Blind Source Separation of Moving Sources Using Sparsity-Based Source Detection and Tracking

Maja Taseska, *Student Member, IEEE*, and Emanuël A. P. Habets, *Senior Member, IEEE*

***Abstract*—Sparsity-based blind source separation (BSS) algorithms in the short time–frequency (TF) domain have received a lot of attention due to their versatility and noise reduction capabilities. In most of these algorithms, the estimation of the BSS filters relies on the accurate association of each time–frequency bin to the dominant source at that bin. The TF bin associations are then used to estimate the statistics of the source signals, and BSS is achieved by optimal spatial filters computed using the estimated statistics. The main objective of this paper is to apply such a framework to scenarios with an unknown number of moving sources. While state-of-the-art approaches employ online clustering algorithms to solve the problem for moving sources, we propose an approximate Bayesian tracker and perform the association of each TF bin to the dominant source using the tracker's measurement-to-source association probabilities. Therefore, the choice of the underlying narrowband models and measurements for the tracker as well as the resulting tracking algorithm constitute the main contributions of this paper. The TF bin associations obtained from the tracker are then used to estimate the statistics of the source signals. The performance of the resulting BSS filters is compared to the performance of state-of-the-art sparsity-based and independent vector analysis-based BSS algorithms. Our proposed approach targets scenarios with at least two spatially separated microphone arrays, with known microphone positions and relative orientations. The framework also allows for efficient management of a time-varying number of sources.**

***Index Terms*—Spatial filtering, source separation, acoustic source tracking, PSD matrix estimation.**

I. Introduction

SIGNALS received at the microphones in hands-free systems often contain multiple speech signals and background noise. Whether the signals are fed to a speech recognizer or used for communication, signal enhancement is necessary to separate speakers and reduce noise, often without knowledge of the source locations and signal statistics. In the last two decades, sparsity-based BSS received increasing attention [1], [2] as a versatile approach to joint BSS and noise reduction. Sparsity in the short-time Fourier transform (STFT) domain implies that even in the presence of concurrent speakers, the energy of only one speaker is dominant at a given time-frequency (TF) bin [2]. Hence, each TF bin can be associated to the dominant source, resulting in so-called TF masks. The TF mask for a given source contains values between 0 and 1, where the value at a given TF bin approaches 1 if the source is dominant at that bin, and zero otherwise.

Originally, to separate the sources, estimated TF masks were applied as gains to one of the microphone signals [2]. In the last decade however, with the availability of microphone arrays with a larger number of microphones, several researchers proposed to use the TF masks only as means to estimate the power spectral density (PSD) matrices of the signals and to compute spatial filters for BSS [3]–[6]. We refer to such TF mask-informed spatial filters as informed spatial filters (ISFs). In addition to sparsity, ISFs exploit the spatial diversity of the microphone array and provide good signal quality even in reverberant and multi-talk scenarios. It was recently shown in [7] that multiplying one of the microphone signals by an estimated TF mask, leads to a lesser speech recognition improvement compared to a minimum variance distortionless response (MVDR) filter that uses the same TF mask for PSD estimation [7]. The objective of this paper is to estimate the TF masks and use informed MVDR filters for BSS of an unknown number of moving sources.

TF masks can be estimated by clustering of features extracted at each TF bin (referred to as *narrowband* features) in J classes, where J is the number of sources. Common features include interaural phase and level differences [2], STFT-domain signal vectors [3], [4], [6], [8], narrowband direction of arrivals (DOAs) [9], [10], or in multi-array systems, inter-array attenuation [11], phase ratios of microphone pairs [12], and narrowband positions [5]. An important practical criterion when choosing features and models is the ability of the resulting system to handle moving sources. In systems based on location features such as DOAs and positions [9], [10], [13], effective heuristics have been devised to manage a time-varying number of moving sources. However, the algorithms based on online clustering in [9], [10], and [13] are generally sub-optimal. To estimate the TF masks via optimal Bayesian tracking of sources, the authors in [14] use a wrapped Kalman filter. Although the authors in [14] suggest probabilistic data association (PDA) for the measurement-to-source association, it should be noted that the single measurement per source model of PDA [15] is not valid for narrowband features (we clarify this point in Section IV). Time-difference-of-arrival (TDOA) has also been

Manuscript received June 16, 2017; revised September 30, 2017 and November 22, 2017; accepted November 28, 2017. Date of publication December 7, 2017; date of current version January 25, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hiroshi Saruwatari. *(Corresponding author: Maja Taseska.)*

The authors are with the International Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), Erlangen 91058, Germany (e-mail: maja.taseska@audiolabs-erlangen.de; emanuel.habets@audiolabs-erlangen.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2780993

2329-9290 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

used for Bayesian tracking [16]. As it represents a fullband feature (i.e., at each time frame a single TDOA is estimated), it is not applicable for TF mask estimation.

In this paper, to estimate the TF masks of moving sources, we propose an approximate Bayesian tracker that is explicitly based on a model where multiple measurements per source are possible in a given time frame. In contrast, in existing Bayesian multi-source trackers, the underlying assumption is that each source can generate at most one measurement per frame [15]–[17]. However, in the TF domain, it is clear that one source can be dominant at multiple frequencies, and hence, can generate multiple measurements per frame. This property of the model has notable implications for the development of a suitable tracking algorithm, and standard multi-source trackers such as the joint probabilistic data association (JPDA) filter [15] are no longer applicable. The development of an approximate Bayesian tracker suitable for this type of models and its application for TF mask estimation represents the major contribution of this work. In the proposed tracker, we use narrowband position estimates as input measurements, which have been used in our previous work for clustering-based BSS [5], [13]. The overall measurement model includes the narrowband positions and the STFT domain signal vectors. Such an augmented model that includes the signal vectors allows us to quantify the speech presence uncertainty at each TF bin using common statistical models [18], and to estimate the background noise statistics. We show that the associations of each TF bin to the dominant source obtained at the tracker output provide accurate TF masks which are used to estimate spatial filters and separate moving sources. It should be noted that the proposed approach does not suffer from the frequency permutation problem commonly found in convolutive BSS algorithms, while the Markovian property of the speaker motion model employed in the tracker ensures that the source association is consistent across time frames.

As the narrowband position measurements for the tracker are obtained by triangulation of multiple DOA estimates, the proposed system requires at least two spatially separated arrays with known orientations and microphone locations. As we do not target ad-hoc scenarios, we assume that all signals are synchronized at a central processor. Synchronization is addressed in [19], while microphone localization is addressed in [20] and references therein. Although multiple arrays are required to obtain the position estimates, the ISF filters for BSS can be computed using only a subset of microphones. For instance, computing a spatial filter using only the nearest microphone array for each source, can lead to better separated signal quality in practice due to the higher signal-to-noise ratio at the nearest microphones. Given the estimated source locations by the tracker, the issue of microphone subset selection [21], [22] in our system is solved by choosing the array that is nearest to the estimated source location.

The rest of the paper is organized follows: in Section II, we give an overview of ISF-based BSS. In Section III, we formulate the tracking problem and relate the measurement-to-source association to TF mask estimation. In Section IV, we derive the multi-source tracker which provides the measurement-to-source associations and the source position estimates. Furthermore, we discuss the relation between the proposed tracker and other trackers from the literature. A method for track management (detecting appearing and disappearing sources) is discussed in Section V. In Section VI we provide a comprehensive performance evaluation and comparison of the proposed BSS approach to state-of-the-art BSS approaches.

II. Source Separation by Informed Spatial Filters

A. Signal Model

Throughout the paper, we assume that time-domain signals are windowed and transformed to the STFT domain. If the number of microphones is M and the total number of speakers at time τ is J_τ, the microphone signal $\boldsymbol{y} \in \mathbb{C}^{M\times 1}$ at time τ and frequency bin k is given by

$$\boldsymbol{y}(\tau,k) = \sum_{j=1}^{J_\tau} \boldsymbol{s}_j(\tau,k) + \boldsymbol{v}(\tau,k), \tag{1}$$

where $\boldsymbol{s}_j$ is the j-th speaker vector and $\boldsymbol{v}$ contains background and sensor noise. Denoting by $A_{jm}(\tau,k)$ the non-zero acoustic transfer function between the j-th source position at time τ and the m-th microphone, the relative transfer function (RTF) vector with respect to the m-th microphone is defined as (we omit the indices τ and k for brevity)

$$\boldsymbol{g}_{jm} = \left[\frac{A_{j1}}{A_{jm}}, \ldots, \frac{A_{j(m-1)}}{A_{jm}},\ 1,\ \frac{A_{j(m+1)}}{A_{jm}} \ldots \frac{A_{jM}}{A_{jm}}\right]. \tag{2}$$

The signal $\boldsymbol{s}_j$ of the j-th source is related to the signal S_{jm} at the m-th microphone via the RTF vector as follows

$$\boldsymbol{s}_j(\tau,k) = \boldsymbol{g}_{jm}(\tau,k)\, S_{jm}(\tau,k), \quad \forall m \in [1, M]. \tag{3}$$

The PSD matrix of the microphone signals is defined as $\boldsymbol{\Phi}_{\boldsymbol{y}}(\tau,k) = \mathrm{E}\left[\boldsymbol{y}(\tau,k)\boldsymbol{y}^{\mathrm{H}}(\tau,k)\right]$, where $\mathrm{E}[\cdot]$ denotes statistical expectation, and $\boldsymbol{\Phi}_{\boldsymbol{s}_j}$ and $\boldsymbol{\Phi}_{\boldsymbol{v}}$ are defined similarly. As the signals are assumed mutually uncorrelated, the PSD matrices satisfy $\boldsymbol{\Phi}_{\boldsymbol{y}} = \sum_{j=1}^{J} \boldsymbol{\Phi}_{\boldsymbol{s}_j} + \boldsymbol{\Phi}_{\boldsymbol{v}}$. Furthermore, for each source j, the corresponding PSD matrix $\boldsymbol{\Phi}_{\boldsymbol{s}_j}$ is given by

$$\boldsymbol{\Phi}_{\boldsymbol{s}_j}(\tau,k) = \phi_{jm}(\tau,k)\, \boldsymbol{g}_{jm}(\tau,k)\, \boldsymbol{g}_{jm}^{\mathrm{H}}(\tau,k), \tag{4}$$

where $\phi_{jm}(\tau,k) = \mathrm{E}\left[|S_{jm}(\tau,k)|^2\right]$. Although due to source movements, the RTF vectors $\boldsymbol{g}_{jm}(\tau,k)$ are time-dependent, it is assumed that each $\boldsymbol{g}_{jm}$ varies slowly compared to the speech PSD ϕ_{jm} (or in other words, that the spectral properties of the speech signal vary faster than their spatial properties).

The objective in this paper is to compute a separation filter $\boldsymbol{h}_{jm}(\tau,k)$ for each source j, and to obtain an estimate of S_{jm} as follows (the reference microphone m can be different for different sources)

$$\widehat{S}_{jm}(\tau,k) = \boldsymbol{h}_{jm}^{\mathrm{H}}(\tau,k)\, \boldsymbol{y}(\tau,k). \tag{5}$$

Finally, the short-time spectra $\widehat{S}_{jm}$ are transformed by an inverse STFT, and the time-domain separated signals are reconstructed by the overlap-add method [23]. Note that if the possibly time-varying number of sources J_τ is not known in advance, it needs to be estimated from the data. In the framework developed in the rest of the paper, we also propose a way to estimate the time-varying number of sources.

B. Source Separation Using Informed MVDR Filters

Using the RTF vectors defined in (2), a minimum variance distortionless response (MVDR) filter to extract the j-th source at the m-th microphone is obtained by solving the following optimization problem for each TF bin (τ, k) [24]

$$\boldsymbol{h}_{jm} = \arg\min_{\boldsymbol{h}} \boldsymbol{h}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{s}_j} \boldsymbol{h}, \quad \text{subject to } \boldsymbol{h}^{\mathrm{H}} \boldsymbol{g}_{jm} = 1, \tag{6}$$

where $\boldsymbol{\Phi}_{\tilde{s}_j} = \sum_{j' \neq j} \boldsymbol{\Phi}_{s_{j'}} + \boldsymbol{\Phi}_{\boldsymbol{v}}$ is the undesired signal PSD matrix with respect to source j. The solution to (6) is given by [24], [25]

$$\boldsymbol{h}_{jm} = (\boldsymbol{g}_{jm}^{\mathrm{H}} \boldsymbol{\Phi}_{\tilde{s}_j}^{-1} \boldsymbol{g}_{jm})^{-1} \boldsymbol{\Phi}_{\tilde{s}_j}^{-1} \boldsymbol{g}_{jm}. \tag{7}$$

In practice, the PSD matrices $\boldsymbol{\Phi}_{s_j}(\tau, k)$ for each j are estimated from the microphone signals, by first estimating $\boldsymbol{\Phi}_{\boldsymbol{v}}$ and $\boldsymbol{\Phi}_{s_j+\boldsymbol{v}} = \boldsymbol{\Phi}_{s_j} + \boldsymbol{\Phi}_{\boldsymbol{v}}$ by recursive temporal averaging as

$$\begin{aligned} \widehat{\boldsymbol{\Phi}}_{s_j+\boldsymbol{v}} &= \alpha_j(\tau)\, \widehat{\boldsymbol{\Phi}}_{s_j+\boldsymbol{v}}(\tau-1) + [1-\alpha_j(\tau)]\, \boldsymbol{y}(\tau)\boldsymbol{y}^{\mathrm{H}}(\tau) \\ \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}} &= \alpha_0(\tau)\, \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(\tau-1) + [1-\alpha_0(\tau)]\, \boldsymbol{y}(\tau)\boldsymbol{y}^{\mathrm{H}}(\tau), \end{aligned} \tag{8}$$

and setting $\widehat{\boldsymbol{\Phi}}_{s_j} = \widehat{\boldsymbol{\Phi}}_{s_j+\boldsymbol{v}} - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}$. To highlight the main concept of informed MVDR filtering for source separation, it is crucial to note that the recursions in (8) for the different PSD matrices differ only in the averaging constant α_j. The ability to estimate different PSD matrices by only changing the averaging constant at each TF bin is due to the underlying sparsity assumption of speech signals in the STFT domain, i.e., that at each TF bin, the energy from only one source is dominant [2]. The averaging parameter α_j should ensure that $\widehat{\boldsymbol{\Phi}}_{s_j}$ is only updated if source j is dominant at bin (τ, k). The sparse signal model and the dominant source detection are formalized in Section II-C, while the rest of the paper develops the approximate Bayesian tracker required to perform dominant source detection in our framework.

Using the rank-one model, the RTF vector $\boldsymbol{g}_{jm}$ required for the MVDR filter is obtained as the m-th column of $\widehat{\boldsymbol{\Phi}}_{s_j}$ divided by its m-th entry [26]. Although the choice of RTF estimator often affects the signal quality [26], such investigation is beyond the scope of this paper. Finally, to justify the rank-one assumption for the recursively averaged PSD matrix estimate $\widehat{\boldsymbol{\Phi}}_{s_j}$ in moving source scenarios, consider a typical frame length and a source velocity of 64 ms and 0.5–1 m/s, respectively. As the source travels only 3.2–6.4 cm within a frame, the RTF vectors of neighboring frames are highly aligned. Therefore, the PSD matrix estimated by recursive temporal averaging with a sufficiently large time constant tends to have only one dominant eigenvalue and can be approximated by a rank-one matrix.

C. Sparse Model and Dominant Source Detection

To formalize the dominant source detection, it is common to define a discrete random variable (RV) $Z_{\tau k}$ with support $[0, J_\tau]$ and let $z_{\tau k}$ denote the realization of $Z_{\tau k}$, indicating the dominant source at bin (τ, k), such that

$$\begin{aligned} z_{\tau k} &= j, \quad \text{for } j > 0 \quad \text{if the } j\text{-th source is dominant,} \\ z_{\tau k} &= 0, \quad \text{if noise is dominant.} \end{aligned} \tag{9}$$

Denoting the posterior probability of $Z_{\tau k}$ by $p(Z_{\tau k} \mid \mathcal{Y}_{1:\tau})$, a maximum a-posteriori (MAP) estimate of the dominant source index $z_{\tau k}$ at TF bin (τ, k) is given by

$$\hat{z}_{\tau k} = \arg\max_{z}\ p(Z_{\tau k} = z \mid \mathcal{Y}_{1:\tau}), \tag{10}$$

where $\mathcal{Y}_{1:\tau}$ contains all signals up to τ. Note that the presented model assumes that at a given TF bin, the same source is dominant at all arrays. Although this assumption might be violated in practice, we explain in Section III-B that such a violation is not critical.

Using the estimated dominant source index $\hat{z}_{\tau k}$, the following binary TF masks can be defined for each source

$$\mathcal{M}_j(\tau, k) = \begin{cases} 1, & \text{if } z_{\tau k} = j \\ 0, & \text{if } z_{\tau k} \neq j. \end{cases} \tag{11}$$

Given the TF masks, the relation between the dominant source index and the averaging parameters α_j for the PSD matrices in (8) can be explicitly stated as follows

$$\alpha_j(\tau, k) = 1 - (1 - \tilde{\alpha}_s)\, \mathcal{M}_j(\tau, k), \tag{12}$$

where $\tilde{\alpha}_s \in (0, 1)$ is a constant, that determines the value of α_j when source j is dominant.

Clearly, accurate estimation of the dominant source index $z_{\tau k}$ is required to obtain accurate PSD matrix estimates. Considering moving source scenarios, and motivated by the optimal properties of Bayesian trackers, we approach the problem from a multi-source tracking point of view, where the estimation of $z_{\tau k}$ is known as *data association* or *measurement-to-source association*. In Section III, we formulate the tracking problem that provides the required dominant source indices and in Section IV, we develop a tracking algorithm for this problem.

III. Formulation of the Tracking Problem

Sections III-A and III-B follow the standard presentation of Bayesian trackers, where we first define the state and the measurement models in Section III-A, and describe the extraction of the measurements from the microphone signals in Section III-B. In Section III-C, we propose an augmented measurement model, with the objective to provide a unified treatment of speech presence uncertainty in the measurements. Note that the benefits of using augmented features for detection of noisy measurements in multi-source trackers was discussed in [27]. Nonetheless, the particular choice of augmented measurements and models for the application to speech source tracking represents one of the contributions of this paper. Finally, the objective of choosing appropriate measurement models is to be able to parametrize and evaluate the posterior probability of the dominant source index, as detailed in Section III-D.

Notation: Probability distributions of discrete RVs are denoted by $p(\cdot)$, whereas probability densities of continuous RVs, or mixed joint densities of a continuous and a discrete RV are denoted by $f(\cdot)$. The time and frequency indices τ and k of the state-space variables are denoted in the subscript.

A. State and Measurement Models

Let the vector $\boldsymbol{x}_{\tau j}$ denote the position (state) of the j-th source at time τ, in an arbitrary 2-dimensional (2D) Cartesian coordinate system. The source movement is modeled as a Gaussian

random walk with a covariance matrix $\boldsymbol{Q}_j$, i.e.,

$$f\left(\boldsymbol{x}_{(\tau+1)j}\right) = \mathcal{N}\left(\boldsymbol{x}_{(\tau+1)j}\,;\, \boldsymbol{x}_{\tau j},\, \boldsymbol{Q}_j\right),\ \forall j \in [1, J_\tau], \tag{13}$$

where $\boldsymbol{Q}_j$ is a diagonal matrix with equal entries on the diagonal. The value on the diagonal relates to the source speed and the length of the STFT frame. To obtain the measurement-to-source associations at each frame τ, we need to estimate the number of sources J_τ and their states $\mathcal{X}_\tau = \{\boldsymbol{x}_{\tau 1}, \boldsymbol{x}_{\tau 2}, \ldots \boldsymbol{x}_{\tau J_\tau}\}$, by using measurements extracted from $\boldsymbol{y}(\tau, k)$. Commonly used measurements for tracking, such as DOAs and TDOAs, are non-linearly related to the states. In this work, we propose a narrowband position measurement $\boldsymbol{r}_{\tau k}$, which represents an estimate of the dominant source position at TF bin (τ, k) obtained by several non-linear processing steps from the microphone signal vector $\boldsymbol{y}(\tau, k)$, detailed in Section III-B. Not to confuse the measurements with the states $\boldsymbol{x}_{\tau j}$, the measurements are denoted by $\boldsymbol{r}_{\tau k}$, although they are both RVs in the state-space. Due to noise and reverberation, the measurements are assumed to be corrupted by Gaussian noise. Considering that the noise and reverberation are rather diffuse, the measurements are assumed uniformly distributed when speech is absent. The measurement model is written as

$$f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k}) = \begin{cases} \mathcal{N}\left(\boldsymbol{r}_{\tau k}\,;\, \boldsymbol{x}_{\tau z_{\tau k}}, \boldsymbol{\Sigma}_{\tau z_{\tau k}}\right), & \text{if } z_{\tau k} > 0 \\ \mathcal{U}\left(\boldsymbol{r}_{\tau k}\right), & \text{if } z_{\tau k} = 0, \end{cases} \tag{14}$$

where $\mathcal{N}(\boldsymbol{r}; \boldsymbol{x}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{x}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathcal{U}(\boldsymbol{r})$ is a uniform distribution on an x-y slice of the room. Note that (14) corresponds to a standard Gaussian model with an additional clutter measurement model, which is also used in JPDA [15]. In contrast to common tracking systems, the noise covariance $\boldsymbol{\Sigma}_{\tau z_{\tau k}}$ in our framework is source- and time-dependent, and needs to be estimated online from the data.

B. Extraction of Position Measurements

An efficient method to obtain the 2×1 narrowband position measurement vectors $\boldsymbol{r}_{\tau k}$ from the $M \times 1$ complex-valued signals $\boldsymbol{y}(\tau, k)$ is by the following non-linear processing steps: splitting $\boldsymbol{y}(\tau, k)$ in multiple vectors corresponding to the different sub-arrays, estimating DOAs at each array, and triangulating DOA vectors. For 2D spatial processing as done in this work, DOAs at two arrays suffice to obtain $\boldsymbol{r}_{\tau k}$, for instance, by choosing at each TF bin the two arrays with the largest signal amplitude at that bin. In this manner, the assumption that for a given TF bin the same source is dominant at all arrays, mentioned in Section II-C, can be relaxed to only require that the same source is dominant at the two arrays with largest signal amplitude. Let $\boldsymbol{d}_1, \boldsymbol{d}_2$ denote the locations of the two arrays chosen for triangulation at a TF bin (τ, k), $\boldsymbol{e}_{\boldsymbol{d}_1,\tau k} = [\cos\theta_{\boldsymbol{d}_1,\tau k},\ \sin\theta_{\boldsymbol{d}_1,\tau k}]$ and $\boldsymbol{e}_{\boldsymbol{d}_2,\tau k} = [\cos\theta_{\boldsymbol{d}_2,\tau k},\ \sin\theta_{\boldsymbol{d}_2,\tau k}]$ the corresponding DOA vectors, and $\theta_{\boldsymbol{d}_1,\tau k}$ and $\theta_{\boldsymbol{d}_2,\tau k}$ the DOAs in radians. The measurement $\boldsymbol{r}_{\tau k}$ is obtained as the intersection of the rays defined by $(\boldsymbol{d}_1, \boldsymbol{e}_{\boldsymbol{d}_1,\tau k})$ and $(\boldsymbol{d}_2, \boldsymbol{e}_{\boldsymbol{d}_2,\tau k})$. In this work, the narrowband DOA estimates at each TF bin are obtained by considering the instantaneous phase differences observed between different microphone pairs, and by least-squares fitting of the DOA, so that the best least-squares approximation of the theoretical phase differences is obtained (assuming monochromatic plane waves propagating in an anechoic, homogeneous medium). Details of this DOA estimator can be found in [28], however, note that any narrowband estimator can be used in general.

As an alternative to triangulation, a 2D steered response power (SRP), a commonly used localisation method in tracking [17], can be computed at each frequency to obtain a position estimate. For accurate localization, the SRP needs to be evaluated on a dense grid, which can be prohibitive for real-time BSS. Moreover, for spatially separated arrays, the SRP might not be appropriate due to the possibly low signal correlation and different source DOAs at the arrays. Besides the low computational complexity, another advantage of using triangulation is the inherent property to discard large number of outliers: two DOA vectors intersect to provide a position only if their inner product is positive.

Note that the model can be extended to 3D space by estimating the DOAs in 3D, and adding the z-coordinate in the state and measurement vectors. The triangulation can then be done, for instance, by finding the point that minimizes the sum of distances from the rays defined by the DOA vectors.

C. Augmented Measurement Model

Probabilistic models of location-related measurements, as the one described in Section III-A, are typical in the tracking literature [15]. However, in speech applications, where the number of noise-dominated TF bins is large due to the speech sparsity, a model based only on location leads to frequent misclassification of noise-dominated TF bins as speaker-dominated ones. To have a more accurate clutter model, we propose the following augmented measurement that includes the raw signal vectors

$$o_{\tau k} = \{\boldsymbol{r}_{\tau k},\ \boldsymbol{y}(\tau, k)\}. \tag{15}$$

The potential of augmented measurement models for robust multi-source tracking was suggested in [15]. Our motivation to include the signal vector in the augmented measurement is the fact that the signal vectors are commonly used in the multi-channel speech processing literature to build Gaussian signal models that allow for speech presence probability estimation and detection of noisy TF bins [18]. The objective is, therefore, to develop a tracker which utilizes the properties of the signal vector to detect noisy TF bins.

To define the likelihood of observing a given augmented measurement, we note that the 2×1 position estimate $\boldsymbol{r}_{\tau k}$ is obtained by several highly non-linear processing steps from the $M \times 1$ complex-valued signal $\boldsymbol{y}(\tau, k)$, and hence, we can assume that $\boldsymbol{r}_{\tau k}$ and $\boldsymbol{y}(\tau, k)$ are independent RVs. Using the independence, the augmented measurement likelihood can be written as follows

$$f(o_{\tau k} \mid Z_{\tau k}) = f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k})\, f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k}), \tag{16}$$

where $f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k})$ is the standard non-augmented measurement likelihood used in Bayesian trackers and defined previously in (14). For the likelihood $f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k})$ of the signal vector, we used the well-established multivariate Gaussian model [18],

i.e., (τ and k omitted for brevity)

$$f(\boldsymbol{y} \mid Z = 0) = (\pi^M \det[\boldsymbol{\Phi}_{\boldsymbol{v}}])^{-1} \, \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1} \boldsymbol{y}}, \tag{17a}$$

$$f(\boldsymbol{y} \mid Z \neq 0) = (\pi^M \det[\boldsymbol{\Phi}_{\boldsymbol{y}}])^{-1} \, \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}} \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1} \boldsymbol{y}}, \tag{17b}$$

where the PSD matrices $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $\boldsymbol{\Phi}_{\boldsymbol{v}}$ were defined in Section II-A.

D. Derivation of the Dominant Source Label Probability

Using the models introduced in Sections III-A and III-C, we are now able to write the posterior distribution of the dominant source label as $p(Z_{\tau k} \mid o_{\tau k}) \equiv p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))$, and express it using the Bayes theorem as follows

$$p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k)) = \frac{f(\boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k) \mid Z_{\tau k}) \; p(Z_{\tau k})}{f(\boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))}. \tag{18}$$

As the likelihood $f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k})$ in (17) was only provided for $Z_{\tau k} > 0$ and $Z_{\tau k} = 0$ rather than across the full support of $Z_{\tau k}$, we can not directly evaluate (18). To evaluate $p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))$, we further assume that

$$p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k), Z_{\tau k} \neq 0) = p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, Z_{\tau k} \neq 0), \tag{19}$$

which means that the signal $\boldsymbol{y}(\tau, k)$ does not contribute to discrimination between speakers. This assumption is justified, as the motivation to include $\boldsymbol{y}(\tau, k)$ in the augmented measurement was discrimination between noise and speech, rather than discrimination between different speakers.

To derive $p(Z_{\tau k} = z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))$ for $z_{\tau k} \neq 0$, using basic probability axioms we can write (for clarity, we omit τ and k from the following equations)

$$\begin{aligned} p(Z = z \mid \boldsymbol{r}, \boldsymbol{y}) &= p(Z = z \mid \boldsymbol{r}, \boldsymbol{y}, Z \neq 0) \cdot p(Z \neq 0 \mid \boldsymbol{r}, \boldsymbol{y}) \\ &= p(Z = z \mid \boldsymbol{r}, Z \neq 0) \cdot p(Z \neq 0 \mid \boldsymbol{r}, \boldsymbol{y}), \end{aligned} \tag{20}$$

where we used the assumption in (19) to obtain the equality on the second line. The first term in the product in (20) can be expressed using the Bayes theorem and the likelihood models defined in (14) as follows

$$p(Z \mid \boldsymbol{r}, Z \neq 0) = \frac{p(Z \mid Z \neq 0) \, \mathcal{N}(\boldsymbol{r}; \boldsymbol{x}_{\tau Z}, \boldsymbol{\Sigma}_{\tau Z})}{\sum_{z'=1}^{J_\tau} p(Z = z' \mid Z \neq 0) \, \mathcal{N}(\boldsymbol{r}; \boldsymbol{x}_{\tau z'}, \boldsymbol{\Sigma}_{\tau z'})}. \tag{21}$$

The second term in (20), namely $p(Z \neq 0 \mid \boldsymbol{r}, \boldsymbol{y})$, represents the speech presence probability, and can be expressed using the Bayes theorem as

$$p(Z \neq 0 \mid \boldsymbol{r}, \boldsymbol{y}) = \frac{f(\boldsymbol{r}, \boldsymbol{y} \mid Z \neq 0) \; p(Z \neq 0)}{f(\boldsymbol{r}, \boldsymbol{y})}. \tag{22}$$

Recalling the independence of the two components of the augmented measurements, we can express $f(\boldsymbol{r}, \boldsymbol{y})$ as

$$\begin{aligned} f(\boldsymbol{r}, \boldsymbol{y}) = \; & f(\boldsymbol{r} \mid Z \neq 0) \; f(\boldsymbol{y} \mid Z \neq 0) \; p(Z \neq 0) \\ & + f(\boldsymbol{r} \mid Z = 0) \; f(\boldsymbol{y} \mid Z = 0) \; p(Z = 0), \end{aligned} \tag{23}$$

and substituting everything in (22), we obtain an expression for the speech presence probability $p(Z \neq 0 \mid \boldsymbol{r}, \boldsymbol{y})$ in terms of the likelihoods: $f(\boldsymbol{y} \mid Z \neq 0)$, given by (17b), $f(\boldsymbol{y} \mid Z = 0)$, given by (17a), $f(\boldsymbol{r} \mid Z = 0)$, given by the uniform distribution in (14), and $f(\boldsymbol{r} \mid Z \neq 0) = 1 - f(\boldsymbol{r} \mid Z = 0)$.

Therefore, to finally evaluate the posterior distribution of the dominant source index, $p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))$, it remains to obtain $p(Z \neq 0)$ and $p(Z \mid Z \neq 0)$. For the prior speech presence probability, $p(Z \neq 0)$, we used the approach from [29]. As the prior did not notably affect the BSS and can also be a fixed value, we refer the reader to [29], [30] for details. For $p(Z \mid Z \neq 0)$ we use a uniform distribution $p(Z \mid Z \neq 0) = J_\tau^{-1}$, as we assume no prior information regarding the activity of the different sources in the room. The remaining challenging problem in the framework and the evaluation of $p(Z_{\tau k} \mid \boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k))$, is estimating the source states, and the associated estimation error covariances which parametrize the Gaussian distributions in (21).

E. Summary

In this section, we proposed an augmented measurement model which allows us to express the posterior probability of the dominant source label $p(Z_{\tau k} \mid o_{\tau k})$. Previously, in Section II, we discussed that the dominant source label is used directly to obtain TF masks for each source, according to (10) and (11). Hence, by estimating $p(Z_{\tau k} \mid o_{\tau k})$, we are able to estimate the source PSD matrices using standard recursive averaging, according to (8) and (12), and compute source separation filters. As shown in Section III-D, to evaluate $p(Z_{\tau k} \mid o_{\tau k})$, the time-varying parameters of the Gaussian distributions $\mathcal{N}(\boldsymbol{r}; \boldsymbol{x}_{\tau z}, \boldsymbol{\Sigma}_{\tau z})$ for $z \in [1, J_\tau]$ remain to be estimated. As the means and covariances of the Gaussian distributions represent estimates of the source positions and the associated estimation error covariances, this problem can be solved by developing a multi-source tracker, consistent with the proposed models.

IV. Proposed Tracking Framework

In this section, the main objective is to develop a tracking algorithm which can estimate the source states (positions) at each time frame and the associated estimation error covariance matrices. As we discussed in detail in the previous section, these quantities are needed to enable estimation of the dominant source label, and subsequent estimation of PSD matrices and separation filters for the different sources.

Notation: The set of frequency bins with valid measurements at time τ is $\mathcal{K}_\tau$ (as mentioned in Section III-B, frequency bins are discarded as outliers when the dot product of the two DOA vectors used for triangulation is negative). Sets of RVs at time τ are denoted by: $\mathcal{X}_\tau = \{\boldsymbol{x}_{\tau j}\}_{j \in J_\tau}$ are the source states, $\mathcal{Z}_\tau = \{z_{\tau k}\}_{k \in \mathcal{K}_\tau}$ are the dominant source labels, and $\mathcal{O}_\tau = \{\{\boldsymbol{r}_{\tau k}, \boldsymbol{y}(\tau, k)\}\}_{k \in \mathcal{K}_\tau}$. The joint distributions of the RVs in the sets are denoted by $f(\mathcal{X}_\tau)$, $f(\mathcal{O}_\tau)$, and $p(\mathcal{Z}_\tau)$. To distinguish the true source locations (states) $\boldsymbol{x}_{\tau j}$ from the estimated ones by the tracker, the latter are denoted by $\hat{\boldsymbol{x}}_{\tau j}$.

A. Formulation of Tracking as a Missing Data Problem

Assuming that the signals at different frequencies are uncorrelated (common assumption for speech in the STFT-domain), the measurement-to-source associations across the different frequencies at a given time τ are mutually independent. This

assumption has important implications for the development of our tracker: i) in contrast to typical Bayesian trackers where a source generates at most one measurement, in our model, a source can generate multiple measurements per frame (the source can be dominant at multiple frequencies); ii) while an independent associations model would render the well-known JPDA tracker prohibitively complex (see Section IV-C), it allows instead for a formulation where the dominant source index $Z_{\tau k}$ is treated as a hidden variable that can be jointly estimated with the states. This concept is also central to the probabilistic multi-hypothesis tracker (PMHT) proposed in [31].

To obtain the state estimates $\hat{\boldsymbol{x}}_{\tau j}$ when receiving the measurement set $\mathcal{O}_\tau$ at time τ, we assume that the state estimates $\hat{\boldsymbol{x}}_{(\tau-1)j}$ for the $J_{\tau-1}$ sources from the previous frame are given, with respective error covariances $\boldsymbol{P}_{(\tau-1)j}$. The goal is to maximize the following joint distribution, or its logarithm, with respect to $\mathcal{X}_\tau$

$$\begin{aligned}&\ln\big[f(\mathcal{X}_\tau, \mathcal{X}_{\tau-1}, \mathcal{Z}_\tau, \mathcal{O}_\tau)\big]\\ &\quad= \ln\big[f(\mathcal{X}_{\tau-1})\, f(\mathcal{X}_\tau \mid \mathcal{X}_{\tau-1})\, f(\mathcal{O}_\tau \mid \mathcal{X}_\tau, \mathcal{Z}_\tau)\, p(\mathcal{Z}_\tau)\big]. \quad (24)\end{aligned}$$

The factorization follows from the independence of $\mathcal{X}_\tau$ and $\mathcal{Z}_\tau$, and the Markovian property of the model. Using the independence of the source tracks, the terms in (24) can be written as follows

$$f(\mathcal{X}_{\tau-1}) = \textstyle\prod_{j=1}^{J_{\tau-1}} \mathcal{N}(\boldsymbol{x}_{(\tau-1)j}; \hat{\boldsymbol{x}}_{(\tau-1)j}, \boldsymbol{P}_{(\tau-1)j}) \quad (25a)$$

$$f(\mathcal{X}_\tau \mid \mathcal{X}_{\tau-1}) = \textstyle\prod_{j=1}^{J_{\tau-1}} \mathcal{N}(\boldsymbol{x}_{\tau j}; \hat{\boldsymbol{x}}_{(\tau-1)j}, \boldsymbol{P}_{(\tau-1)j} + \boldsymbol{Q}_j), \quad (25b)$$

where the covariance term $\boldsymbol{Q}_j$ in $f(\mathcal{X}_\tau \mid \mathcal{X}_{\tau-1})$ is added due to the source motion uncertainty, defined in (13). Using the independence of associations across measurements, and the independence of $\boldsymbol{y}(\tau, k)$ and $\boldsymbol{r}_{\tau k}$, we can write

$$\begin{aligned}f(\mathcal{O}_\tau \mid \mathcal{X}_\tau, \mathcal{Z}_\tau) = &\textstyle\prod_{k\in\mathcal{K}_\tau} f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{\tau Z_{\tau k}})\\ &\times\ f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k}, \boldsymbol{x}_{\tau Z_{\tau k}}). \quad (26)\end{aligned}$$

The likelihood terms in $f(\mathcal{O}_\tau \mid \mathcal{X}_\tau, \mathcal{Z}_\tau)$ were defined in Section III-D: according to the model in (17), the signal $\boldsymbol{y}(\tau, k)$ does not depend on the source state, and hence $f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k}, \boldsymbol{x}_{\tau Z_{\tau k}}) \equiv f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k})$. Regarding the term $f(\hat{\boldsymbol{r}}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{\tau Z_{\tau k}})$, we have $f(\hat{\boldsymbol{r}}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{t Z_{tk}}) \equiv f(\hat{\boldsymbol{r}}_{\tau k} \mid Z_{\tau k})$, where $f(\hat{\boldsymbol{r}}_{\tau k} \mid Z_{\tau k})$ was defined in (14), and the dependency on the source state was implicitly included in the mean of the Gaussian distribution.

Finally, substituting (25) and (26) in (24), and omitting $f(\mathcal{X}_{\tau-1})$, $p(\mathcal{Z}_\tau)$ and $f(\boldsymbol{y}(\tau, k) \mid Z_{\tau k})$, which do not depend on the states at time τ, the function to be maximized with respect to the states $\mathcal{X}_\tau$ is

$$\mathcal{J}(\mathcal{X}_\tau) = \sum_{j=1}^{J_{\tau-1}} \ln f(\boldsymbol{x}_{\tau j} \mid \hat{\boldsymbol{x}}_{(\tau-1)j}) + \sum_{k\in\mathcal{K}_\tau} \ln f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{\tau j}). \quad (27)$$

As $Z_{\tau k}$ is unknown, the cost function (27) can not be directly maximized. Instead, we start with an initial guess of $\mathcal{X}_\tau$, denoted by $\mathcal{X}'_\tau$, and maximize the conditional expectation of (27). The conditional expectation is taken with respect to to the dominant source index posterior probability distribution $p(\mathcal{Z}_\tau \mid \mathcal{O}_\tau, \mathcal{X}'_\tau)$, which is evaluated using the initial state estimates. Setting the initial estimates $\mathcal{X}'_\tau$ to the previous state estimates $\widehat{X}_{\tau-1}$, mimics a Bayesian one-step prediction, and the conditional expectation of (27) is given by

$$\begin{aligned}Q(\mathcal{X}_\tau \mid \widehat{\mathcal{X}}_{\tau-1}) = &\sum_{j=1}^{J_{\tau-1}} \ln f(\boldsymbol{x}_{\tau j} \mid \hat{\boldsymbol{x}}_{(\tau-1)j})\\ &+ \sum_{j=1}^{J_{\tau-1}} \sum_{k\in\mathcal{K}_\tau} p(\mathcal{Z}_{\tau k} = j \mid o_{\tau k}, \hat{\boldsymbol{x}}_{(\tau-1)j}) \ln f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{\tau j}). \quad (28)\end{aligned}$$

It can be recognized that besides mimicking a traditional Bayesian one-step prediction, the described procedure represents an iteration of the expectation maximization (EM) algorithm [32], and hence it is guaranteed that the new state estimates $\hat{\mathcal{X}}_\tau$ which maximize (28), will increase the likelihood (27). The steps to compute $p(\mathcal{Z}_{\tau k} \mid o_{\tau k}, \hat{\boldsymbol{x}}_{(\tau-1)j})$ were described in Section III-D, where the dependency on the states was implicit via the Gaussian likelihood in (21). It is very important to highlight the reason for including the state $\hat{\boldsymbol{x}}_{(\tau-1)j}$ in the notation $p(\mathcal{Z}_{\tau k} \mid o_{\tau k}, \hat{\boldsymbol{x}}_{(\tau-1)j})$ explicitly, unlike in the dominant source index posterior distribution $p(\mathcal{Z}_{\tau k} \mid o_{\tau k})$ that is used to update the source PSD matrices. Namely, the former, known as the measurement-to-source association probability in the tracking literature, provides a way for the tracker to probabilistically associate the different measurements to the sources based on previous state estimates, and update the state estimates at the current frame. The latter on the other hand, i.e., our dominant source index posterior distribution used to update PSD matrices, needs to be evaluated using the updated state and noise covariances at time τ.

For brevity of the subsequent derivation of the updated state estimates, we introduce the following notation

$$\beta_{\tau kj} = p(\mathcal{Z}_{\tau k} = j \mid o_{\tau k}, \hat{\boldsymbol{x}}_{(\tau-1)j}), \ \ \xi_{\tau j} = \textstyle\sum_{k\in\mathcal{K}_\tau} \beta_{\tau kj}. \quad (29)$$

Maximization of (28) reduces to independent maximization of the following function defined for each $j \in [1, J_{\tau-1}]$,

$$Q_j(\boldsymbol{x}_{\tau j}) = \ln f(\boldsymbol{x}_{\tau j} \mid \hat{\boldsymbol{x}}_{(\tau-1)j}) + \sum_{k\in\mathcal{K}_\tau} \beta_{\tau kj} \ln\big[f(\boldsymbol{r}_{\tau k} \mid Z_{\tau k}, \boldsymbol{x}_{\tau j})\big]. \quad (30)$$

Substituting the Gaussian distributions from (25a), and setting the gradient with respect to $\boldsymbol{x}_{\tau j}$ to zero, we obtain the following estimate of the source state $\boldsymbol{x}_{\tau j}$ (derivation in the appendix)

$$\hat{\boldsymbol{x}}_{\tau j} = \hat{\boldsymbol{x}}_{(\tau-1)j} + \boldsymbol{G}_{\tau j}(\tilde{\boldsymbol{r}}_{\tau j} - \hat{\boldsymbol{x}}_{(\tau-1)j}), \quad (31)$$

where $\boldsymbol{G}_{\tau j}$, $\tilde{\boldsymbol{r}}_{\tau j}$, and $\tilde{\boldsymbol{\Sigma}}_{\tau j}$ are defined as follows

$$\boldsymbol{G}_{\tau j} = \boldsymbol{P}^{(j)}_{\tau|\tau-1}(\boldsymbol{P}^{(j)}_{\tau|\tau-1} + \tilde{\boldsymbol{\Sigma}}_{\tau j})^{-1}, \quad (32)$$

$$\tilde{\boldsymbol{r}}_{\tau j} = \frac{1}{\xi_{\tau j}} \sum_{k\in\mathcal{K}_\tau} \beta_{\tau kj}\, \boldsymbol{r}_{\tau k}, \ \ \tilde{\boldsymbol{\Sigma}}_{\tau j} = \frac{1}{\xi_{\tau j}} \boldsymbol{\Sigma}_{\tau j}, \quad (33)$$

where we denoted $\boldsymbol{P}^{(j)}_{\tau|\tau-1} = \boldsymbol{P}_{(\tau-1)j} + \boldsymbol{Q}_j$, for consistency with the standard notation of prediction error covariance [15], and $\boldsymbol{\Sigma}_{\tau j}$ is the measurement noise covariance matrix. Equation (31) has the form of a standard Kalman filter, with the noise covariance $\boldsymbol{\Sigma}_{\tau j}$ scaled by $\xi_{\tau j}$, and the measurement given by a weighted sum of all measurements, where the weight depends on the association probabilities. Large $\xi_{\tau j}$ indicates that more

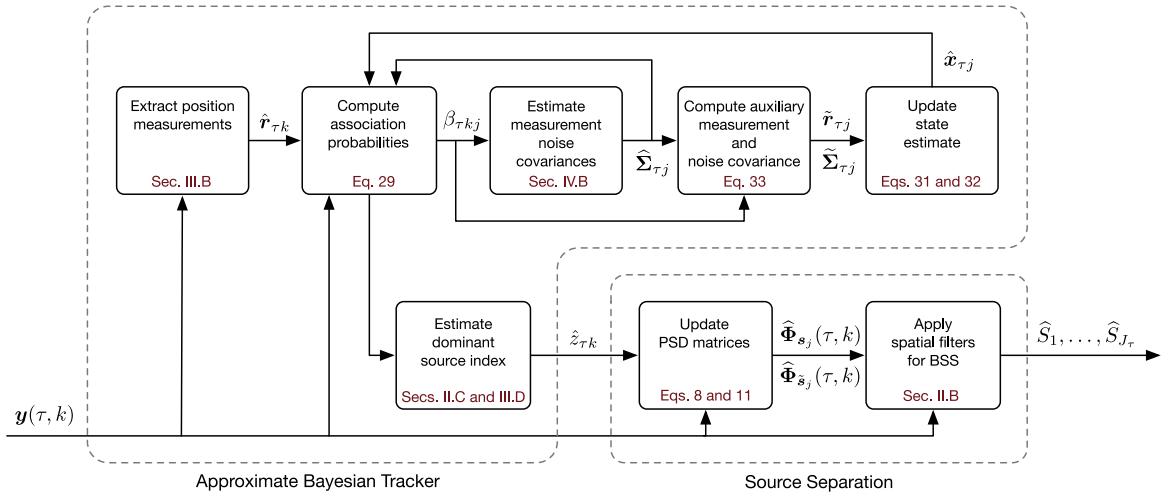

Fig. 1. The processing blocks of the proposed ISF-based system for blind source separation.

measurements in a frame originate from source j. As a result, $\boldsymbol{\Sigma}_{\tau j}$ is reduced via the scaling (33) and the filter puts more emphasis on the measurements. For robustness, the state of source j is updated only if $\xi_{\tau j}$ exceeds a threshold ξ_{thr}, i.e., there is sufficient evidence that source j is active.

Note that the prediction covariances $\boldsymbol{P}_{\tau|\tau-1}^{(j)}$ empirically assign uncertainty to the state estimates. At frame τ, $\boldsymbol{P}_{\tau|\tau-1}^{(j)}$ is based on the system noise and the number of frames $\Delta\tau_j$ since the last update of the state of source j, i.e.,

$$\boldsymbol{P}_{\tau|\tau-1}^{(j)} = (\Delta\tau_j + 1)\boldsymbol{Q}_j. \tag{34}$$

Hence, after a silent period, the filter puts less emphasis on the predicted state, and more emphasis on the measurements, which is a desired response for robustness to speech pauses. This is in contrast to JPDA, where $\boldsymbol{P}_{\tau|\tau-1}^{(j)}$ represents prediction covariance in a strict statistical sense [15].

B. Estimation of Measurement Noise Covariance Matrices

So far, the covariance matrices $\boldsymbol{\Sigma}_{\tau j}$ were assumed known. In practice, they can be estimated from the measurements from the past L frames and the association probabilities, as done in EM-based clustering [13] frameworks, i.e.,

$$\widehat{\boldsymbol{\Sigma}}_{\tau j} = \frac{\sum_{t,k}\ p(Z_{tk} = j \mid o_{tk})(\boldsymbol{r}_{tk} - \bar{\boldsymbol{r}}_{\tau j})(\boldsymbol{r}_{tk} - \bar{\boldsymbol{r}}_{\tau j})^{\mathrm{T}}}{\sum_{t,k}\ p(Z_{tk} = j \mid o_{tk})}, \tag{35}$$

where the sum is over $t \in [\tau - L, \tau - 1]$, and $\bar{\boldsymbol{r}}_{\tau j}$ is the probabilistically weighted sample mean for the L frames

$$\bar{\boldsymbol{r}}_{\tau j} = \frac{\sum_{t,k}\ p(Z_{tk} = j \mid o_{tk}) \cdot \boldsymbol{r}_{tk}}{\sum_{t,k}\ p(Z_{tk} = j \mid o_{tk})}. \tag{36}$$

However, our experiments indicated that the estimation of $\widehat{\boldsymbol{\Sigma}}_{\tau j}$ as given by (35) and (36) is not always robust in real rooms with moderate reverberation times and concurrent moving speakers. While the proposed augmented measurements are robust to noise, the problem is caused by speech-dominated bins, which due to reverberation are inaccurately localized. If such TF occur frequently, they cause an inadequate increase of the noise covariances, which in turn impedes the tracker from updating the states, and increasing the risk of track losses.

To alleviate the problem, we introduce a data-dependent weight $b_{\tau k}(\boldsymbol{r}_{\tau k})$ for each TF-bin as follows: between each two $\boldsymbol{r}_{\tau_1 k_1}$ and $\boldsymbol{r}_{\tau_2 k_2}$, for $\tau_1, \tau_2 \in [\tau - L, \tau - 1]$, compute a distance measure $a(\boldsymbol{r}_{\tau_1 k_1}, \boldsymbol{r}_{\tau_2 k_2})$, and assign a weight to each point as the sum of the distances from other points i.e.,

$$a(\boldsymbol{r}_{\tau_1 k_1}, \boldsymbol{r}_{\tau_2 k_2}) = \mathrm{e}^{-\|\boldsymbol{r}_{\tau_1 k_1} - \boldsymbol{r}_{\tau_2 k_2}\|_2}, \tag{37}$$

$$b(\boldsymbol{r}_{\tau k}) = \sum_{\boldsymbol{r}' \neq \boldsymbol{r}_{\tau k}} a(\boldsymbol{r}_{\tau k}, \boldsymbol{r}'), \tag{38}$$

and normalize the weights $b(\boldsymbol{r}_{\tau k})$ such that they sum to 1. The weight $b(\boldsymbol{r}_{\tau k})$ indicates the average proximity to other points and increases robustness against isolated outliers. Thus, instead of using $p(Z_{\tau k} = j \mid o_{\tau k})$ for the noise covariance matrix updates in (35) and (36), we use the modified weight

$$p_{j\tau k} = p(Z_{\tau k} = j \mid o_{\tau k}) \cdot b_{\tau k}(\boldsymbol{r}_{\tau k}). \tag{39}$$

Our experiments showed that in scenarios with several concurrent sources in reverberant rooms, $b(\boldsymbol{r}_{\tau k})$ notably increases the robustness of the tracker.

The processing blocks of the full tracking and BSS system for a given number of sources are illustrated in Fig. 1.

C. Relation to JPDA and PMHT Trackers

To elaborate why the well-known JPDA tracker is not suited for our narrowband models, recall that the JPDA assumes at most one measurement per source at time τ. To spot the implications, consider the JPDA state estimate $\hat{\boldsymbol{x}}_{\tau j}$ that is given by

$$\hat{\boldsymbol{x}}_{\tau j}^{\text{JPDA}} = \mathrm{E}\left[\boldsymbol{x}_{\tau j} \mid \mathcal{O}\right] = \textstyle\sum_{\boldsymbol{Z}_\tau} \mathrm{E}\left[\boldsymbol{x}_j \mid \boldsymbol{Z}_\tau, \mathcal{O}\right]\ p(\boldsymbol{Z}_\tau \mid \mathcal{O}), \tag{40}$$

where the association events are given by the $K_\tau \times J_\tau$ random matrix $\boldsymbol{Z}_\tau$ whose support is the set of all matrices with at most one entry equal to 1 per row and column. If each source produces multiple measurements, the support of $\boldsymbol{Z}_\tau$ contains all matrices with maximum one 1 per row, and an arbitrary number of ones per column. The cardinality of such support is very large (the number of ways to distribute K_τ balls in $J_\tau + 1$ boxes) and evaluating (40) is not manageable in real-time.

Algorithm 1: Implementation of the Proposed Detection, Tracking and BSS System.

1: **Given from frame** $\tau-1$**:** states $\{\hat{\boldsymbol{x}}_{\tau-1,j}\}_{j=1:J_{\tau-1}}$, measurement error covariance matrices $\{\widehat{\boldsymbol{\Sigma}}_{\tau-1,j}\}_{j=1:J_{\tau-1}}$, vector containing the τ_{ttl} counters for each source.
2: **do for each frame** τ:
3: Estimate narrowband DOAs for each array and each frequency bin k.
4: Estimate the posterior speech presence probability for each frequency bin k (Section III-D). The required likelihoods of the STFT signal vectors are obtained using the noise PSD matrix from the previous frame.
5: Update the noise PSD matrix $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}}(\tau,k)$ in (8) for each k, using the posterior speech presence probability from `line 4`.
6: Triangulate the DOAs from the different arrays to obtain the positions measurement set $\mathcal{K}_\tau$.
7: Propagate the state estimates for each source j (prediction step): $\hat{\boldsymbol{x}}_{\tau,j} \equiv \hat{\boldsymbol{x}}_{\tau-1,j}$.
8: Use the current state and noise covariance estimates of the tracker to compute $\mathcal{P}_\tau(\boldsymbol{r})$ in (43).
9: Use the position estimates from last T frames to compute $\mathcal{L}_\tau(\boldsymbol{r})$ in (42).
10: Compute $\mathcal{J}_\tau(\boldsymbol{r}) = \mathcal{L}_\tau(\boldsymbol{r}) \cdot \mathcal{P}_\tau(\boldsymbol{r})$ and detect a new source with a state at the maximum of $\mathcal{J}_\tau(\boldsymbol{r})$, if $\max[\mathcal{J}(\boldsymbol{r})] > \chi_{\text{thr}}$.
11: Update the state prediction error covariance $\boldsymbol{P}^{(j)}_{\tau|\tau-1}$ in (34) for each source j.
12: Update the measurement noise covariance matrices $\widehat{\boldsymbol{\Sigma}}_{\tau j}$ for each source j (Section IV-B).
13: Compute the association probabilities $\beta_{\tau kj}$ and the auxiliary values $\xi_{\tau j}$ according to (29), for each source j.
14: Update the source state estimates $\hat{\boldsymbol{x}}_{\tau j}$ according to (31)-(33) for each source j.
15: Check if tracks need to be merged using a Mahalanobis distance-based merger, as done in [5] (merge if two estimated states have a Mahalanobis distance smaller than a threshold).
16: Remove any source for which the time-to-live τ_{ttl} has exceeded the threshold of 2.5 seconds. (Section V-B).
17: Re-compute the association probabilities using the updated states and covariances, and the speech presence probability from `line 7`. These are our required dominant source index posterior probabilities.
18: At each frequency k, find the source j_k^* with largest association probability.
19: Select a reference microphone $m_{\tau j}$ for each source j, such that the reference is closest to the state estimate $\hat{\boldsymbol{x}}_{\tau j}$.
20: Update the PSD matrices $\hat{\boldsymbol{\Phi}}_{\boldsymbol{s}_{j_k^*}}(\tau,k)$ and the RTF vectors $\boldsymbol{g}_{j_k^* m_{\tau j_k^*}}$ for all k.
21: Update the PSD matrices $\hat{\boldsymbol{\Phi}}_{\tilde{\boldsymbol{s}}_j}(\tau,k)$ at each k for $j \neq j_k^*$.
22: Estimate $\widehat{S}_1, \ldots \widehat{S}_{J_\tau}$ using the updated informed MVDR filters (Section II-B).
23: **end for**

Instead, a model where a source can generate multiple measurements per frame allows to formulate the tracking as a missing data problem, as in the PMHT proposed in [31]. The source states in PMHT are found by solving similar optimisation problem as (27), jointly across frames. As batch processing is not suitable for online BSS, our approach can be considered as an online variant of the PMHT. Originally used for applications where single measurement per source was expected, the PMHT received criticism due to its model violation [33]. A related criticism is the so-called *hospitality*, meaning that multiple measurements decrease the noise covariance $\tilde{\boldsymbol{\Sigma}}_{\tau j}$, in contrast to the JPDA, where multiple measurements increase the innovations covariance [15]. However, for narrowband models used in our framework, PMHT paradigm is well-suited, as multiple measurements indicate likely source activity in a frame and the reduction of noise covariance allows the tracker to promptly update the state.

A PMHT drawback which also affects our system is the larger tendency to track losses compared to JPDA. This stems from the fact that the association probabilities $\beta_{\tau kj}$ are computed with the noise covariance $\widehat{\boldsymbol{\Sigma}}_{\tau j}$, while in JPDA the noise plus the prediction covariance is used (innovation covariance). Thus, after speech pauses, JPDA considers measurements from a wider area and is less prone to lost tracks. Therefore, to increase the robustness to speech pauses, we added the prediction covariance $\boldsymbol{P}^{(j)}_{\tau|\tau-1}$ when evaluating the association probabilities as follows

$$\beta_{\tau kj} = \mathcal{N}\left(\boldsymbol{r}_{\tau k}\,;\, \boldsymbol{x}_{\tau j},\, \widehat{\boldsymbol{\Sigma}}_{\tau j} + \boldsymbol{P}^{(j)}_{\tau|\tau-1}\right). \tag{41}$$

V. Track Management

In the previous sections, we presented the processing steps for a multi-source tracker, which similarly as the JPDA and PMHT, estimates the tracks of a known number of sources. In practice, a tracking system should provide a robust track management as well, which detects and discards sources as they appear and disappear. In this section, we present a track management system for the proposed tracker, which detects new sources by utilising both the measurements, as well as the estimated states and covariances of the sources in track. A pseudo-code of the implementation of the full framework, including the source detection and removal mechanisms, is given in Algorithm 1.

A. Source Detection

To detect new sources, we propose a function $\mathcal{J}_\tau(\boldsymbol{r}) = \mathcal{L}_\tau(\boldsymbol{r}) \times \mathcal{P}_\tau(\boldsymbol{r})$ which is designed to reflect the confidence that a new source appears at position $\boldsymbol{r}$. The first term, $\mathcal{L}_\tau(\boldsymbol{r})$, is a low-resolution grid where each cell is associated with the number of narrowband position estimates in that cell, collected over

the last T frames, i.e.,

$$\mathcal{L}_\tau(\boldsymbol{r}) = \sum_{t=\tau-T+1}^{\tau} \sum_{k\in\mathcal{K}_t} I(\boldsymbol{r}, \boldsymbol{r}_{tk}), \tag{42}$$

where $I(\boldsymbol{r}, \boldsymbol{r}_{tk}) = 1$ if $\boldsymbol{r}_{tk}$ is within the cell centered at $\boldsymbol{r}$, and 0 otherwise. Next, to avoid duplicate tracks and de-emphasize regions with already tracked sources we propose the following function for the second term $\mathcal{P}_\tau(\boldsymbol{r})$

$$\mathcal{P}_\tau(\boldsymbol{r}) = \sum_{j=1}^{J_\tau} 1 - \mathrm{e}^{\frac{1}{2}(\boldsymbol{r}-\hat{\boldsymbol{x}}_{\tau j})^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{\tau j}+\boldsymbol{P}_{\tau|\tau-1})^{-1}(\boldsymbol{r}-\hat{\boldsymbol{x}}_{\tau j})}. \tag{43}$$

$\mathcal{P}_\tau(\boldsymbol{r})$ is subsequently normalize so that $\max[\mathcal{P}(\boldsymbol{r})] = 1$. Note that the terms of $\mathcal{J}(\boldsymbol{r})$ resemble a likelihood term (observed data) and a prior term (information from existing tracks). A new source is declared if $\max[\mathcal{J}(\boldsymbol{r})] > \chi_{\text{thr}}$, with initial state at the location of the maximum and noise covariance equal to a diagonal matrix with equal entries on the diagonal. A good threshold χ_{thr} is found by empirical examination, where it is important to promptly detect sources, even at the cost of false alarms, as false tracks can be easily discarded (see Section V-B). Values $\chi_{\text{thr}} = [1, 3]$ have shown to work well in many scenarios, where in adverse conditions, lower values are preferred.

In the rare case of duplicate tracks assigned to one source, a merging mechanism based on Mahalanobis distance is implemented, similarly as in [5]. Note that with such mechanism, tracks of crossing speakers will be merged as well. If the sources move apart, the tracks split by re-detecting one of the speakers as a new source. Although in certain applications it might be desired to maintain separate tracks while crossing, the system in this work was developed for BSS, where closely located speakers cannot be separated based on spatial information alone, and hence the merging process is justified.

B. Source Removal

As described in Section IV, a source track is updated only if $\xi_{j\tau} > \xi_{\text{thr}}$, i.e., if there is sufficient evidence that source j is active. Let τ_{ttl} (known as *time-to-live* in the tracking literature) denote the number of consecutive frames a source can be silent before removed. The value τ_{ttl} should be chosen such that short speech pauses do not affect the source track. However, as the source can not be tracked when inactive for a longer period, it is reasonable to discard the track, and re-detect the source when speech is resumed. In this work, we set $\tau_{\text{ttl}} = 78$, corresponding to 2.5 seconds.

VI. Performance Evaluation

To evaluate the proposed BSS approach for moving sources, we used simulated and measured data. The measurement-to-source association accuracy is evaluated in Section VI-B, and the objective quality of the separated signals is evaluated in Sections VI-C and VI-D, and compared to different state-of-the-art approaches. Due to space constraints, we do not evaluate the tracking performance in terms of error between the true and estimated source trajectories. However, the reader can find further experiments regarding this aspect at *https://www.audiolabs-erlangen.de/resources/2017-IEEE-BSS-tracking*

A. Experimental Setup

Measurements were conducted in a room with reverberation $T_{60} \approx 0.3$ s, using three circular arrays with diameter 2.7 cm and three omnidirectional DPA microphones per array (model d:screet SMK-SC4060). Speech samples (male and female speech in English, German and French) were emitted using Focal loudspeakers (model CMS40). In addition to the sensor noise present in the signals, air conditioner noise was recorded during speech absence, and added to the microphone signals with a given signal-to-noise ratio (SNR). In the simulations, a room with the same array geometry was simulated at different reverberation times, where clean speech signals were convolved with room impulse responses for moving sources simulated using [34]. Diffuse babble noise signals at the microphones were generated according to [35], and scaled to achieve a specified SNR. In addition, uncorrelated Gaussian sensor noise was added with a speech-to-uncorrelated noise ratio of approximately 35 dB in all experiments. In each of the experiments, the sources are continuously active (with typical short speech pauses) for 20 seconds and traverse the indicated trajectories multiple times with velocities 0.12-0.4 m/s. The SNRs with respect to the different sources are in the range [1, 6] dB, and the signal-to-interference ratios (SIRs) are in the range [−5.5, −0.5] dB.

The processing was done at a sampling rate of 16 kHz, with an STFT frame size of 64 ms with 50% overlap, windowed by a Hamming window. The averaging constants for the PSD matrix estimation, α_s and α_v, were set to 0.75 and 0.98, respectively (corresponding to time constants of 0.11 s and 1.58 s). Note that as the PSD matrices are updated at each TF bin, the informed MVDR filters are also re-computed at each TF bin. Although such implementation provides the fastest filter adaptation to moving sources, depending on the available computing power, the filters can also be implemented with a smaller update rate, at the cost of a reduced adaptation speed. The room was uniformly sampled with 10 samples per meter to evaluate the function $\mathcal{J}(\boldsymbol{r})$ for new source detection. The number of frames T considered for source detection and L for estimation of the measurement noise covariance matrix were $T = 10$ and $L = 30$, respectively. Note that as spatial aliasing in the DOA estimates for the given array geometry occurs around 7 kHz, the signals were band-limited to 7 kHz before processing.

B. Association Accuracy and Detection Delay

To evaluate the performance for different reverberation levels, we used simulations for this experiment. All sources have approximately equal power, where the SNR with respect to each source is approximately 9 dB.

The measurement-to-source association is evaluated in terms of false positive rate (FPR), and false negative rate (FNR), which

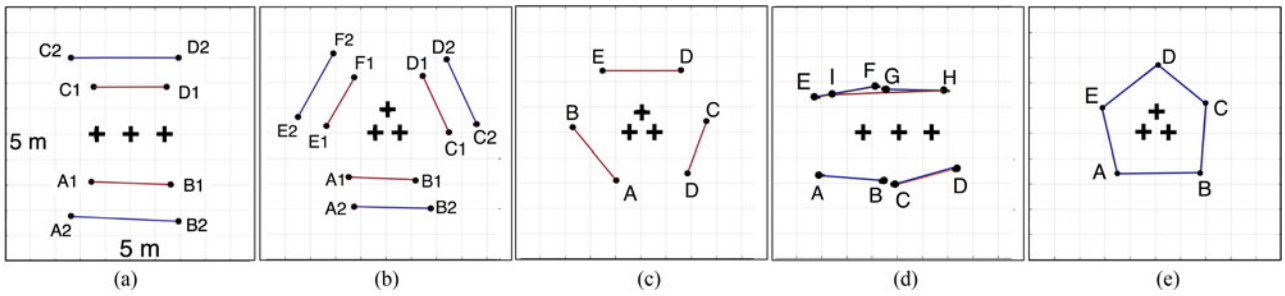

Fig. 2. Scenarios used for evaluation. The line segments denote source trajectories (traversed multiple times back and forth), and the crosses denote microphone arrays. Line segments of different color correspond to different simulations.

for each source j are defined as

$$\mathrm{FPR}(j) = {\sum_{t,k}[\hat{z}_{tk}=j \wedge z_{tk}\neq j]}\Big/{\sum_{t,k}[z_{tk}\neq j]},$$
$$\mathrm{FNR}(j) = {\sum_{t,k}[\hat{z}_{tk}\neq j \wedge z_{tk}= j]}\Big/{\sum_{t,k}[z_{tk}= j]}, \quad (44)$$

where $\sum_{t,k}[\cdot]$ denotes a sum over all TF bins of the value of the logical expression in the brackets. The true value of z_{tk} indicates the source with maximum instantaneous power at TF bin (t, k). To obtain the final measure, the FPRs and FNRs are averaged across all the sources in a given experiment.

As a state-of-the-art framework for online TF mask estimation of moving sources, which is an equivalent problem as the measurement-to-source association, we consider the DOA-based algorithm by Loesch and Yang proposed in [10], and denoted by L-Y in the following. To have a more fair comparison, where the array arrangement is adapted to the particular framework, an additional array setup is simulated for L-Y to capture the acoustic scene, where all microphones from the distributed arrays are now placed in a single compact array with the same diameter as the other arrays. The simulated scenarios are illustrated in Fig. 2. The array used for L-Y is placed at the location of the middle array in Fig. 2(a), and at the centroid of the triangle defined by the three arrays in Fig. 2(b) and (c). In addition, to avoid possible initialization problems, the algorithm L-Y was initialized with the true angle of each source at the moment when the source appears.

In the description of L-Y in [10] no information is provided regarding the detection of noisy TF bins. Therefore, we employed a simple energy-based voice activity detector (VAD), as follows: from a noise-only period of 5 seconds we computed the average noise power, and whenever the instantaneous power at a given TF bin is not at least 6 dB higher than the average noise power, we declare that TF bin as a noise-dominated, i.e. $\hat{z}_{tk} = 0$. In contrast, one of the advantages of our proposed framework is the fact that by using the augmented measurements, the detection of noisy TF bins is inherently included in the tracking system.

We simulated the following four scenarios for evaluation:

Setup 1: $T_{60} = 0.2$ s, two sources traversing the trajectories A_1-B_1, C_1-D_1 in Fig. 2 (a), with velocity ≈ 0.2 m/s.

Setup 2: $T_{60} = 0.2$ s, three sources traversing A_2-B_2, C_2-D_2, E_2-F_2 in Fig. 2 (b) with velocity ≈ 0.25 m/s.

TABLE I
FALSE POSITIVE AND FALSE NEGATIVE RATES OF THE PROPOSED AND THE STATE-OF-THE-ART FRAMEWORKS FOR MEASUREMENT-TO-SOURCE ASSOCIATION

	FPR		FNR	
Setup	L-Y	Proposed	L-Y	Proposed
1	0.05	**0.01**	**0.50**	0.68
2	0.07	**≈0**	**0.59**	0.85
3	0.20	**0.01**	**0.61**	0.80
4	0.26	**0.01**	**0.59**	0.82

The best result is shown in bold.

Setup 3: $T_{60} = 0.2$ s, three sources traversing A-B, C-D, E-F in Fig. 2 (c) with velocity ≈ 0.2 m/s.

Setup 4: $T_{60} = 0.4$ s, three sources traversing A_1-B_1, C_1-D_1, E_1-F_1 in Fig. 2 (b), with a velocity of $\approx$0.31 m/s. Note that our implementation of L-Y was unable to track the sources when they traversed the trajectories farther from the arrays (A_2-B_2 and C_2-D_2 in Fig. 2(a)).

The setups cover different array geometries, different distances of the sources from the arrays, and a setup where the source locations are less suitable for triangulation (Fig. 2(c)). The FPR and the FNR are summarized in Table I. Distinctively, the proposed framework provides a FPR of at most 0.01 in all cases, while the FPR of L-Y notably increases when the sources are farther away from the array and when the T_{60} increases. Although the FNR of the L-Y and the proposed framework is respectively higher than 0.5 and 0.68, it should be noted that the FNR is less critical than the FPR [36]. While false positives introduce errors in the RTF vectors causing speech distortion, false negatives only indicate that the PSD matrices are not updated as frequently as they could if detection was accurate. Although this leads to sub-optimal undesired signal reduction and could be improved if the FNR is reduced, it does not introduce severe distortion to the estimated source signals. Our experiments showed that while FPR of 0.1 already causes audible distortion, FNRs can reach up to 0.9 while still providing good signal quality.

For completeness, we present the detection delay of the proposed track management mechanism, when multiple sources appear simultaneously in different acoustic conditions, while keeping the parameters of the track management fixed. First we consider scenarios from Fig. 2(a), for $T_{60} = 0.2$ s and $T_{60} = 0.4$ s,

TABLE II
DETECTION DELAY IN SECONDS USING THE SETUP IN FIG. 2(A)

	Scenario 1		Scenario 2	
T_{60} [s]	Source 1 (A_1-B_1)	Source 2 (C_1-D_1)	Source 1 (A_2-B_2)	Source 2 (C_2-D_2)
0.2	0.054 s	0.030 s	0.057 s	0.220 s
0.4	0.056 s	0.032 s	0.059 s	0.160 s

The velocity of all sources is ≈ 0.2 m/s.

TABLE III
DETECTION DELAY IN SECONDS USING THE SETUP IN FIG. 2(B), SOURCES FARTHER FROM THE ARRAY

T_{60}[s]	Velocity	Source 1 (A_2-B_2)	Source 2 (C_2-D_2)	Source 3 (E_2-F_2)
0.2	0.25 m/s	0.058 s	0.220 s	0.320 s
0.4	0.31 m/s	0.057 s	0.032 s	0.103 s
0.6	0.43 m/s	0.058 s	0.030 s	0.300 s

with source velocities of approximately 0.2 m/s. As shown in Table II, when the sources are nearer to the arrays, both are detected almost instantaneously, and when they are farther away, a delay of only 0.2 s is introduced for one of the sources. The delay is also shown when the three sources traversing the trajectories farther from the arrays in Fig. 2(b) appear simultaneously. The results in Table III indicate prompt source detection in all scenarios, without notable increase of the detection delay for the different reverberations and velocities. Note that the detection delay of a given source also depends on the relative position of the source with respect to the arrays. In particular, certain positions are more sensitive to errors in the DOA estimates, for instance those where the DOA vectors intersect at small angles. Due to this reason, in our experiments, source 1 is generally detected with a larger delay than the other sources.

C. Evaluation of Separated Signals Using Simulations

In the simulation-based experiments, we can compute oracle ISFs whose PSD matrices are updated with ideal measurement-to-source associations. Moreover, using true source locations, we can steer a delay-and-sum beamformer (DSB) towards each source as a baseline. The oracle ISFs, the DSBs, and the ISFs from the proposed framework are computed using only the three microphones from the nearest array, whereas the ISFs obtained using L-Y are computed using all nine microphones from the compact array.

1) Performance Measures: For each source, the reference m is time-varying depending on the estimated location. We used the following performance measures:

i) *Speech distortion (SD) index* v_{sd}, attains values in [0, 1] Values close to zero indicate low distortion. For the i-th segment, the SD index is given by

$$v_{\mathrm{sd},j}(i) = {}^{\langle |s_{jm}(t)-\tilde{s}_{jm}(t)|^2\rangle}\!/_{\langle |s_{jm}(t)|^2\rangle},\ t \in \big((i-1)T,\ iT\big], \quad (45)$$

where $s_{jm}(t)$ is the clean time-domain signal of source j at the m-th microphone and $\tilde{s}_{jm}(t)$ is the time-domain signal of source j after applying the separation filter for source j. To consider only segments where source j is present, the median of segment-wise signal power was computed, and the segments by 10 dB lower than the median were excluded.

TABLE IV
EVALUATION RESULTS OF THE DIFFERENT BASELINES AND THE PROPOSED BSS FRAMEWORK IN SIMULATED SCENARIOS

Setup		Oracle	DSB	L-Y	Proposed
1	v_{SD}	0.03	**0.01**	0.04	0.04
	IR [dB]	16.0	1.5	**15.0**	**15.0**
	NR [dB]	5.8	0.2	3.9	**5.0**
	Δ_{PESQ}	0.95	0.05	0.61	**0.82**
	Δ_{STOI}	0.21	0.03	0.15	**0.19**
2	v_{SD}	0.03	**0.01**	0.06	0.04
	IR [dB]	15.0	1.0	**13.3**	13.0
	NR [dB]	5.8	0.2	3.7	**5.8**
	Δ_{PESQ}	0.81	0.05	0.53	**0.60**
	Δ_{STOI}	0.25	0.03	0.21	**0.22**
3	v_{SD}	0.04	**0.01**	0.07	0.05
	IR [dB]	14.9	1.0	**13.3**	12.8
	NR [dB]	5.7	0.2	3.7	**5.8**
	Δ_{PESQ}	0.80	0.05	0.40	**0.51**
	Δ_{STOI}	0.24	0.03	0.17	**0.19**
4	v_{SD}	0.07	**0.03**	0.18	0.13
	IR [dB]	9.3	0.8	**8.4**	**8.4**
	NR [dB]	4.8	0.2	2.9	**5.6**
	Δ_{PESQ}	0.51	0.03	0.21	**0.31**
	Δ_{STOI}	0.20	0.02	0.13	**0.15**

The best result (not considering the oracle filter) is indicated in bold.

ii) *Interference reduction* Δ_{IR} *and noise reduction* Δ_{NR},

$$\Delta_{\mathrm{NR}}(i) = 10\log_{10}{}^{\langle |v_m(t)|^2\rangle}\!/_{\langle |\tilde{v}_m(t)|^2\rangle},\ t \in \big((i-1)T,\ iT\big]$$
$$\Delta_{\mathrm{IR}}(i) = 10\log_{10}{}^{\langle |\sum_{j'\neq j} s_{j'm}(t)|^2\rangle}\!/_{\langle |\sum_{j'\neq j} \tilde{s}_{j'm}(t)|^2\rangle}, \quad (46)$$

where $v_m(t)$ denotes the time-domain noise signal at the m-th microphone and tilde denotes filtering by the separation filter for the source that is evaluated (Δ_{IR} and Δ_{NR} are computed for each source separately).

iii) Improvement in terms of short-time objective intelligibility (STOI) [37] and perceptual evaluation of speech quality (PESQ) scores [38], compared to the reference mixture.

2) Discussion: The objective results for the same setups as in Section VI-B are summarized in Table IV. All results are averaged across all sources in the given scenario. Further evaluation using measured data, where the extracted signal for each source is evaluated separately, is provided in Section VI-D. The insufficient ability of the DSB to reduce undesired speakers is clearly demonstrated. Therefore, even if a perfect tracker would be provided, fixed spatial filters do not achieve sufficiently good source separation. Comparing the proposed and the L-Y approach, we note that the proposed approach achieves by 2-3 dB better noise reduction, which can be attributed to the good accuracy when detecting noisy bins using the proposed augmented measurements compared to a simple energy-based detector. The lower FNR of L-Y than the proposed approach is manifested in the slightly better interference reduction of L-Y (at most 0.5 dB), while the critically large FPR of L-Y manifests itself in the speech distortion, as well as in the worse noise reduction

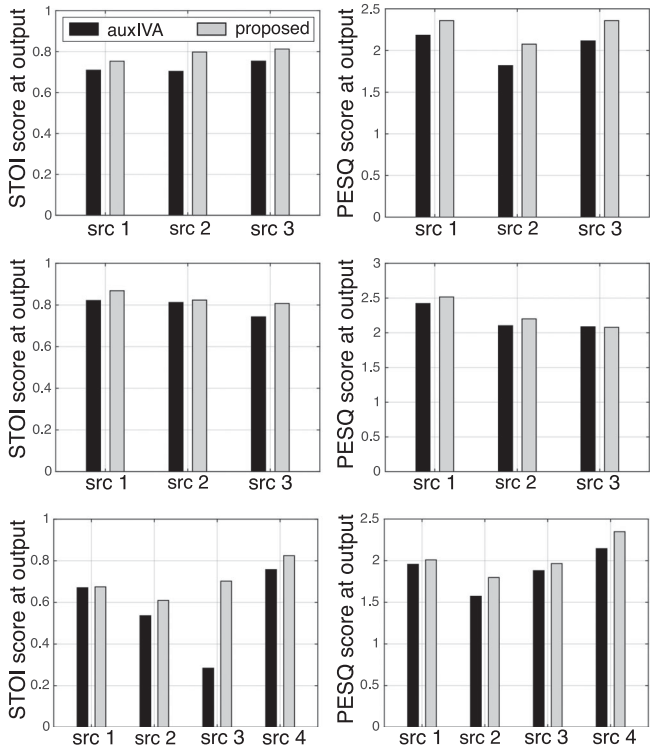

Fig. 3. STOI and PESQ scores of the separated source signals at the output of the proposed BSS system and at the output of auxIVA. Top: Setup 1; middle: Setup 2; bottom: Setup 3.

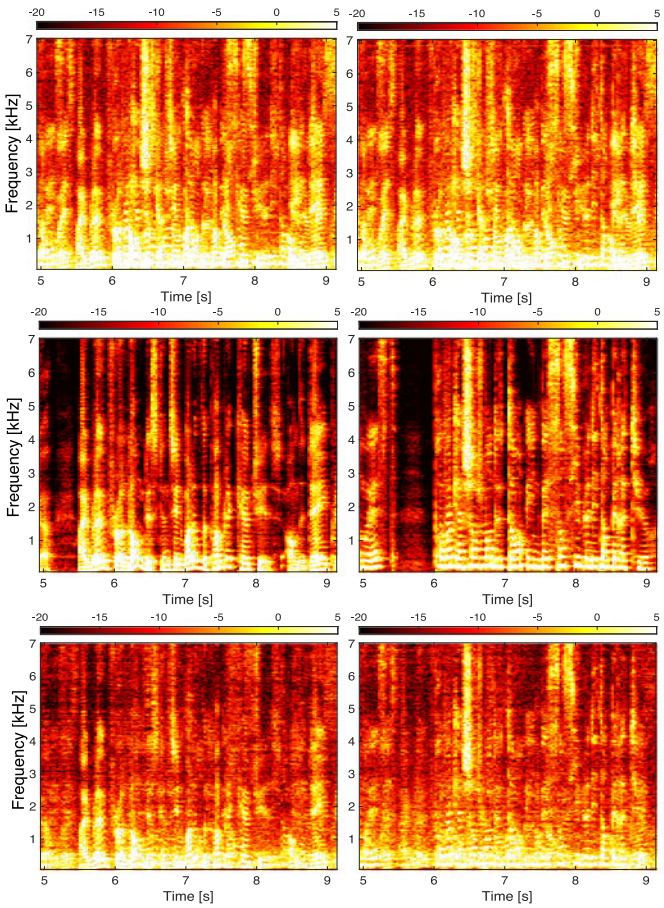

Fig. 4. Spectrograms from Setup 1. Top: mixtures; middle: clean speech signals; bottom: separated signals by the proposed BSS framework.

performance. As a result, the proposed method outperforms L-Y in terms of PESQ and STOI scores.

For higher T_{60}, the performance of all filters, including the oracle ones, deteriorates. The spatial filters have limited ability to reduce interferers as the RTFs of the sources are not fully captured within an STFT frame. Finally, note that as the reverberant signal is used as a reference when computing the SD, the increase in the SD index is partially due to dereverberation.

D. Evaluation of Separated Signals Using Measurements

Using data from real measurements, we compare the proposed BSS system to a powerful state-of-the-art BSS based on auxiliary function-based independent vector analysis, proposed in [39] and denoted by auxIVA in the following. Three experimental setups were evaluated, where the velocities of all sources were approximately 0.3 m/s.:

Setup 1: Fig. 2(d): three sources, female English speaker traverses A-B, male German speaker traverses C-D, and male French speaker traverses H-I.

Setup 2: Fig. 2(e): three sources, female French speaker traverses B-C, male English speaker traverses A-B, male French speaker traverses E-D.

Setup 3: Fig. 2(d): four sources, male German speaker traverses A-B, female English speaker traverses C-D, male English speaker traverses G-H and male French speaker traverses E-F.

As the processed signals by auxIVA (using the implementation by the authors in [39]) provided better results when using only the microphones from one array, we applied auxIVA with each of the three arrays, and manually picked the best result for each source. In contrast, our proposed framework selects the reference array depending on the estimated source location, thereby providing a solution to the microphone subset selection problem [21], [22]. Due to the possibly different reference microphones for the proposed system and auxIVA, we do not provide the PESQ and STOI improvements with respect to the reference, but rather the final scores of the separated signals. Note that we did not have the filtered versions of the clean signals by the auxIVA, and hence, we were unable to compute the SD, the NR, and the IR. Nonetheless, the PESQ and STOI scores in Fig. 3 indicate that the proposed approach consistently outperforms auxIVA. It is worthwhile mentioning that the proposed approach provided superior background noise reduction: one advantage of spatial filtering-based BSS compared to IVA is the fact that noise reduction is explicitly addressed by the spatial filters. For an impression of the audio quality, the audio files of this experiment are available online at *https://www.audiolabs-erlangen.de/resources/2017-IEEE-BSS-tracking*.

Spectrograms of signal segments from the mixtures, the clean speech signals, and the separated signals at the output of the proposed BSS system are illustrated in Fig. 4 for two out of the three sources in the measured Setup 1.

VII. Conclusion

The main objective of this paper was the application of informed spatial filters to the problem of blind source separation of moving sources. The major challenge in such a system is to accurately estimate the statistics of each source, which are required for the computation of optimal separation filters. It is well-known that for accurate statistics estimation, each TF bin needs to be associated to the dominant source at that bin. To solve the latter problem, we proposed a multi-source tracking framework based on a narrowband measurement model. Distinguishing properties of our system compared to existing systems are the unified treatment of speech uncertainty via augmented measurements and the formulation of the tracking as a hidden data problem which follows from the properties of the narrowband model. To achieve BSS, the measurement-to-source association from the tracker is used to estimate the signal statistics and compute spatial filters for BSS. Evaluation of the BSS performance in different acoustic conditions demonstrated the advantages of the proposed system compared to a similar state-of-the-art BSS framework where the TF bin-to-source association is done using online clustering of narrowband DOA estimates. In addition, an improved objective quality of the separated source signals was obtained compared to a powerful state-of-the-art BSS approach based on independent vector analysis.

Appendix
Maximization of (30) With Respect to $\boldsymbol{x}_{\tau j}$

As the maximization is done for each source independently, we omit the source index. First, we substitute the Gaussian distributions given by (25a). Rather than maximizing (30), we can also minimize the cost function

$$\mathcal{J}(\boldsymbol{x}_\tau) = (\boldsymbol{x}_\tau - \hat{\boldsymbol{x}}_{\tau-1})^{\mathrm{T}} \boldsymbol{P}_\tau^{-1} (\boldsymbol{x}_\tau - \hat{\boldsymbol{x}}_{\tau-1}) + \sum_k \beta_k (\boldsymbol{r}_{\tau k} - \boldsymbol{x}_\tau)^{\mathrm{T}} \boldsymbol{\Sigma}_\tau^{-1} (\boldsymbol{r}_{\tau k} - \boldsymbol{x}_\tau). \quad (47)$$

Setting the derivative of $\mathcal{J}(\boldsymbol{x}_\tau)$ to zero we obtain

$$\frac{d\mathcal{J}}{d\boldsymbol{x}_\tau} \propto \boldsymbol{P}_\tau^{-1}(\boldsymbol{x}_\tau - \hat{\boldsymbol{x}}_{\tau-1}) - \sum_k \beta_k \boldsymbol{\Sigma}_\tau^{-1}(\boldsymbol{r}_{\tau k} - \boldsymbol{x}_\tau) \stackrel{!}{=} 0, \quad (48)$$

which can be easily rearranged and solved for $\boldsymbol{x}_\tau$

$$\boldsymbol{x}_\tau = \left(\boldsymbol{P}_\tau^{-1} + \sum_k \beta_k \boldsymbol{\Sigma}_\tau^{-1} \right)^{-1} \left(\boldsymbol{P}_\tau^{-1} \boldsymbol{x}_{\tau-1} + \boldsymbol{\Sigma}_\tau^{-1} \sum_k \beta_k \boldsymbol{r}_k \right).$$

To arrive to the more insightful formula for $\boldsymbol{x}_\tau$, we use the definitions in (33) and rewrite the solution above as follows

$$\boldsymbol{x}_\tau = \left(\boldsymbol{P}_\tau^{-1} + \widetilde{\boldsymbol{\Sigma}}_\tau^{-1} \right)^{-1} \left(\boldsymbol{P}_\tau^{-1} \boldsymbol{x}_{\tau-1} + \widetilde{\boldsymbol{\Sigma}}_\tau^{-1} \tilde{\boldsymbol{r}}_k \right). \quad (49)$$

Next, we invoke one of Searle's matrix identities [40]

$$(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} = \boldsymbol{A}(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{B}, \quad (50)$$

and by substituting $\boldsymbol{P}_\tau^{-1}$ and $\widetilde{\boldsymbol{\Sigma}}_\tau^{-1}$ for $\boldsymbol{A}$ and $\boldsymbol{B}$, (49) can be rewritten as

$$\boldsymbol{x}_\tau = \boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} \widetilde{\boldsymbol{\Sigma}}_\tau \boldsymbol{P}_\tau^{-1} \hat{\boldsymbol{x}}_{\tau-1} + \boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} \tilde{\boldsymbol{r}}_\tau.$$

Next, we add and subtract $\boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} \boldsymbol{x}_{\tau-1}$ on the right-hand side which allows us to write

$$\boldsymbol{x}_\tau = \boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} (\boldsymbol{I} + \widetilde{\boldsymbol{\Sigma}}_\tau \boldsymbol{P}_\tau^{-1}) \hat{\boldsymbol{x}}_{\tau-1} + \boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} (\tilde{\boldsymbol{r}}_\tau - \hat{\boldsymbol{x}}_{\tau-1}). \quad (51)$$

Finally, using basic matrix identities we can write

$$\boldsymbol{I} + \widetilde{\boldsymbol{\Sigma}}_\tau \boldsymbol{P}_\tau^{-1} = \boldsymbol{P}_\tau^{-1} (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau) = [\boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1}]^{-1}, \quad (52)$$

and by substituting (52) in (51) we obtain the desired result

$$\boldsymbol{x}_\tau = \hat{\boldsymbol{x}}_{\tau-1} + \boldsymbol{P}_\tau (\boldsymbol{P}_\tau + \widetilde{\boldsymbol{\Sigma}}_\tau)^{-1} (\tilde{\boldsymbol{r}}_\tau - \hat{\boldsymbol{x}}_{\tau-1}). \quad (53)$$

Acknowledgment

The authors would like to thank Prof. Dr. N. Ono from the National Institute of Informatics, Tokyo, Japan, for processing our data using their framework for online auxiliary function based independent vector analysis.

References

[1] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. New York, NY, USA: Springer, 2007.
[2] Ò. Yilmaz and S. Rickard, "Blind separation of speech mixture via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[3] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1913–1928, Sep. 2013.
[4] D. H. Tran Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2010, pp. 1–4.
[5] M. Taseska and E. A. P. Habets, "Informed spatial filtering with distributed arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1195–1207, Jul. 2014.
[6] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 385–389.
[7] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Dec. 2015, pp. 436–443.
[8] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, May 2011.
[9] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1900–1912, Sep. 2011.
[10] B. Loesch and B. Yang, "Online blind source separation based on time-frequency sparseness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2009, pp. 117–120.
[11] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 354–367, Feb. 2014.
[12] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1692–1703, Oct. 2015.
[13] M. Taseska and E. A. P. Habets, "An online EM algorithm for source extraction using distributed microphone arrays," in *Proc. Eur. Signal Process. Conf.*, Marrakech, Morocco, Sep. 2013, pp. 1–5.
[14] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2233–2243, Dec. 2014.
[15] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. Storrs, CT, USA: YBS, 1995.

[16] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Multimodal Technologies for Perception of Humans (Volume 4122 of Lecture Notes in Computer Science)*, R. Stiefelhagen and J. Garofolo, Eds. pp. 137–150. Berlin, Germany: Springer, 2007.
[17] F. C. Fallon and J. S. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, May 2012.
[18] M. Souden, J. Chen, J. Benesty, and S. Affès, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.
[19] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sep. 2014, pp. 183–187.
[20] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, Feb. 2012.
[21] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 661–676, May 2011.
[22] V. M. Tavakoli, J. R. Jensen, M. G. Christensen, and J. Benesty, "A framework for speech enhancement with ad hoc microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1038–1051, Jun. 2016.
[23] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
[24] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[25] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
[26] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 544–548.
[27] Y. Bar-Shalom, X. Tian, and P. K. Willett, *Multitarget-Multisensor Tracking: Principles and Techniques*. Bradford, U.K.: Yaakov Bar-Shalom, 2011.
[28] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. V-33–V-36.
[29] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, Sep. 2012, pp. 1–4.
[30] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.
[31] R. L. Streit and T. E. Luginbuhl, "Probabilistic multi-hypothesis tracking," Naval Undersea Warfare Center Division, Newport, RI, USA, NUWC-NPT Tech. Rep. 10,428, Feb. 1995.
[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[33] P. Willett, Y. Ruan, and R. Streit, "PMHT: Problems and some solutions," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 38, no. 3, pp. 738–754, Jul. 2002.
[34] E. A. P. Habets, "MATLAB implementation for: Signals of moving sources captured at microphones," Available. [Online]: https://github.com/ehabets/Signal-Generator
[35] 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.
[36] 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.
[37] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[38] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T, Geneva, Switzerland, 2001.
[39] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, May 2014.
[40] S. R. Searle, *Matrix Algebra Useful for Statistics*. New York, NY, USA: Wiley, 1982.

Maja Taseska (S'13) received the B.Sc. degree in electrical engineering from the Jacobs University, Bremen, Germany, and the M.Sc. degree from the Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany in 2010 and 2012, respectively. She then joined the International Audio Laboratories Erlangen, as a Ph.D. Candidate and Researcher in the field of informed spatial filtering for speech enhancement. Her research interests include microphone array signal processing, optimal filtering, source localization, tracking, and separation, and multimodal data analysis.

Emanuël A. P. Habets (S'02–M'07–SM'11) received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, The Netherlands, and the M.Sc. and Ph.D. degrees in electrical engineering from Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 1999, 2002, and 2007, respectively.

From 2007 to 2009, he was a Postdoctoral Fellow with the Technion–Israel Institute of Technology and with the Bar-Ilan University, Ramat Gan, Israel. From 2009 to 2010, he was a Research Fellow in the Communication and Signal Processing Group, Imperial College London, London, U.K. He is currently an Associate Professor with the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS), Erlangen, Germany, and the Head of the Spatial Audio Research Group, Fraunhofer IIS, Erlangen, Germany. His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, and echo reduction), and sound localization and tracking.

Dr. Habets was a Member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control in Eindhoven, The Netherlands, the General Cochair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics in New Paltz, New York, USA, and the General Cochair of the 2014 International Conference on Spatial Audio in Erlangen, Germany. He was a Member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013–2015), a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and the EURASIP *Journal on Advances in Signal Processing*, and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (2013–2017). He was the recipient, with S. Gannot and I. Cohen, of the 2014 IEEE Signal Processing Letters Best Paper Award. He is currently a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the Vice-Chair of the EURASIP Special Area Team on Acoustic, Sound, and Music Signal Processing, and the Editor-in-Chief for the EURASIP *Journal on Audio, Speech, and Music Processing*.