular error (see (12)); and $\pi(n)$ denotes the n^{th} element of each subset $\pi \in \mathbf{\Pi}$. For $N(t) > K(t)$, the OSPA distance is evaluated as $\text{OSPA}(\mathbf{\Phi}(t), \hat{\mathbf{\Phi}}(t))$ [170]. The impact of the choice of p and c is discussed in [168]. In this paper, $c = 30°$.

To provide further insight into the OSPA measure, we note that the term $\frac{1}{K(t)} \min_{\pi \in \mathbf{\Pi}_{K(t)}} \sum_{n=1}^{N(t)} d_c(\phi_n(t), \hat{\phi}_{\pi(n)}(t))^p$ evaluates the average angular error by comparing each angle estimate against every ground-truth source angle. The OSPA is therefore agnostic of the estimate-to-source association. The cardinality error is evaluated as $K(t) - N(t)$. The order parameter, p, determines the weighting of the angular error relative to the cardinality error.

Due to the dataset size of the LOCATA corpus, a comprehensive analysis of the OSPA at each time stamp for each submission, task, array, and recording is impractical. Therefore, the analysis of the LOCATA challenge results is predominantly based on the mean and variance of the OSPA across all time stamps and recordings for each task.

VII. Evaluation Results

The following section presents the performance evaluation for the LOCATA challenge submissions using the measures detailed in Section VI. The evaluation in Section VII-A focuses on the single-source tasks 1, 3 and 5. Section VII-B presents the results for the multi-source tasks 2, 4 and 6.

The evaluation framework establishes an assignment between each ground-truth source location and a source estimate for every time stamp during voice-active periods in each recording, submission, task, and array (see Section VI). The azimuth error in (12a) between associated source-to-track pairs is averaged over all time stamps and all recordings. The resulting average azimuth errors for each task, submission, and array are provided in Table III. The baseline (BL) corresponds to the MUSIC implementation as detailed in [19]. One submission (ID 5) is not included in the discussion as details of the method are not available at the time of writing. Two further submissions (ID 13 and ID 14) are also not included due to inconclusive results.

A. Single-Source Tasks 1, 3, 5

1) Task 1 - Azimuth Accuracy: For Task 1, involving a single, static source and a static microphone array, average azimuth accuracies of around 1° can be achieved (see Table III). Notably, Submission 3 results in 1.0° using the DICIT array by combining TDE with a particle filter for tracking; Submission 11 results in an average azimuth accuracy of 0.7° using the robot head; and Submission 12 achieves an accuracy of 1.1° using the Eigenmike. Submissions 11 and 12 are MUSIC implementations, applied to the microphone signals in the STFT domain and domain of spherical harmonics, respectively.

A possible reason for the performance of Submissions 11 and 12 is that MUSIC does not suffer from spatial aliasing if applied to arrays that incorporate a large number of microphones. As such, the overall array aperture can be small for low noise levels. Therefore, the performance of the two MUSIC-based Submissions 11 (robot head) and 12 (Eigenmike) is comparable. Moreover, for the Eigenmike, Submission 12 (1.1°) leads to improvements of the SRP-based Submissions 6 (6.4°) and 7 (7.0°).

For the pseudo-intensity-based approaches that were applied to the Eigenmike, Submission 10 achieves an azimuth accuracy of 8.9° by extracting pseudo-intensity vectors from the first-order ambisonics and applying a particle filter for tracking. Submission 15, which extracts the pseudo-intensity from the signals in the domain of spherical harmonics and applies subspace-based processing, results in 8.1°. The pseudo-intensity-based Submissions 10 and 15 lead to a performance degradation of approximately 7°, compared to the MUSIC-based Submission 12, also applied in the domain of spherical harmonics. The reduced accuracy may be related to the resolution of the spatial spectra provided by the pseudo-intensity-based approaches compared to MUSIC. The spatial spectrum is computed using MUSIC by scanning each direction in a discrete grid, specified by the steering vector. In contrast, pseudo-intensity-based approaches approximate the spatial spectrum by effectively combining the output of three dipole beamformers, steered along the x-, y-, and z-axis relative to the array. Therefore, compared to MUSIC, pseudo-intensity approaches evaluate a coarse approximation of the spatial spectrum, but require reduced computational load.

A performance degradation from the 12-channel robot head to the 32-channel Eigenmike is observed for the submissions that involved both arrays. For ground-truth acquisition using

TABLE III
AVERAGE AZIMUTH ERRORS DURING VAP. SUBMISSIONS CORRESPONDING TO MINIMUM AVERAGE ERRORS ARE HIGHLIGHTED IN BOLD FONT. COLUMN COLOUR INDICATES TYPE OF ALGORITHM, WHERE WHITE INDICATES FRAMEWORKS INVOLVING ONLY DOA ESTIMATION (SUBMISSION IDS 1, 6, 9, 11, 12, 15, 16 AND THE BASELINE (BL)), AND GREY INDICATES FRAMEWORKS THAT COMBINE DOA ESTIMATION WITH SOURCE TRACKING (SUBMISSION IDS 2, 3, 4, 7, 8, 10)

Task		Array	Submission ID													
			1	2	3	4	6	7	8	9	10	11	12	15	16	BL
Single Source	1	Robot Head	-	-	-	2.1	1.5	1.8	-	-	-	**0.7**	-	-	-	4.2
		DICIT	-	-	**1.0**	-	-	2.2	-	9.1	-	-	-	-	-	12.3
		Hearing Aids	**8.5**	-	-	-	-	-	8.7	-	-	-	-	-	-	15.9
		Eigenmike	-	-	-	-	6.4	7.0	-	-	8.9	-	**1.1**	8.1	-	10.2
	3	Robot Head	-	-	-	4.6	3.2	**3.1**	-	-	-	-	-	-	-	9.4
		DICIT	-	-	**1.8**	-	-	4.5	-	-	-	-	-	-	-	13.9
		Hearing Aids	-	-	-	-	-	-	**7.2**	-	-	-	-	-	-	16.0
		Eigenmike	-	-	-	-	**8.1**	9.3	-	-	11.5	-	-	-	-	17.6
	5	Robot Head	-	-	-	4.9	**2.2**	3.7	-	-	-	-	-	-	-	5.4
		DICIT	-	-	**2.7**	-	-	3.4	-	-	-	-	-	-	-	13.4
		Hearing Aids	-	-	-	-	-	-	**11.8**	-	-	-	-	-	-	14.6
		Eigenmike	-	-	-	-	**6.3**	7.5	-	-	-	-	-	-	-	12.9
Multiple Sources	2	Robot Head	-	-	-	3.8	-	-	-	-	-	**2.0**	-	-	-	9.0
		DICIT	-	-	-	-	-	-	-	-	-	-	-	-	-	**11.0**
		Hearing Aids	-	-	-	-	-	-	-	-	-	-	-	-	-	**15.6**
		Eigenmike	-	-	-	-	-	-	-	-	7.3	-	**1.4**	-	7.1	10.2
	4	Robot Head	-	9.4	-	**6.0**	-	-	-	-	-	-	-	-	-	9.2
		DICIT	-	13.5	-	-	-	-	-	-	-	-	-	-	-	**12.9**
		Hearing Aids	-	13.8	-	-	-	-	-	-	-	-	-	-	-	**13.7**
		Eigenmike	-	12.8	-	-	-	-	-	-	**9.0**	-	-	-	-	11.8
	6	Robot Head	-	-	-	**8.1**	-	-	-	-	-	-	-	-	-	8.5
		DICIT	-	-	-	-	-	-	-	-	-	-	-	-	-	**13.9**
		Hearing Aids	-	-	-	-	-	-	-	-	-	-	-	-	-	**13.9**
		Eigenmike	-	-	-	-	-	-	-	-	-	-	-	-	-	**12.9**

the OptiTrack system, the reflective markers were attached to the shockmount of the Eigenmike, rather than the baffle of the array, to minimize shadowing and scattering effects, see [17], [18]. Therefore, a small bias in the DoA estimation errors is possible due to rotations of the array within the shockmount. Nevertheless, this bias is expected to be significantly smaller than some of the errors observed for the Eigenmike in Table III. Possible reasons are that 1) the irregular array topology of the robot head may lead to improved performance for some of the algorithms, or that 2) the performance improvements in localization accuracy may be related to the larger array aperture of the robot head, compared to the Eigenmike. However, with the remaining uncertainty regarding the actual implementation of the algorithms, conclusions remain somewhat speculative at this point.

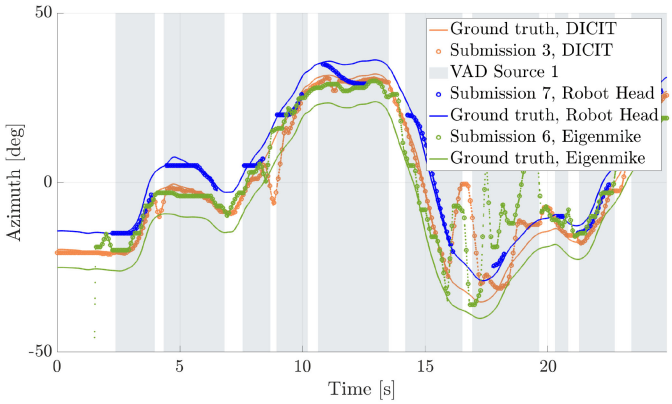
Submission 6, applying SRP-PHAT to a selection of microphone pairs, results in average azimuth errors of 1.5° using the robot head and 6.4° using the Eigenmike. Similar results of 1.8° and 7.0° for the robot head and Eigenmike, respectively, are obtained using Submission 7, which combine an SRP beamformer for localization with a Kalman filter for tracking. Therefore, the SRP-based approaches in Submissions 6 and 7, applied without and with tracking, respectively, lead to comparably accurate results.

Table III also highlights a significant difference in the performance results between the approaches submitted to Task 1 using the DICIT array. Submission 3 achieves an average azimuth accuracy of 1.0° by combining GCC-PHAT with a particle filter. Submission 7, combining SRP beamforming and a Kalman filter, results in a small degradation to 2.2° in average azimuth accuracy. Submission 9 leads to a decreased accuracy of 9.1°. Submission 3 uses the subarray of microphone pairs corresponding to 32 cm spacings to exploit spatial diversity between the microphones; Submission 7 uses the 7-microphone linear subarray at the array centre; Submission 9 uses three microphones at the centre of the array, with a spacing of 4 cm, to form two microphone pairs. A reduction of the localization accuracy can therefore be intuitively expected for Submission 9, compared to Submissions 3 and 7, due to a) the reduced number of microphones, and b) the reduced inter-microphone spacing, and hence reduced spatial diversity of the sensors.

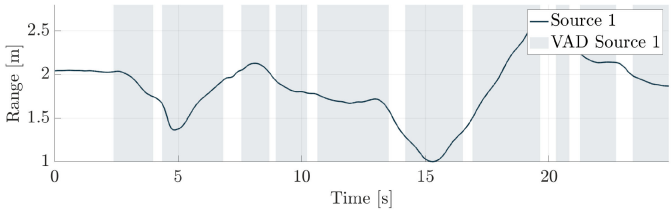
For the hearing aids in Task 1, both Submissions 1 and 8 result in comparable azimuth errors of 8.5° and 8.7° respectively. The recordings for the hearing aids were performed separately from the remaining arrays, and are therefore not directly comparable to the results for other arrays. Nevertheless, a reduction in azimuth accuracy for the hearing aids is intuitively expected due to the small number of microphones integrated in each of the arrays.

To conclude, we note that the results for the static single-source Task 1 indicate a comparable performance between the submissions that incorporate localization and those submissions that combine localization with source tracking. Since the source is static, long blocks of data can be used for localization. Furthermore, temporal averaging can be applied across data blocks. Therefore, since a dynamical model is not required for the static single-source scenario, localization algorithms can apply smoothing directly to the DoA estimate, without the need for explicit source tracking.

2) Task 3 - Azimuth Accuracy: In the following, $\mathcal{S}_{135} = \{3, 4, 6, 7, 8\}$ denotes the set of submissions that were evaluated for Tasks 1, 3 and 5. For Task 3, involving a single, moving



(a) Azimuth ground-truth and estimates



(b) Ground-truth range between source and robot head

Fig. 4. Azimuth estimates for Task 3, recording 4 for (a) azimuth estimates for Submissions 3, 6, 7. As a reference, the ground-truth range between the robot head and the source is shown in (b).

source, a small degradation is observed in the azimuth error over $\mathcal{S}_{135}$ from 4.3° for Task 1 to 5.5° for Task 3. For example, Submission 7 leads to the lowest average absolute error in azimuth with only 3.1° for Task 3 using the robot head, corresponding to a degradation of 1.3° compared to Task 1. The accuracy of Submission 3 reduces from 1.0° for Task 1 to 1.8° for Task 3.

The reduction in azimuth accuracy from static single-source Task 1 to moving single-source Task 3 is similar for all submissions. Trends in performance between approaches for each array are identical to those discussed for Task 1. The overall degradation in performance is therefore related to differences in the scenarios between Task 1 and Task 3. Recordings from human talkers are subject to variations in the source orientation and source-sensor distance. The orientation of sources directed away from the microphone array leads to a decreased direct-path contribution to the received signal. Furthermore, with increasing source-sensor distance, the noise field becomes increasingly diffuse. Hence, reductions in the Direct-to-Reverberant Ratio (DRR) [23] due to the source orientation, as well as the CDR due to the source-sensor distance, result in increased azimuth estimation errors.

To provide further insight into the results for Task 3, Fig. 4 provides a comparison for recording 4 of the approaches leading to the highest accuracy for each array, i.e., Submission 7 using the robot head, Submission 3 using the DICIT array, and Submission 6 using the Eigenmike. For Submission 7, accurate and smooth tracks of the azimuth trajectories are obtained during VAPs. Therefore, diagonal unloading SRP beamforming clearly provides power maps of sufficiently high resolution to provide accurate azimuth estimates whilst avoiding systematic false detections in the directions of early reflections. Moreover, application of the Kalman filter provides smooth azimuth trajectories.

Similar results in terms of the azimuth accuracy are obtained for Submission 3, combining GCC-PHAT with a particle filter for the DICIT array. However, due to the lack of a VAD, temporary periods of track divergence can be observed for Submission 3 around periods of voice inactivity, i.e., between [3.9, 4.4] s and [8.5, 9.2] s.

For the voice-active period between [16.9, 19.6] s, the results of Submission 7 are affected by a significant number of missing detections, whilst the results for Submission 3 exhibits diverging track estimates. Fig. 4(b) provides a plot of the range between the source and robot head, highlighting that the human talker is moving away from the arrays between [15.1, 20] s. Therefore, the Cross-Power Spectral Density (CPSD)-based VAD algorithm of Submission 7 results in missing detections of voice activity with decreasing CDR. For Submission 3 and 6, that do not involve a VAD, the negative DRR leads to missing and false DoA estimates in the direction of early reflections. Therefore, increasing DoA estimation errors are observed in voice-active periods during which the source-sensor distance increases beyond 2 m.

3) Task 5 - Azimuth Accuracy: The mean azimuth accuracy over $\mathcal{S}_{135}$, averaged over the corresponding submissions and arrays, decreases from 5.5° for Task 3, using static arrays, to 9.7° for Task 5, using moving arrays. Despite the reduced number of submissions for Task 5, the overall performance trends are similar to those in Task 1 and Task 3 (see Table III).

The trend of an overall performance degradation is related to the increasingly challenging conditions. Similar to Task 3, the motion of the source and arrays lead to time-varying source-sensor distances and source orientations relative to the array. Furthermore, due to the motion of the array, it is crucial that the microphone signals in Task 5 are processed over analysis windows of sufficiently short duration.

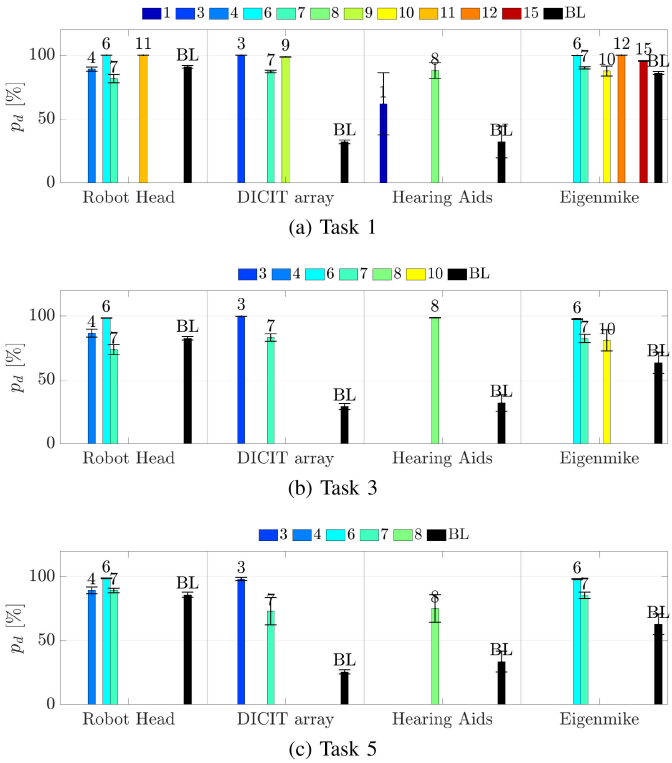
4) Tasks 1, 3, 5. Impact of Gating on Azimuth Accuracy: To illustrate the effect of gating on the evaluation results, the evaluation was repeated without gating by assigning each source to its closest estimate.[2] Table IV provides the difference in the average azimuth errors with and without gating. In Table IV, entries with value 0.0 indicate that evaluation with and without gating lead to the same result. Entries with values greater than 0.0 highlight that the azimuth error increases without gating, i.e., the submitted results are affected by outliers outside of the gating collar.

For the majority of submissions, a gating threshold of 30° results in improved azimuth accuracies in the range of 0.1° to 4° across Tasks 1, 3 5. A significant number of outliers are observed for Submissions 1, 7 and 8. To reflect outliers in the analysis of the results, evaluation measures, such as the FAR and probability of detection, are required in addition to the average azimuth error.

[2]Even though Tasks 1, 3 and 5 correspond to single-source scenarios, gating and association is required for evaluation, since azimuth estimates corresponding to multiple source IDs were provided for some submissions.

TABLE IV
DIFFERENCE IN AVERAGE AZIMUTH ERRORS WITH AND WITHOUT GATING, EVALUATED FOR SINGLE-SOURCE TASKS 1, 3, 5 FOR ALL SUBMISSIONS AND THE BASELINE (BL). SUBMISSIONS UNAFFECTED BY GATING, AND HENCE OUTLIERS, ARE HIGHLIGHTED IN BOLD FONT

Task		Array	Submission ID													
			1	2	3	4	6	7	8	9	10	11	12	15	16	BL
Single Source	1	Robot Head	-	-	-	**0.0**	**0.0**	**0.0**	-	-	-	**0.0**	-	-	-	0.2
		DICIT	-	-	**0.0**	-	-	**0.0**	-	0.5	-	-	-	-	-	49.6
		Hearing Aids	42.3	-	-	-	-	-	**4.0**	-	-	-	-	-	-	49.2
		Eigenmike	-	-	-	-	0.1	0.1	-	-	**0.0**	-	**0.0**	**0.0**	-	0.4
	3	Robot Head	-	-	-	**0.0**	1.2	**0.0**	-	-	-	-	-	-	-	3.4
		DICIT	-	-	**0.0**	-	-	**0.0**	-	-	-	-	-	-	-	63.2
		Hearing Aids	-	-	-	-	-	-	**0.4**	-	-	-	-	-	-	46.8
		Eigenmike	-	-	-	-	0.6	**0.2**	-	-	1.6	-	-	-	-	8.3
	5	Robot Head	-	-	-	**0.1**	0.8	1.2	-	-	-	-	-	-	-	1.8
		DICIT	-	-	**0.6**	-	-	16.7	-	-	-	-	-	-	-	53.8
		Hearing Aids	-	-	-	-	-	-	**12.7**	-	-	-	-	-	-	43.7
		Eigenmike	-	-	-	-	**1.1**	1.9	-	-	-	-	-	-	-	14.9

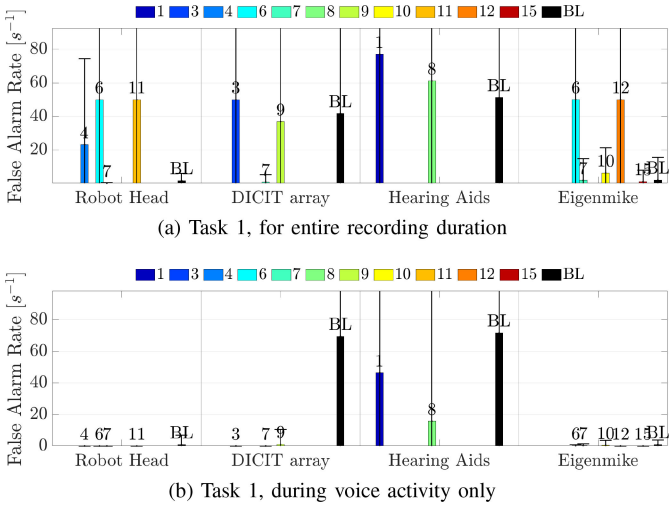


(a) Task 1

(b) Task 3

(c) Task 5

Fig. 5. Probability of detection (bars) and standard deviation over recordings (whiskers) for Tasks 1, 3, 5, for each submission and array. Legends indicate the submission IDs available for each of the tasks.



(a) Task 1, for entire recording duration

(b) Task 1, during voice activity only

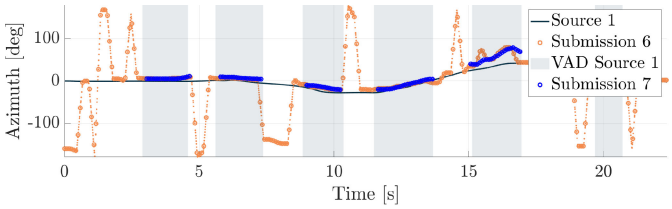
Fig. 6. FAR for Task 1 involving single static loudspeakers (a) for entire recording duration, and (b) during voice-activity periods only.

5) Completeness and Ambiguity: As detailed in Section VI, the track cardinality and probability of detection are used as evaluation measures of the track completeness. For single-source scenarios, the track completeness quantifies the robustness of localization and tracking algorithms against changes in the source orientation and source-sensor distance. Furthermore, the FAR is used as an evaluation measure of the track ambiguity, quantifying the robustness against early reflections and noise in the case of the single-source scenarios.

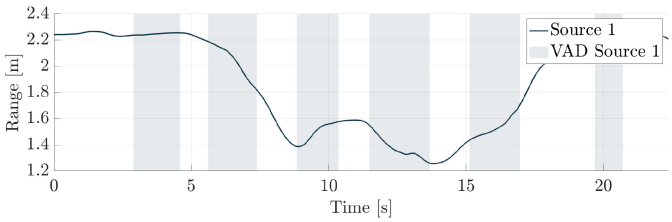
The probability of detection and FAR, averaged over all recordings in each task, are shown in Fig. 5 and Fig. 6, respectively. The results indicate that the probability of detection between Tasks 1, 3 and 5 remains approximately constant, with a trend towards a small reduction in p_d, when changing from static to dynamic sources.

The results also highlight that Submissions 11 and 12, corresponding to the highest average azimuth accuracy for Task 1 using the robot head and Eigenmike (see Section VII-A1), exhibit 100% probability of detection. The same submissions also correspond to a comparatively high FAR of 50 false estimates per second, averaged across all recordings for Task 1 and evaluated for the full duration of each recording (see Fig. 6(a)). These results are indicative of the fact that Submissions 11 and 12 do not incorporate VAD algorithms. For comparison, Fig. 6(b) depicts the average FARs for Task 1 evaluated during voice-activity only. The results in Fig. 6(b) clearly highlight a significant reduction in the FAR for Submissions 3, 6, 11, which do not incorporate VAD.

Fig. 7(a), selected from Submission 6 for Task 3 and recording 2, shows that estimates during periods of voice inactivity are affected by outliers, which are removed from the measure for azimuth accuracy due to the gating process, and are accounted for in the FAR. The majority of DoA estimates provided during voice-activity correspond to smooth tracks near the ground-truth source azimuth. In the time interval [15.1, 17] s, the estimates

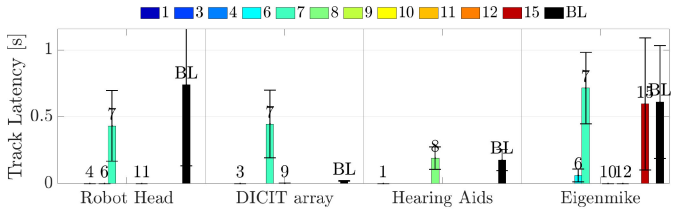


(a) Azimuth ground-truth for Source 1 and estimates

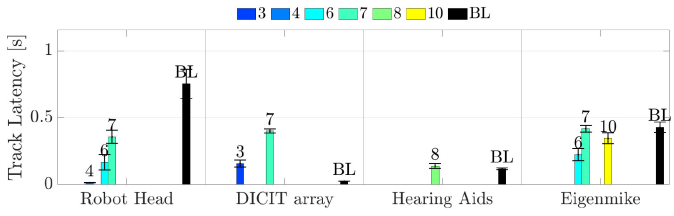


(b) Ground-truth source-sensor range

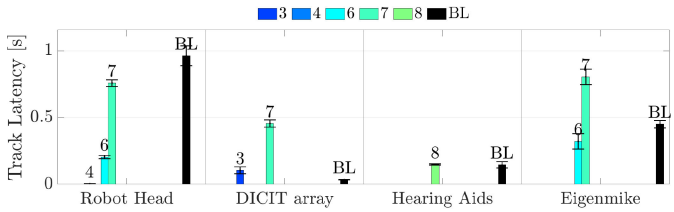
Fig. 7. Comparison of (a) azimuth estimates for Task 3, recording 2 using the Eigenmike for Submissions 6, 7, and (b) ground-truth range between the source and the Eigenmike array origin. Results indicate outliers during voice inactivity for Submission 6 and temporary track divergence during voice activity between [15.1, 17] s for Submissions 6 and 7.



(a) Task 1



(b) Task 3



(c) Task 5

Fig. 8. Track latency (bars) and standard deviation over recordings (whiskers) for Tasks 1, 3 and 5, for each submission and array. Legends indicate the submission IDs available for each of the tasks.

exhibit a temporary period of track divergence. The results for Submission 7 in Fig. 7(a) highlight that outliers during voice inactivity are avoided since the submission incorporates VAD. The results also indicate diverging track estimates in the interval [15.1, 17] s. The track divergence affecting both submissions is likely caused by the time-varying source-sensor geometry due to the motion of the source. Fig. 7(b) highlights that the source is moving away from the array after 13 s. As the source orientation is directed away from the array, the contribution of the direct-path signal decreases, resulting in reduced estimation accuracy in the source azimuth. The reduction in azimuth accuracy eventually results in false estimates outside of the gating threshold.

6) Timeliness: The track latency is used as an evaluation measure of the timeliness of localization and tracking algorithms. Therefore, the track latency quantifies the sensitivity of algorithms to speech onsets, and the robustness against temporal smearing at speech endpoints.

Fig. 8 shows the track latency, averaged across all recordings for Tasks 1, 3 and 5. Submissions 1, 3, 6, 8, 9, 11 and 12 do not incorporate VAD. Hence, estimates are provided at every time stamp for all recordings. Submissions 3 and 8 incorporate tracking algorithms, where the source estimates are propagated through voice-inactive periods by track prediction. Submissions 1, 11 and 12, submitted for only the static tasks, estimate the average azimuth throughout the full recording duration and extrapolate the estimates across all time steps.

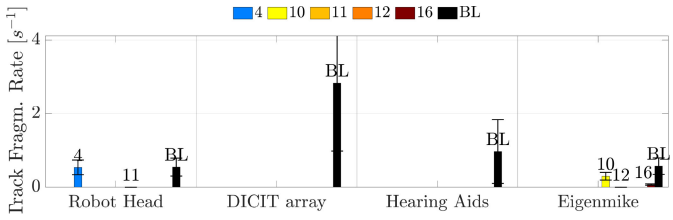
Therefore, for Task 1, Submissions 1, 3, 11 and 12 correspond to 0 s track latency throughout. However, these algorithms also correspond to high FARs, when the FAR is evaluated across voice-active and inactive periods (see Fig. 6(a)). Submissions 3 and 8, which do not involve a VAD and were submitted to the tasks involving moving sources, result in track latencies of below 0.2 s for Tasks 3 and 5, where the extrapolation of tracks outside of VAPs is non-trivial.

Submission 4 incorporates a VAD that estimates voice activity as a side-product of the variational EM algorithm for tracking. The results show that Submission 4 effectively detects speech onsets, leading to negligible track latencies across Tasks 1, 3 and 5. Submission 10, incorporating the noise Power Spectral Density (PSD)-based VAD of [171], detects speech onsets accurately in the static source scenario in Task 1. However, the track latency for Task 3, involving a moving source, increases to 0.35 s. It is important to note that Submissions 7 and 10 incorporate Kalman or particle filters with heuristic approaches to track initialization. Therefore, it is likely that track initialization rules - rather than the VAD algorithms - lead to delays in the confirmation of newly active sources.

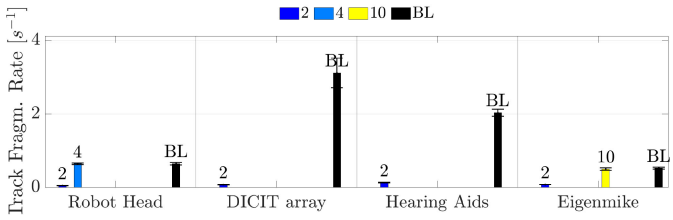
B. Multi-Source Tasks 2, 4, 6

1) Accuracy: For the multi-source Tasks 2, 4 and 6, the results in Table III indicate similar trends as discussed for the single-source Tasks 1, 3 and 5. However, the overall performance of all submissions for Tasks 2, 4 and 6 is decreased compared to Tasks 1, 3 and 5.

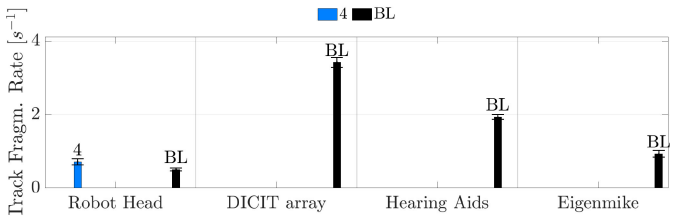
The reduction in azimuth accuracy is due to the adverse effects of interference from multiple simultaneously active sound sources. Due to the broadband nature of speech, the speech signals of multiple talkers often correspond to energy in the overlapping time-frequency bins, especially for talkers with similar voice pitch. Therefore, localization approaches that rely



(a) Task 2, Track Fragmentation Rate



(b) Task 4, Track Fragmentation Rate



(c) Task 6, Track Fragmentation Rate

Fig. 9. Track fragmentation rate (bars) and standard deviation over recordings (whiskers) for Tasks 2, 4, 6, for each submission and array.

on the W-disjoint orthogonality of speech may result in biased estimates of the DoA (see, e.g., Submission 4).

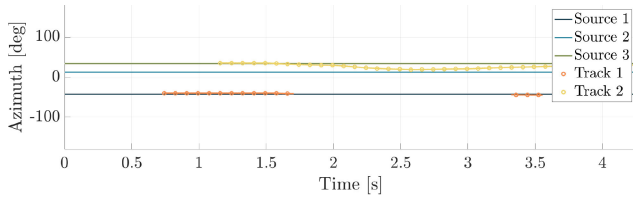
Robustness against interference can be achieved by incorporating time-frequency bins containing the contribution of a single source only, e.g., at the onset of speech. For example, Submission 11 and 12 incorporate the Direct Path Dominance (DPD)-test in [112], and result in azimuth accuracies of 2.0° and 1.4°, respectively, for the robot head and Eigenmike in Task 2, compared to 0.7° and 1.1° in Task 1.

An increasing number of sources also results in an increasingly diffuse sound field in reverberant environments. For data-dependent beamforming techniques [1], the directivity pattern of the array is typically evaluated based on the signal and noise levels. For increasing diffuse noise, it is therefore expected that the performance of beamforming techniques decreases in multi-source scenarios.

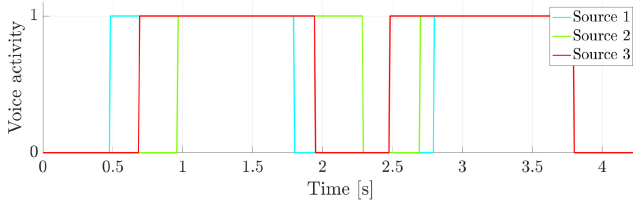
In addition to a reduction in the angular accuracy, ambiguities arising in scenarios involving multiple, simultaneously active sound sources result in missing and false DoA estimates, affecting the completeness, continuity, and ambiguity of localization and tracking approaches.

2) Continuity: The TFR is used as an evaluation measure for track continuity (see Section VII). Fig. 9 provides the TFRs for Tasks 2, 4 and 6 for each array and submission and averaged over the recordings.

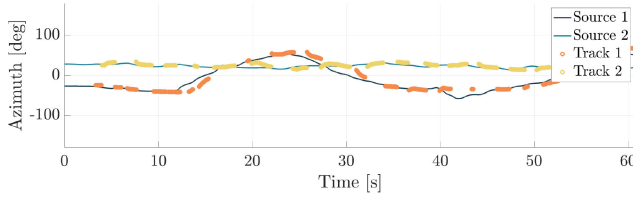
The results indicate that the subspace-based Submissions 11, 12 and 16 are robust to track fragmentation. Although the submissions rely on the assumption of W-disjoint orthogonal



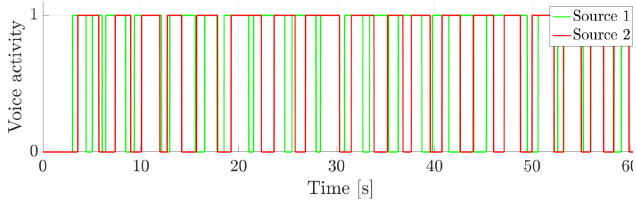
(a) Azimuth estimates: Task 2, Recording 5



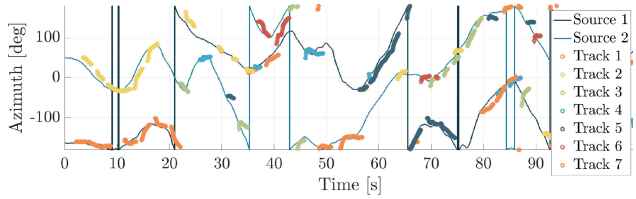
(b) VAD: Task 2, Recording 5



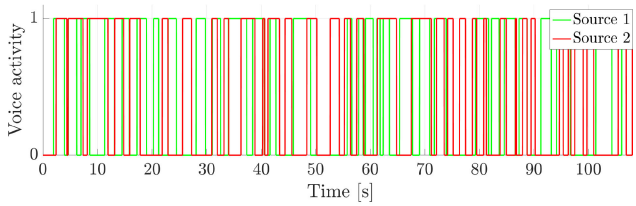
(c) Azimuth estimates: Task 4, Recording 4



(d) VAD: Task 4, Recording 4



(e) Azimuth estimates: Task 6, Recording 2



(f) VAD: Task 6, Recording 2

Fig. 10. Azimuth estimates and VAD for Submission 4 using the robot head for (a)–(b) Task 2, (c)–(d) Task 4, and (e)–(f) Task 6.

sources, localization is performed only on a subset of frequency bins that correspond to the contribution of a single source. In contrast, BSS-based approaches assume that the W-disjoint orthogonality applies to all frequency bands required for the reconstruction of the source signals.

TABLE V
AVERAGE OSPA RESULTS. COLUMN COLOUR INDICATES TYPE OF ALGORITHM, WHERE WHITE INDICATES FRAMEWORKS INVOLVING ONLY DOA ESTIMATION (SUBMISSION IDS 11, 12, 16 AND THE BASELINE (BL)), AND GREY INDICATES FRAMEWORKS THAT COMBINE DOA ESTIMATION WITH SOURCE TRACKING (SUBMISSION IDS 2, 4, 10)

Task	Array	Submission ID													
		2		4		10		11		12		16		BL	
		p		p		p		p		p		p		p	
		1	5	1	5	1	5	1	5	1	5	1	5	1	5
2	Robot Head	-	-	17.5	22.4	-	-	12.4	17.6	-	-	-	-	19.5	23.8
	DICIT	-	-	-	-	-	-	-	-	-	-	-	-	26.6	28.0
	Hearing Aids	-	-	-	-	-	-	-	-	-	-	-	-	26.1	27.7
	Eigenmike	-	-	-	-	17.5	22.3	-	-	12.2	17.3	12.4	18.2	21.5	25.0
4	Robot Head	13.8	18.9	13.5	16.4	-	-	-	-	-	-	-	-	16.3	18.9
	DICIT	15.6	20.0	-	-	-	-	-	-	-	-	-	-	25.8	26.6
	Hearing Aids	15.2	19.6	-	-	-	-	-	-	-	-	-	-	27.7	28.1
	Eigenmike	14.6	19.3	-	-	13.1	16.4	-	-	-	-	-	-	18.4	20.8
6	Robot Head	-	-	13.8	15.0	-	-	-	-	-	-	-	-	14.8	15.8
	DICIT	-	-	-	-	-	-	-	-	-	-	-	-	24.8	25.2
	Hearing Aids	-	-	-	-	-	-	-	-	-	-	-	-	25.2	25.8
	Eigenmike	-	-	-	-	-	-	-	-	-	-	-	-	21.1	21.7

The advantage of subspace-based processing for robustness against track fragmentation is reinforced when comparing the results for Submission 10, based on pseudo-intensity vectors for ambisonics, against Submission 16, using subspace pseudo-intensity vectors in the domain of spherical harmonics. The azimuth accuracies of both submissions are comparable, where Submission 10 results in an average azimuth error of 7.3° and Submission 16 leads to 7.1° in Task 2. In contrast, Submission 10 leads to 0.3 fragmentations per second, whereas Submission 16 exhibits only 0.07 fragmentations per second.

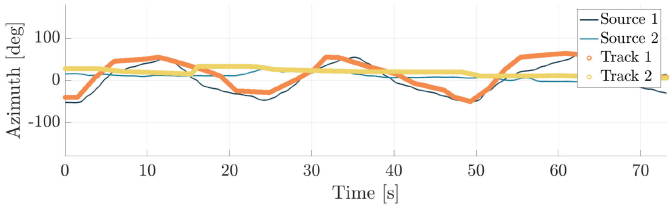
Comparing the results for static Task 2 against the moving-source Task 4 and the fully dynamic Task 6, the results in Fig. 9 highlight increasing TFRs across submissions. For example, Submission 4, the only approach that was submitted for all three multi-source tasks, corresponds to 0.53 fragmentations per second for Task 2, involving multiple static loudspeakers, to 0.64 fragmentations per second for Task 4, involving multiple moving human talkers, and to 0.71 fragmentations per second for Task 6 involving multiple moving human talkers and moving arrays. The increasing TFR is due to the increasing spatio-temporal variation of the source azimuth between the three tasks. Task 2 corresponds to constant azimuth trajectories of the multiple static loudspeakers, observed from static arrays (see Fig. 10(a), showing the azimuth estimates for Task 2, recording 5). The motion of the human talkers that are observed from static arrays in Task 4 correspond to time-varying azimuth trajectories within limited intervals of azimuth values. For example, for Task 4, recording 4 shown in Fig. 10(c), source 1 is limited to azimuth values in the interval between [6, 24]°, whilst source 2 is limited between $[-66, 50]^\circ$. The motion of the moving sources and moving arrays in Task 6 result in azimuth trajectories that vary significantly between $[-180, 180]^\circ$ (see Fig. 10(e) for the azimuth estimates provided for Task 6, recording 2). Furthermore, the durations of recordings for Task 4 and Task 6 are substantially longer than those for Task 2. As to be expected, periods of speech inactivity and the increasing time-variation of the source azimuth relative to the arrays result in increasing TFRs when comparing Task 2, Task 4, and Task 6.

3) OSPA - Accuracy vs. Ambiguity, Completeness and Continuity: The results for the OSPA measure, averaged over all recordings for the multi-source Tasks 2, 4 and 6, is summarized for order parameters $p = \{1, 5\}$ (see (13)) in Table V. In contrast to the averaged azimuth errors in Table III, the OSPA results trade off the azimuth accuracy against cardinality errors, and hence false and missing track estimates. For example, the results for Task 2 in Table III indicate a significant difference in the results for Submission 12 (1.4°) and Submission 16 (7.1°). In contrast, due to false track estimates during periods of voice inactivity, Table V highlights only a small difference between the OSPA for Submissions 12 and 16.

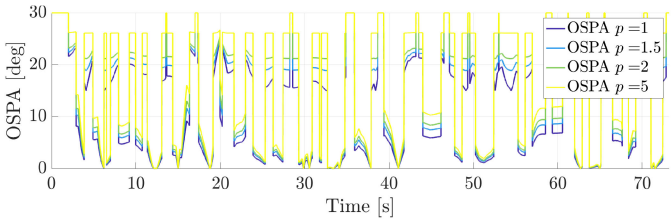
To provide intuitive insight into the OSPA results and the effect of the order parameter, p, Fig. 11 compares the azimuth estimates obtained using Submissions 2 and 10 for the Eigenmike, Task 4, Recording 1.

The results highlight distinct jumps of the OSPA between periods during which a single source is active and the onsets of periods of two simultaneously active sources. During periods of voice inactivity, detection errors in the onsets of speech lead to errors corresponding to the cutoff threshold of $c = 30^\circ$. Therefore, the cardinality error dominates the OSPA when $N(t) = 0$ and $K(t) > 0$. During VAPs where $N(t) = K(t)$, the OSPA is dominated by the angular error between each estimate and the ground-truth direction of each source, resulting in values in the range of [0, 20]°. For $N(t) = K(t)$, the order parameter, p, does not affect the results since the cardinality error is $K(t) - N(t) = 0$. During periods where $K(t) < N(t)$, the cardinality error causes the OSPA to increase to between [15, 30]°. The OSPA increases with the order parameter p.

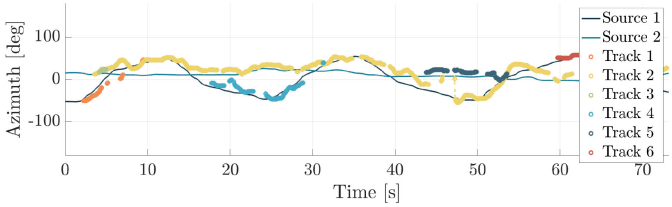
The results highlight that both approaches are affected by cardinality errors, indicated by jumps in the OSPA. For Submission 10, which incorporates VAD, the cardinality errors arise predominantly due to missing detections and broken tracks (see Fig. 11(d)). In contrast, Submission 2 is mainly affected by false estimates during voice inactivity. Since Submission 2 does not involve a VAD, tracks are propagated through periods of voice inactivity using the prediction step of the tracking filter. Temporary periods of track divergence therefore lead to



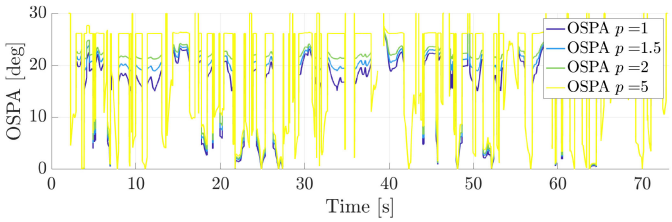
(a) Submission 2, Eigenmike



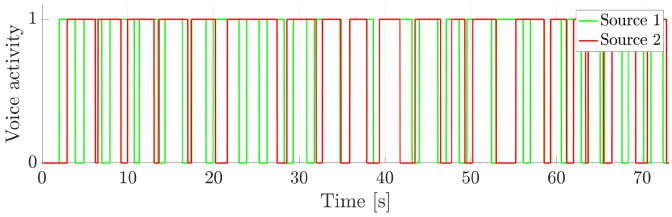
(b) Submission 2, Eigenmike



(c) Submission 10, Eigenmike



(d) Submission 10, Eigenmike



(e) VAD, Eigenmike

Fig. 11. Azimuth trajectories and corresponding OSPA metric for recording 1 of Task 4 for (a)–(b) Submission 2 using the Eigenmike, (c)–(d) Submission 10 using the Eigenmike. The VAD periods are shown in (e).

estimates that are classified as false estimates by gating and data association.

VIII. Conclusion

The open-access LOCATA challenge data corpus of real-world, multichannel audio recordings and open-source evaluation software provides a framework to objectively benchmark state-of-the-art localization and tracking approaches. The challenge consists of six tasks, ranging from the localization and tracking of a single static loudspeaker using static microphone arrays to fully dynamic scenes involving multiple moving sources and microphone arrays on moving platforms. Sixteen state-of-the-art approaches were submitted for participation in the LOCATA challenge, one of which needed to be discarded for evaluation due to the lack of documentation. Seven submissions corresponded to sound source localization algorithms, obtaining instantaneous estimates at each time stamp of a recording. The remaining submissions combined localization algorithms with source tracking, where spatio-temporal models of the source motion are applied in order to exploit constructively knowledge of the history of the source trajectories. The submissions incorporated localization algorithms based on time-delay estimation, subspace processing, beamforming, classification, and deep learning. Source tracking submissions incorporated the Kalman filter and its variants, particle filters, variational Bayesian approaches and PHD filters.

The controlled scenarios of static single-source in Task 1 are used to evaluate the robustness of the submissions against reverberation and noise. The results highlighted azimuth estimation accuracies of up to approximately 1.0° using the pseudo-spherical robot head, spherical Eigenmike and planar DICIT array. For the hearing aids, recorded separately but in the same environment, the average azimuth error was 8.5°. Interference from multiple static loudspeakers in Task 2 leads to only small performance degradations of up to 3° compared to Task 1. Variations in the source-sensor geometries due to the motion of the human talkers (Tasks 3 and 4), or the motion of the arrays and talkers (Tasks 5 and 6) affect predominantly the track continuity, completeness and timeliness.

The evaluation also provides evidence for the intrinsic suitability of a given approach for particular arrays or scenarios. For static scenarios (i.e., Tasks 1 and 2), subspace approaches demonstrated particularly accurate localization using the Eigenmike and the robot head incorporating a large number of microphones. Time delay estimation combined with a particle filter resulted in the highest azimuth estimation accuracy for the planar DICIT array. Tracking filters were shown to reduce FARs and missing detections by exploiting models of the source dynamics. Specifically, the localization for moving human talkers in Tasks 3–6 benefits from the incorporation of tracking in dynamic scenarios, resulting in azimuth accuracies of up to 1.8° using the DICIT array, 3.1° using the robot head, and 7.2° using the hearing aids.

Results for the Eigenmike highlighted that localization using spherical arrays benefits from signal processing in the domain of spherical harmonics. The results also indicated that the number of microphones in an array, to some extent, can be traded off against the array aperture. This conclusion is underpinned by the localization results for the 12-microphone robot head that consistently outperformed the 32-microphone Eigenmike for approaches evaluated for both arrays. Nevertheless, increasing microphone spacings also lead to increasingly severe effects of spatial aliasing. As a consequence, all submissions for the 2.24 m-wide DICIT array used subarrays of at most 32 cm inter-microphone spacings.

Several issues remain open challenges for localization and tracking approaches. Intuitively, localization approaches benefit from accurate knowledge of the onsets and endpoints of speech

to avoid false estimates during periods of speech inactivity. Several approaches therefore incorporated voice activity detection based on power spectral density estimates, zero-crossing rates, or by implicit estimation of the onsets and endpoints of speech from the latent variables estimated within a variational Bayesian tracking approach. For the single-source scenarios, particularly low track latency was achieved by the submission based on implicit estimation of the voice activity periods. However, for the multi-source scenarios, approaches incorporating voice activity detection led to increased track fragmentation rates.

Morover, whereas sufficiently long frames are required to address the non-stationarity of speech, dynamic scenes involving moving sources and/or sensors require sufficiently short frames to accurately capture the spatio-temporal variation of the source positions. Therefore, in dynamic scenes, estimation errors due to the non-stationarity of speech must be traded off against biased DoA estimates due to spatio-temporal variation in the source-sensor geometries when selecting the duration of the microphone signals used for localization. In combination with the adverse effects of reverberation and noise, non-stationary signals in dynamic scenes therefore often lead to erroneous, false, missing, spurious DoA estimates in practice.

To conclude, current research is predominantly focused on static scenarios. Only a small subset of the approaches submitted to the LOCATA challenge address the difficult real-world tasks involving multiple moving sources. The challenge evaluation highlighted that there is significant room for improvement, and hence substantial potential for future research. Except for localizing a single static source in not too hostile scenarios none of the problems is robustly solved to the extent desirable for, e.g., informed spatial filtering with high spatial resolution. Therefore, research on appropriate localization and tracking techniques remains an open challenge and the authors hope that the LOCATA dataset and evaluation tools will be found useful to also evaluate future progress.

Inevitably, there are substantial practical limitations in setting up data challenges. In the case of LOCATA, it has resulted in the use of only one acoustic environment because of the need for spatial localization of the ground-truth. Future challenges may beneficially explore variation in performance across different environments.

ACKNOWLEDGMENT

The authors would like to thank all participants of the LOCATA challenge for their submissions and feedback; Claas-Norman Ritter for his contributions to the recordings of the LOCATA corpus; Prof. Verena Hafner for providing access to the facilities at Humboldt-Universität zu Berlin; the anonymous reviewers for their positive and helpfully detailed comments that led to significant improvements of this manuscript; and the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and the IEEE SPS Challenges and Data Collections subcommittee for the support of the LOCATA challenge.

REFERENCES

[1] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2004.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer, 2008.

[4] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[5] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Dec. 2009, pp. 253–256.

[6] K. Reindl, S. Meier, H. Barfuß, and W. Kellermann, "Minimum mutual information-based linearly constrained broadband signal extraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 6, pp. 1096–1108, Jun. 2014.

[7] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 320–332, Feb. 2017.

[8] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[9] C. Evers and P. A. Naylor, "Optimized self-localization for SLAM in dynamic scenes using probability hypothesis density filters," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 863–878, Feb. 2018.

[10] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

[11] K. Harada, E. Yoshida, and K. Yokoi, *Motion Planning for Humanoid Robots*. Berlin, Germany: Springer, 2014.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 64, no. 4, pp. 943–950, Apr. 1979.

[13] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[14] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Prague, Czech Republic, May 2011, pp. 129–132.

[15] H. W. Löllmann *et al.*, *IEEE-AASP Challenge on Source Localization and Tracking: Data Corpus*, Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3630470

[16] C. Evers *et al.*, *IEEE-AASP Challenge Source Localization and Tracking: MATLAB Evaluation Framework*, Jan. 2020. [Online]. Available: https://github.com/cevers/sap_locata_eval

[17] H. W. Löllmann *et al.*, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, Sheffield, U.K., Jul. 2018, pp. 410–414.

[18] H. W. Löllmann *et al.*, *IEEE-AASP Challenge Source Localization and Tracking: Documentation for Participants*. Apr. 2018. [Online]. Available: www.locata-challenge.org.

[19] C. Evers *et al.*, "LOCATA challenge - Evaluation tasks and measures," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Tokyo, Japan, Sep. 2018, pp. 565–569.

[20] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Antibes, France, Sep. 2014, pp. 40–44.

[21] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*, S. Bengio and H. Bourlard, Eds. Berlin, Germany: Springer, 2005, pp. 182–195.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Scottsdale, AZ, USA, Dec. 2015, pp. 504–511.

[23] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[24] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, Jan. 2016. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-016-0306-6#citeas

[25] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. IEEE Work. Autom. Speech Recognit. Understanding*, Dec. 2015, pp. 275–282.

[26] X. A. Pineda *et al.*, "RAVEL: An annotated corpus for training robots with audiovisual abilities," *J. Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 79–91, 2013.

[27] A. Deleforge and R. Horaud, "Learning the direction of a sound source using head motions and spectral features," INRIA, Research Report RR-7529, Feb. 2011. [Online]. Available: https://hal.inria.fr/inria-00564708

[28] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen and J. Garofolo, Eds. Berlin, Germany: Springer, 2007, pp. 1–44.

[29] M. Krindis *et al.*, "An audio-visual database for evaluating person tracking algorithms," in *Proc. IEEE Intl. Conf. Acoust., Speech Signal Process.*, Philadelphia, PA, USA, 2005, pp. ii/237-ii/240.

[30] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Madrid, Spain, Oct. 2018, pp. 5735–5742.

[31] R. N. Jazar, *Theory of Applied Robotics: Kinematics, Dynamics, and Control*, 2nd ed. Berlin, Germany: Springer, 2010.

[32] mh acoustics, *EM32 Eigenmike microphone array release notes (v17.0)*, Oct. 2013. [Online]. Available: www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf

[33] Q. V. Nguyen, F. Colas, E. Vincent, and F. Charpillet, "Motion planning for robot audition," *Auton. Robots*, vol. 43, no. 8, pp. 2293–2317, Dec. 2019.

[34] OptiTrack, *Product Information about OptiTrack Flex13*, Feb. 2018. [Online]. Available: http://optitrack.com/products/flex-13/, http://optitrack.com/products/flex-13/

[35] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1803–1814, Dec. 2014.

[36] V. Tourbabin and B. Rafaely, "Optimal design of microphone array for humanoid-robot audition," in *Proc. Israeli Conf. Robot.*, Herzliya, Israel, Mar. 2016, (abstract).

[37] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "WOZ acoustic data collection for interactive TV," *Lang. Resour. Eval.*, vol. 44, no. 3, pp. 205–219, Sep. 2010.

[38] C. Veaux, J. Yamagishi, and K. MacDonald, "English multi-speaker corpus for CSTR voice cloning toolkit," Jan. 2017. [Online]. Available: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[39] S. Ağcaer and R. Martin, "Binaural source localization based on modulation-domain features and decision pooling," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

[40] 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*, Tokyo, Japan, Sep. 2018.

[41] X. Qian, A. Cavallaro, A. Brutti, and M. Omologo, "LOCATA challenge: Speaker localization with a planar array," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

[42] X. Li, Y. Ban, L. Girin, X. A. Pineda, and R. Horaud, "A cascaded multiple-speaker localization and tracking system," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

[43] 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*, Tokyo, Japan, Sep. 2018.

[44] D. Salvati, C. Drioli, and G. L. Foresti, "Localization and tracking of an acoustic source using a diagonal unloading beamforming and a Kalman filter," in *Proc. LOCATA Challenge Workshop - Satellite Event IWAENC*, Tokyo, Japan, Sep. 2018.

[45] L. D. Mosgaard, D. P. Garcia, T. B. Elmedyb, M. J. Pihl, and P. Mowlaee, "Circular statistics-based low complexity DOA estimation for hearing aid application," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

[46] J. Pak and J. W. Shin, "LOCATA challenge: A deep neural networks-based regression approach for direction-of-arrival estimation," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

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

[48] L. Madmoni, H. B. 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*, Tokyo, Japan, Sep. 2018.

[49] K. Nakadai, K. Itoyama, K. Hoshiba, and H. G. Okuno, "MUSIC-based sound source localization and tracking for tasks 1 and 3," in *Proc. LOCATA Challenge Workshop - Satell. Event IWAENC*, Tokyo, Japan, Sep. 2018.

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

[51] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[52] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. IEEE Intl. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, Mar. 2017.

[53] C. Evers, E. A. P. Habets, S. Gannot, and P. A. Naylor, "DoA reliability for distributed acoustic tracking," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1320–1324, Sep. 2018.

[54] A. Brendel and W. Kellermann, "Learning-based acoustic source-microphone distance estimation using the coherent-to-diffuse power ratio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 61–65.

[55] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 87–112, 2015.

[56] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, 2017.

[57] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," in *Proc. Wireless Acoust. Sensor Netw. Appl.*, May 2017.

[58] M. Souden, J. Benesty, and S. Affes, "Broadband source localization from an eigenanalysis perspective," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1575–1587, Aug. 2010.

[59] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[60] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, 2005.

[61] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[62] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2171–2186, Nov. 2016.

[63] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2007, pp. 18–21.

[64] B. Berdugo, M. A. Doron, J. Rosenhouse, and H. Azhari, "On direction finding of an emitting source from time delays," *J. Acoust. Soc. Amer.*, vol. 105, no. 6, pp. 3355–3363, 1999.

[65] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[66] H. Cao, Y. T. Chan, and H. C. So, "Maximum likelihood TDOA estimation from compressed sensing samples without reconstruction," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 564–568, May 2017.

[67] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 1976–1990, Nov. 2018.

[68] X. Li, Y. Ban, L. Girin, X. A. Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 88–103, Mar. 2019.

[69] J. Traa and P. Smaragdis, "A wrapped Kalman filter for azimuthal speaker tracking," *IEEE Signal Process. Lett.*, vol. 20, no. 12, Dec. 2013.

[70] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1983.

[71] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Amer.*, vol. 62, no. 1, pp. 157–167, 1977.

[72] F. L. Wightman and D. J. Kistler, "The dominant role of lowfrequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, 1992.

[73] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[74] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[75] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 611–623, Mar. 2017.

[76] M. Farmani, M. S. Pedersen, Z. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1275–1289, Jul. 2018.

[77] E. L. Benaroya, N. Obin, M. Liuni, A. Roebel, W. Raumel, and S. Argentieri, "Binaural Localization of multiple sound sources by non-negative tensor factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1072–1082, Jun. 2018.

[78] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Intl. Conf. Mach. Learning*, New York, NY, USA, 2005, pp. 792–799.

[79] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J. Acoust. Soc. Amer.*, vol. 112, no. 4, pp. 1583–1596, 2002.

[80] T. V. den Bogaert, E. Carette, and J. Wouters, "Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna," *Int. J. Audiology*, vol. 50, no. 3, pp. 164–176, 2011.

[81] H. Wallach, "The role of head movement and vestibular and visual cues in sound localization," *J. Exp. Psychol.*, vol. 27, pp. 339–368, 1940.

[82] J. Burger, "Front-back discrimination of the hearing systems," *Acta Acustica United Acustica*, vol. 8, no. 5, pp. 301–302, 1958.

[83] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," *J. Acoust. Soc. Amer.*, vol. 42, no. 2, pp. 489–493, 1967.

[84] F. L. Wightman and D. J. Kistler, "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2841–2853, 1999.

[85] S. Perrett and W. Noble, "The contribution of head motion cues to localization of low-pass noise," *Perception Psychophys.*, vol. 59, no. 7, pp. 1018–1026, Jan. 1997.

[86] D. M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *J. Acoust. Soc. Amer.*, vol. 31, no. 7, pp. 977–986, 1959.

[87] P. A. Hill, P. A. Nelson, O. Kirkeby, and H. Hamada, "Resolution of front–back confusion in virtual acoustic imaging systems," *J. Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 2901–2910, 2000.

[88] U.-H. Kim, K. Nakadai, and H. G. Okuno, "Improved sound source localization and front-back disambiguation for humanoid robots with two ears," in *Recent Trends in Applied Artificial Intelligence*, M. Ali, T. Bosse, K. V. Hindriks, M. Hoogendoorn, C. M. Jonker, and J. Treur, Eds. Berlin, Germany: Springer, 2013, pp. 282–291.

[89] W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," *Signal Process.*, pp. 577–590, 1973.

[90] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inf. Theory*, vol. 19, no. 5, pp. 608–614, Sep. 1973.

[91] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1210–1217, Oct. 1983.

[92] M. Taseska and E. A. P. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1291–1304, Jul. 2016.

[93] M. Omologo, P. G. Svaizer, and R. D. Mori, *Acoustic Transduction*. San Diego, CA, USA: Academic, 1998, pp. 23–69.

[94] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. Acoust., Speech, Signal Processs.*, vol. 13, no. 4, pp. 593–606, Jul. 2005.

[95] D. Salvati, C. Drioli, and G. L. Foresti, "A low-complexity robust beamforming using diagonal unloading for acoustic source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 609–622, Mar. 2018.

[96] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer Topics in Signal Processing. Berlin, Germany: Springer, 2015.

[97] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[98] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, 2004.

[99] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3351–3361, Jul. 2016.

[100] D. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Berlin, Germany: Springer, 2016.

[101] 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, Lang. Process.*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[102] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. Int. Symp. Signal Process. Appl.*, Jul. 2003, vol. 2, pp. 411–414.

[103] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[104] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[105] R. J. Weiss, M. I. Mandel, and D. P. Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Commun.*, vol. 53, no. 5, pp. 606–621, 2011.

[106] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[107] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[108] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Intl. Conf. Acoust., Speech Signal Process.*, Mar. 2005.

[109] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas NV, USA, Mar. 2008, pp. 5276–5279.

[110] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *J. Soc. Ind. Appl. Math. Ser. B Numerical Anal.*, vol. 2, no. 2, pp. 205–224, 1965.

[111] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2002.

[112] 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, Lang. Process.*, vol. 22, no. 10, Oct. 2014, pp. 1494–1505.

[113] B. Ristic, *Particle Filters for Random Set Models*, 1st ed. New York, NY, USA: Springer, 2013.

[114] R. P. S. Mahler, *Statistical Multisource Multitarget Information Fusion*. Norwood, MA, USA: Artech House, 2007.

[115] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, Santander, Spain, Sep. 2012, pp. 1–6.

[116] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016.

[117] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 2814–2818.

[118] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2386–2390.

[119] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 405–409.

[120] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2217–2221.

[121] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[122] M. R. Bai, S. Lan, and J. Huang, "Time difference of arrival (TDOA)-based acoustic source localization and signal extraction for intelligent audio classification," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, Jul. 2018, pp. 632–636.

[123] F. Grondin, F. Glass, I. Sobieraj, and M. D. Plumbley, "Sound event localization and detection using CRNN on pairs of microphones," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, New York, NY, USA, Oct. 2019.

[124] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2122–2131, Nov. 2018.

[125] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," in *Proc. ML4Audio Workshop NIPS*, 2017.

[126] D. Hinrichsen and A. J. Pritchard, *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*, Texts in Applied Mathematics. Berlin, Germany: Springer, 2005.

[127] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME–J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[128] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Boston, MA, USA: Artech House, 2004.

[129] 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 Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[130] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2006, Jun. 2006, Art. no. 017021.

[131] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Information Science and Statistics. Berlin, Germany: Springer, 2001.

[132] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.

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

[134] P. Li, R. Goodall, and V. Kadirkamanathan, "Estimation of parameters in a linear state space model using a Rao-blackwellised particle filter," *IEE Proc. - Control Theory Appl.*, vol. 151, no. 6, pp. 727–738, Nov. 2004.

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

[136] A. Levy, S. Gannot, and E. A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1540–1555, Aug. 2011.

[137] T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: Classification, implementation, and strategies," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 70–86, May 2015.

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

[139] M. Bolic and P. M. D. and, "New resampling algorithms for particle filters," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2003, vol. 2, pp. II–589.

[140] M. Halimeh, C. Hümmer, A. Brendel, and W. Kellermann, "Hybrid particle filtering based on an elitist resampling scheme," in *Proc. Sensor Array Multichannel Signal Process. Workshop*, Jul. 2018, pp. 257–261.

[141] M. Halimeh, A. Brendel, and W. Kellermann, "Evolutionary resampling for multi-target tracking using probability hypothesis density filter," in *Proc. Eur. Signal Process. Conf.*, Sep. 2018, pp. 647–651.

[142] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 14–29, Jul. 2016.

[143] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 651–661, Mar. 2017.

[144] D. J. Salmond, "Mixture reduction algorithms for point and extended object tracking in clutter," *IEEE Trans. Aerosp. and Electron. Syst.*, vol. 45, no. 2, pp. 667–686, Apr. 2009.

[145] K. V. Mardia and P. E. Jupp, *Directional Statistics*, vol. 494. Hoboken, NJ, USA: Wiley, 2009.

[146] K. V. Mardia, "Bayesian analysis for bivariate Von Mises distributions," *J. Appl. Statist.*, vol. 37, no. 3, pp. 515–528, 2010.

[147] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. Eur. Signal Process. Conf.*, Aug. 2011, pp. 1347–1351.

[148] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

[149] S. S. Blackman, "Association and fusion of multiple sensor data," in *Multitarget-Multisensor Tracking: Advanced Algorithms*, Y. Bar-Shalom, Ed. Norwood, MA, USA: Artech House, 1990, ch. 6, pp. 187–218.

[150] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[151] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[152] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451–460, 1975.

[153] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Syst. Mag.*, vol. 29, no. 6, pp. 82–100, Dec. 2009.

[154] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, Jul. 1983.

[155] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen and J. Garofolo, Eds. Berlin, Germany: Springer, 2007, pp. 137–150.

[156] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking multiple audio sources with the von mises distribution and variational em," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 798–802, Jun. 2019.

[157] R. P. S. Mahler, *Adv. in Statistical Multisource-Multitarget Information Fusion* . Artech House, 2014.

[158] R. P. S. Mahler, "Statistics 101 for multisensor, multitarget data fusion," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1, Jan. 2004.

[159] R. P. S. Mahler, ""Statistics 102" for multisource-multitarget detection and tracking," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 3, pp. 376–389, Jun. 2013.

[160] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[161] H. Pessentheiner, "Localization, characterization, and tracking of harmonic sources with applications to speech signal processing," Ph.D. dissertation, Graz Univ. Technol., Graz, Austria, Jan. 2017.

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

[163] Y. Liu, W. Wang, J. Chambers, V. Kilic, and A. Hilton, "Particle flow SMC-PHD filter for audio-visual multi-speaker tracking," in *Latent Variable Analysis and Signal Separation*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds. Cham, Switzerland: Springer, 2017, pp. 344–353.

[164] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA, USA: Artech House, 1999.

[165] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, pp. 83–97, Mar. 1955.

[166] R. L. Rothrock and O. E. Drummond, "Performance metrics for multiple-sensor multiple-target tracking," *Proc. SPIE*, vol. 4048, Jul. 2000.

[167] A. A. Gorji, R. Tharmarasa, and T. Kirubarajan, "Performance measures for multiple target tracking problems," in *Proc. Int. Conf. Inf. Fusion*, Chicago (Illinois), USA, Jul. 2011.

[168] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.

[169] A. S. Rahmathullah and A. F. G. Fernández, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *Int. Conf. Inf. Fusion*, Jul. 2017, pp. 1–8.
[170] B. Ristic, B.-N. Vo, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3452–3457, Jul. 2011.
[171] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

Christine Evers (Senior Member, IEEE) received her PhD from the University of Edinburgh in 2010; her MSc degree in Signal Processing and Communications from the University of Edinburgh in 2006; and her BSc degree in Electrical Engineering and Computer Science from Jacobs University, Germany, in 2005. She is a Lecturer in Computer Science at the University of Southampton. She was the recipient of an EPSRC Fellowship, hosted at Imperial College London between 2017–2019. She worked as a Research Associate at Imperial College London between 2014–2016; as a Senior Systems Engineer at Selex Electronic Systems between 2010–2014; and as a Research Fellow at the University of Edinburgh between 2009–2010. Her research interests are on Bayesian learning for machine listening. She is currently member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and serves as an associate editor of the EURASIP Journal on Audio, Speech, and Music Processing.

Heinrich W. Löllmann (Senior Member, IEEE) received the Dipl.-Ing. (univ) degree in electrical engineering in 2001 and the Dr.-Ing. degree in 2011 at RWTH Aachen University. He worked as Scientific Co-Worker at the Institute of Communication Systems and Data Processing of RWTH Aachen University from 2001 to 2012. Since 2012, he is a senior researcher at the Chair of Multimedia Communications and Signal Processing at the Friedrich-Alexander University Erlangen-Nürnberg (FAU). Heinrich Löllmann has authored one book chapter and more than 40 refereed papers. His research focuses on speech and audio signal processing, including filter-bank design, speech dereverberation and noise reduction, estimation of room acoustical parameters, and algorithms for robot audition. He is currently member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing.

Heinrich Mellmann studied Computer Science and Mathematics at the Humboldt-Universität zu Berlin (HUB), and received his degrees Dipl.-Inf. in 2011 and Dipl.-Math. in 2017 respectively. He is currently a Research Assistant at the chair of Adaptive Systems, HUB, pursuing a PhD in the area of cognitive robotics. Over the past years, he was involved in a number of research projects in cognitive robotics. Heinrich Mellmann has been actively involved in the RoboCup initiative since 2005, and has been leading the RoboCup team 'Berlin United' since 2009. His research interests within cognitive robotics focus on spatial perception and decision making in autonomous humanoid robots.

Alexander Schmidt (Member, IEEE) studied Electrical Engineering at Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany, and received his MSc degree in 2015. He is currently a Research Assistant at the Chair of Multimedia Communications and Signal Processing, FAU, pursuing a PhD in the area of multichannel signal enhancement for robot audition. He was with the EU FP7 Project Embodied Audition for Robots (EARS). His special interests lie in the area of (sparse) dictionary learning for signal representation combined with physical-mechanical models.

Hendrik Barfuss received a master degree from Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany, in 2013. He is currently pursuing a Dr.-Ing. degree in Electrical, Electronic and Communication Engineering at the Chair of Multimedia Communications and Signal Processing, FAU. His research interests are in microphone array signal processing, spatial filtering, speech enhancement, and localization.

Patrick A. Naylor (Fellow, IEEE) received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, UK, and the PhD degree from Imperial College London, UK. He is Professor of Speech and Acoustic Signal Processing at Imperial College London. His research interests are in speech, audio and acoustic signal processing. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement. He has also worked on speech dereverberation including blind multichannel system identification and equalization, acoustic echo control, non-intrusive speech quality estimation, and speech production modelling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is currently a member of the Board of Governors of the IEEE SIGNAL PROCESSING SOCIETY AND PRESIDENT OF THE EUROPEAN ASSOCIATION FOR SIGNAL PROCESSING (EURASIP). He was formerly Chair of the IEEE SIGNAL PROCESSING SOCIETY TECHNICAL COMMITTEE ON AUDIO AND ACOUSTIC SIGNAL PROCESSING. He has served as an associate editor of IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor of IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.

Walter Kellermann (Fellow, IEEE) received the Dipl.-Ing. (univ.) degree in Electrical Engineering from FAU, in 1983, and the Dr.-Ing. degree from the Technical University Darmstadt, Germany, in 1988. From 1989 to 1990, he was a postdoctoral Member of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ. He is a Professor for communications at the Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany, since 1999. In 1990, he joined Philips Kommunikations Industrie, Nuremberg, Germany, to work on hands-free communication in cars. From 1993 to 1999, he was a Professor at the Fachhochschule Regensburg, where he also became Director of the Institute of Applied Research in 1997. In 1999, he cofounded DSP Solutions, a consulting firm in digital signal processing, and he joined FAU as a Professor and Head of the Audio Research Laboratory. He authored or coauthored 21 book chapters, 300+ refereed papers in journals and conference proceedings, as well as 70+ patents, and is a co-recipient of ten best paper awards. His current research interests include speech signal processing, array signal processing, adaptive and learning algorithms, and its applications to acoustic humanmachine interfaces. Dr. Kellermann served as an Associate Editor and Guest Editor to various journals, including the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2000 to 2004, the IEEE Signal Processing Magazine in 2015, and presently serves as Associate Editor to the EURASIP Journal on Applied Signal Processing. He was the General Chair of seven mostly IEEE-sponsored workshops and conferences. He served as a Distinguished Lecturer of the IEEE Signal Processing Society (SPS) from 2007 to 2008. He was the Chair of the IEEE SPS Technical Committee for Audio and Acoustic Signal Processing from 2008 to 2010, a Member of the IEEE James L. Flanagan Award Committee from 2011 to 2014, a Member of the SPS Board of Governors (2013-2015), Vice President Technical Directions of the IEEE Signal Processing Society (2016-2018) and is currently a member of the SPS Nominations Appointments Committee (2019-2020). He was awarded the Julius von Haast Fellowship by the Royal Society of New Zealand in 2012 and the Group Technical Achievement Award of the European Association for Signal Processing (EURASIP) in 2015. In 2016, he was a Visiting Fellow at Australian National University, Canberra, Australia. He is an IEEE Fellow.