Blind Separation for Multiple Moving Sources With Labeled Random Finite Sets

Jonah Ong[®], Ba Tuong Vo[®], and Sven Nordholm[®]

Abstract—This paper proposes a novel solution for separating an unknown and time-varying number of moving acoustic sources in a blind setting using multiple microphone arrays. A standard steered-response power phase transform method is applied to extract source position measurements, which inevitably contain noise, false detections, missed detections, and are not labeled with the source identities. The imperfect measurements lead to the space-time permutation problem, as there is no information on how the measurements are associated to the sources in space, nor how the measurements are connected across time, if at all. To solve this problem, a labeled random finite set tracking framework is adopted to jointly estimate the source positions and their labels or identities. Based on these trajectory estimates, a corresponding set of time-varying generalized side-lobe cancellers is constructed to perform source separation. The overall algorithm operates in a block-wise or an online fashion and is scalable with the number of microphone arrays. The quality of the measurements, tracking, and separation, are evaluated respectively, with the OSPA metric, OSPA⁽²⁾ metric, and ITU-T P.835 based listening tests, on both real-world and simulated data.

Index Terms—Blind source separation, multi-object tracking, labeled random finite sets, acoustic localization, spatial filtering.

I. INTRODUCTION

I NMICROPHONE array processing, blind source separation (BSS) is the estimation of source signals, using only the received mixture signals with no information about the original sources and the mixing process [1]. In a realistic auditory scene, one of the main challenges for separating a mixture of concurrent sources is not only that the sources are moving, but also that the number of sources is unknown and time-varying, i.e. new sources can appear and existing sources can disappear or undergo silence periods. For static sources, established solutions to BSS include independent component analysis (ICA) [2], sparseness-based approaches [3], [4], and non-negative matrix factorization (NMF) [5]. These methods can be extended for moving sources by using a block-wise approach wherein moving

The authors are with the Deptartment of Electrical, and Computer Engineering, Curtin University, Bentley, WA 6102, Australia (e-mail: jonahosx25@gmail.com; ba-tuong.vo@curtin.edu.au; s.nordholm @curtin.edu.au).

This article has supplementary downloadable material available at https://doi. org/10.1109/TASLP.2021.3087003, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2021.3087003

sources are assumed to be static within a short time block [6], [7].

An alternative and a more recent block-wise approach is based on tracking of multiple moving sources, and followed by spatial filtering for extracting the signal-of-interest (SOI) from the estimated position/direction at each time [7]–[10]. One of the main difficulties in tracking an unknown number of sources in a reverberant environment is that acoustic localization measurements are subject to noise and false positives or negative, i.e. spurious or missing measurements. Moreover, the more pertinent issue is the space-time permutation problem. As in space, it is not known which measurements are connected to which sources, and in time, it is not known how the measurements are connected across time frames with respect to the sources. Furthermore, the solution must cater for possible appearance of new sources, movement of active or inactive sources, and disappearance of existing sources.

Classical dynamic Bayesian estimation techniques such as the particle filter have been applied to single source tracking in [11]–[13]. For multiple sources, there is uncertainty not only in the source position, but also in the number of sources, and the latter is not accounted for within the classical Bayesian framework [14]. Recent solutions for addressing multiple sources have relied on adaptations of the Rao-Blackwellised Particle Filter (RBPF) [7], [15], the Probabilistic Multiple Hypothesis Tracker (PMHT) [9], and the Joint Probabilistic Data Association (JPDA) filter [16]. The newer RFS framework based on Finite Set Statistics (FISST) [17], offers a principled mechanism to cater for an unknown and time-varying number of sources in a Bayesian setting, and is directly applicable to acoustic tracking [14]. The first RFS based solution for multi-source acoustic tracking was proposed in [18]. Subsequent RFS-based solutions have been proposed for multi-source acoustic tracking with the Probability Hypothesis Density (PHD) filter [8], [19]-[21], the Cardinalized PHD filter [22], the Cardinality-Balanced Multi-Target Multi-Bernoulli filter [23], and the RFS Particle Filter [24].

However, these above methods do not directly estimate source tracks, which are source position estimates associated with a common label. Consequently, they require a post-processing step such as track management to resolve each track individually. These methods are thus suboptimal in the sense that they solve the space-time permutation problem separately. As the spatial filtering module relies on accurate label or identity estimates, the presence of labeling errors results in switching in the separated signal estimates. Solving the space-time permutation problem

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Manuscript received December 16, 2020; revised April 14, 2021; accepted May 31, 2021. Date of publication June 7, 2021; date of current version June 21, 2021. This work was supported by Australian Research Council under Grant DP170104854. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W H Khong (*Corresponding author: Jonah Ong.*)

jointly has the potential to significantly improve tracking performance and hence separation performance. Furthermore, the above mentioned approaches do not scale linearly in the number of arrays used in the system, thereby making them impractical for online implementation when the number of arrays is large.

In this paper, we propose a novel online solution for multiarray BSS with an unknown time-varying number of moving sources in a 3D auditory scene. Our solution follows the approach of first obtaining position measurements, then tracking of multiple sources, and finally separation using spatial filtering, all in an online or block-wise fashion. Source position measurements obtained through Steered-Response Power Phase Transform (SRP-PHAT) [25] exhibit the space-time permutation issue, where it is not known which measurement (if any) is connected to which source at the current time, nor which measurements are connected to the same source across time. This work is the first to formally address the space-time permutation problem, using a labeled random finite set (RFS) approach [26]-[28] to jointly estimate the number of sources, their positions and their labels. The solution invokes the Multi-Sensor Generalized Labeled Multi-Bernoulli (MS-GLMB) tracker [29], which is a tractable linear complexity recursive filter for estimating the source trajectories from raw measurements. The tracking estimates at each frame are used to construct a set of time-varying beamformers, known as the Generalized Side-lobe Canceller (GSC) [30], which are used for multi-source separation. The proposed method is evaluated using real recordings and under different reverberation times via simulation. We use the Optimal Sub-Pattern Assignment (OSPA) which is a metric for two sets of points [31], to evaluate the quality of the array measurements. The tracking performance is evaluated using a variant of the OSPA metric called the $OSPA^{(2)}$, which is a proper metric for two sets of tracks [32]. Finally, we evaluate the separation performance via subjective listening tests according to the ITU-T P.835 methodology [33].

II. PROBLEM FORMULATION AND SOLUTION OVERVIEW

One of the main challenges in BSS for multiple moving sources is the inherent space-time permutation problem, since acoustic localization techniques are generally unable to identify and produce exactly one measurement for each source. It is then necessary to estimate the trajectory of each source from the measurements, which entails knowing when a source appears or enters the scene, disappears or exits the scene, and how its position changes over each time instance. This is effectively an online tracking problem where the objective is to estimate, at each time instance, the number of sources, their positions and unique labels. Knowledge of the correct source positions and their labels is crucial, as it resolves the inherent space-time permutation problem, thereby enabling the application of a set of time-varying spatial filters to achieve source separation. The underlying signal model and overview of the proposed solution are given below.

A. Signal Model

We consider a scenario consisting of N(t) point sources where each source is indexed by $n \in \{1, ..., N(t)\}$ with 3D position denoted by $\alpha_n(t) \in \mathbb{R}^3$ at discrete time instance t. Each source signal is denoted by s_n , and all sources are assumed to be mutually uncorrelated, i.e. the cross power spectral density between two sources is zero. An array indexed by $q \in \{1, ..., Q\}$, comprises M_q microphone elements. The source signals impinge on each microphone element $m \in \{1, ..., M_q\}$ of array q, and are corrupted with non-directional diffuse noise $v^{(q,m)}$. The mixture signal at microphone (q, m) is represented by some mapping function ρ of the source signals $s_1, ..., s_{N(t)}$, source positions $\alpha_1, ..., \alpha_{N(t)}$, and noise $v^{(q,m)}$, evaluated at time t:

$$y^{(q,m)}(t) = \rho\left(s_1, \dots, s_{N(t)}, \alpha_1, \dots, \alpha_{N(t)}, v^{(q,m)}\right)(t).$$
(1)

For stationary sources in an invariant and homogeneous acoustic environment, the mixture signal can be modeled via the sum of the convolutions of source signals and the room impulse response (RIR), which encapsulates the direct path (time-delay) and multipath terms (reflections) between the sources and microphone element (q, m) [34], [35]. However when sources are moving, the effective RIR becomes time-varying. To circumvent this issue, we consider the source signal in blocks of frames:

$$s_n(t) = \sum_{k=1}^{K} s_n(t) w_T \left(t - (k-1)T \right) = \sum_{k=1}^{K} s_{k,n}(t), \quad (2)$$

where w_T is a window function of length T, and k is the index of a time block/frame with length T. Specifically, we assume source stationarity at each frame k of length T, i.e. $\alpha_n(t) = \alpha_{k,n}$ and $N(t) = N_k$ for $t = (k - 1)T, \ldots, kT$. Thus, the signal is filtered by a new RIR for each time frame:

$$y^{(q,m)}(t) \approx \sum_{k=1}^{K} \sum_{n=1}^{N_k} (s_{k,n} * h^{(q,m)}_{k,\alpha_{k,n}})(t) + v^{(q,m)}(t), \quad (3)$$

where * denotes convolution, and $h_{k,\alpha_{k,n}}^{(q,m)}$ denotes the RIR between source n with position $\alpha_{k,n}$ and microphone element (q,m), at frame k. From this representation, each source signal is assumed to be a point source (in a fixed position) in frame k, which is filtered by a linear time-invariant system, where the time invariance is assumed over the block at length T. For tractability reasons, we focus only on the direct path term and approximate the mixture signal as:

$$y^{(q,m)}(t) \approx \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{s_{k,n} \left(t - \tau(\alpha_{k,n}, u^{(q,m)}) \right)}{4\pi ||\alpha_{k,n} - u^{(q,m)}||} + v^{(q,m)}(t),$$
(4)

where $||\cdot||$ is the Euclidean distance, $\tau(\alpha_{k,n}, u^{(q,m)}) \triangleq c^{-1}||\alpha_{k,n} - u^{(q,m)}||$ is the time delay between source n at position $\alpha_{k,n}$ and microphone (q,m) at position $u^{(q,m)} \in \mathbb{R}^3$ (c is the speed of sound). Based on this model, the objective is to estimate the individual source signals for every frame k (frame by frame) using only the mixture signals $y^{(1,1)}, \ldots, y^{(Q,M_Q)}$ and no prior knowledge on the sources.



Fig. 1. Processing Chain for the Proposed Method.

B. Overview of the Proposed Method

The processing chain of the proposed method is depicted in Fig. 1. Raw microphone signals are segmented into frames and transformed into the frequency domain. Then, acoustic localization techniques that rely on source features such as direction-of-arrivals (DOAs), are used to acquire the source position candidates at each frame. The position candidates from each array are subjected to noise (disturbance), they may not reflect a source that is present (false negative), and some may not correspond to any source (false positive). Above all, there is a space-time permutation problem because the acquired position candidates from each array are unidentified (without labels) across time. As a result, there is no trajectory information on the sources, and spatial filtering cannot be applied for source separation. To remedy this, spatial distributions of the position candidates from all arrays are exploited to jointly estimate the number of sources, their positions and labels for each frame. The estimation of the source labels is important because it resolves the permutation ambiguity. Based on this information, a series of time-varying spatial filter can be constructed using the direct path model for source separation. The proposed method can be broken down into 3 stages: signal pre-processing, multi-source tracking and source separation.

In the first stage, raw microphone signals $y^{(1,1)}, \ldots, y^{(Q,M_Q)}$ from all arrays are pre-processed into frames of data in the frequency domain using the short-time Fourier transform (STFT). For each frame, we use the Steered-Response Power Phase Transform (SRP-PHAT) [25], and apply a region search algorithm known as Stochastic Region Contraction (SRC) proposed in [25], to obtain 3D position candidates from each array. Due to noise, false positives, false negatives, and the space-time permutation problem, the obtained source position candidates from all arrays are not fit for spatial filtering to achieve source separation.

In the second stage, we employ a Bayesian state estimation framework that processes the obtained position candidates from all arrays, herein referred to as the multi-array measurements, and produces estimates of the source positions and labels at each frame. The tracking filter works by recursively propagating a posterior density which characterizes the uncertainty of a set of labeled states given all multi-array measurements up to the current time. This framework accounts for noise, false positives and false negatives in the multi-array measurements. Source labels, motions, appearances and disappearances are also incorporated into the formulation. The joint estimation of the source labels and positions resolves the space-time permutation problem. In the third stage, source separation is achieved via constructing a type of spatial filter known as the Generalized Side-lobe Canceller (GSC) for each frame. The GSC aims to emphasize and separate the source of interest while actively cancelling interfering sources. In order to do this, it is necessary to have the estimated source positions and the labels at each frame, which is provided by the proposed tracking solution. In addition, we utilize the GSC signals to construct a time-frequency mask for enhancing the separated signals. Finally, the time-domain separated signals are recovered using the inverse STFT.

III. SIGNAL PRE-PROCESSING

This section describes the segmentation of raw signals into frames of data using the short-time Fourier Transform (STFT), followed by the use of Steered-Response Power Phase Transform (SRP-PHAT) combined with Stochastic Region Contraction (SRC) to obtain the 3D source position candidates. The shortcomings of the obtained position candidates are outlined and discussed.

A. Short-Time Fourier Transform (STFT)

Each raw microphone signal $y^{(q,m)}$ is segmented into $y_1^{(q,m)}, \ldots, y_K^{(q,m)}$ via:

$$y_k^{(q,m)}(t) = y^{(q,m)}(t + (k-1)T)w_T(t),$$
(5)

where w_T is a selected window function of length T. The window function is chosen such that it captures enough information while reducing signal discontinuities at the edges, e.g. a Hann window $w_T(t) = 0.5 - 0.5\cos(2\pi t/T)$, $t = 0, \ldots, T-1$. We denote the discrete short-time Fourier transform of $y_k^{(q,m)}(t)$ by $Y_k^{(q,m)}(\lambda)$ where λ is the frequency bin index. To represent the segmented frequency-domain raw signals from all microphones at array q in a compact form, we stack them into a vector:

$$Y_k^{(q)}(\lambda) = \left[Y_k^{(q,i)}(\lambda)\right]_{i=1}^{M_q} \tag{6}$$

B. Steered-Response Power Phase Transform (SRP-PHAT)

Steered-Response Power Phase Transform (SRP-PHAT) is an acoustic source localization solution well known for its robust performance in adverse acoustic environments [36]. The SRP is the output power of a delay-and-sum beamformer that is steered to a set of source positions which are defined under a specified spatial grid [25]. The Phase Transform (PHAT) is a weighting technique to avoid peak spreading in the SRP by



Fig. 2. SRP-PHAT Measurements.

emphasizing the phase information of the involved signals [25]. Given $Y_k^{(q)}$ received at array q, the spatial power that emanates from the direction of the source location $\alpha_k \in \mathbb{R}^3$ at each frame k, is computed using Steered-Response Power (SRP) with PHAT by [25]:

$$\mathcal{P}_{k}^{(q)}(\alpha) = \sum_{a=1}^{M_{q}-1} \sum_{b=a+1}^{M_{q}} \sum_{\lambda} \frac{Y_{k}^{(q,a)}(\lambda)Y_{k}^{*(q,b)}(\lambda)}{\left|Y_{k}^{(q,a)}(\lambda)Y_{k}^{*(q,b)}(\lambda)\right|} \times e^{j\omega_{\lambda}\left(\tau(\alpha, u^{(q,b)}) - \tau(\alpha, u^{(q,a)})\right)},$$
(7)

where $\omega_{\lambda} = 2\pi(\lambda - 1)F_s/T$ (F_s is the sampling frequency), the PHAT weighting is the inverse magnitude of the frequency components of the involved signals, and the exponential term is responsible for time-aligning the microphone signals based on time-difference-of-arrival. Searching for multiple local maxima of (9) at any frame k corresponds to source position candidates that are present at that time frame. However, this process is computationally expensive as it involves a large search space.

C. Stochastic Region Contraction (SRC)

Using the computationally efficient SRC algorithm [25], the 3D source position candidates are obtained via peak-picking SRP-PHAT for every array with a certain threshold. For each array q, we denote the collection of the position candidates as a measurement set:

$$Z_{k}^{(q)} = \{z_{k,1}^{(q)}, \dots, z_{k,|Z_{k}^{(q)}|}^{(q)}\},\tag{8}$$

where $|Z_k^{(q)}|$ denotes the number of measurements (see Fig. 2). For multiple arrays, we define $Z_k \triangleq (Z_k^{(1)}, \ldots, Z_k^{(Q)})$ as the multi-array measurements. The multi-array measurements are utilized to deduce the optimal positions of the sources. However, due to nonlinearity, noise and reverberation (in real-world conditions), the multi-array measurements have the following issues:

- A measurement $z_k^{(q)}$ obtained from a single array (if it is generated by a source) is noisy after undergoing a highly nonlinear transformation.
- The multi-array measurements contain false positives, which are measurements not generated by any active source; and false negatives, which are missing measurements even when sources are active.
- Furthermore, we are faced with the inherent space-time permutation problem as the multi-array measurements are unordered and have no identities/labels. Specifically, in space, it is not known which individual measurement in

the sets $Z_k^{(1)}, \ldots, Z_k^{(Q)}$ is generated by which source. In time, it is not known how an individual measurement from the sets $Z_k^{(1)}, \ldots, Z_k^{(Q)}$ at the current frame, to the sets $Z_{k+1}^{(1)}, \ldots, Z_{k+1}^{(Q)}$ at the next frame, is connected with respect to an existing source. Also, the appearance of a new active source or the disappearance of an existing active source is unknown.

IV. TRACKING OF MULTIPLE SOURCES

This section presents a labeled RFS solution for estimating the source trajectories from the source measurements thereby addressing the space-time permutation problem. The solution entails the recursive multi-source Bayes filter, which requires specification of the multi-source transition and multi-array likelihood models. A tractable implementation is given in the form of the Multi-Sensor Generalized Labeled Multi-Bernoulli filter. These are summarized as follows.

A. Multi-Source Bayesian Tracking Filter

Given the multi-array measurements $Z_k \triangleq (Z_k^{(1)}, \dots, Z_k^{(Q)})$, the objective is to estimate the number of the sources, their positions and labels at each frame k. In order to do so, it is necessary to have a stochastic model to characterize the timevarying nature of the number of sources and the individual source positions, which arises due to source appearance, disappearance and physical motion. Similarly, it is necessary to have a stochastic model to characterize the multi-array measurements as the number of measurements for each array is also time-varying, partly because the number of sources is time-varying, but also because the measurements are subjected to noise, false negatives and false positives.

A random finite set (RFS) is a natural representation for the collection of source positions (with labels), and for each of the array measurements, because an RFS is essentially a set-valued random variable, wherein the number of points as well as the values of individual points are random [17], [18], [26]. In order to develop an online solution for estimating the number of sources, their positions and labels based on RFS modeling for each frame, we cast the problem into a recursive Bayesian estimation framework.

In the context of this framework, source appearances and disappearances are referred to as source births and deaths respectively, while false negatives and false positives are referred to as missed detections and false detections respectively. Recall that the time permutation problem arises due to source motions, appearances and disappearances, while the space permutation problem arises due to the absence of labels in the array measurements, which are also subjected to noise, missed and false detections. The space-time permutation problem is referred to as the data association problem and can be addressed using the RFS tracking framework. Fig. 3 gives an illustration of the array measurements prior to tracking as well as the desired result after tracking is applied.

Each source at frame k has a state denoted by $x_k \triangleq (x_k, \ell)$, where $x_k \triangleq (\alpha_k, \dot{\alpha}_k)$ is a vector capturing the 3D position and



Fig. 3. Two sources appear at frame k-1 and persist until frame k+1, while a third source appears at k and persists until k+1. (a) Illustration of measurements from two arrays, $Z^{(1)}$ and $Z^{(2)}$ (time subscript k is suppressed). (b) Illustration of desired tracking result to resolve the space-time permutation problem.

velocity of the source, and ℓ is a unique label from a discrete space \mathbb{L} . Note that the velocity component is an auxiliary variable needed for the state transition model in the Bayesian framework. The states of multiple sources at each frame k, are represented as a finite set:

$$X_k = \{x_{k,1}, \dots, x_{k,N_k}\},$$
 (9)

herein referred to as a multi-source state. Note that the existence of unique labels in the multi-source state means that consecutive states with the same label across frames constitute the trajectory of the source movement (see Fig. 3(b)).

The RFS representation of X_k naturally accounts for the movements of active sources, births of new sources and deaths of existing sources, while the RFS representation of the sets $Z_k^{(1)}, \ldots, Z_k^{(Q)}$ naturally accounts for, noise, missed detections, and false detections in the measurements across all arrays. In Bayesian RFS tracking, the aim is to estimate frame-by-frame (recursively) the multi-source state X_k , given the multi-array measurements obtained from the beginning of time up to the current time frame k, i.e. $Z_{1:k} \triangleq (Z_1, \ldots, Z_k)$. The solution is the *multi-object Bayes filter*, which is a recursive mechanism that computes the probability density of X_k given $Z_{1:k}$ [26]. In the context of Bayesian filtering, this probability density is referred as the filtering density denoted by $\pi_{k|k}(X_k|Z_{1:k})$. At any given frame k, all uncertainty in the multi-source state X_k given $Z_{1:k}$, is captured in $\pi_{k|k}(X_k|Z_{1:k})$ [26].

The propagation of the filtering density is a recursive two-step procedure. The first step is the time update of the current filtering density $\boldsymbol{\pi}_{k|k}(\boldsymbol{X}_k|Z_{1:k})$ via [26]:

$$\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k}) = \int \boldsymbol{f}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_{k}) \boldsymbol{\pi}_{k|k}(\boldsymbol{X}_{k}|Z_{1:k}) \delta \boldsymbol{X}_{k}, (10)$$

where the integral is not the usual Euclidean notion of integration, rather it is a set integral defined under Finite Set Statistics (FISST) for dealing with RFSs in a mathematically consistent manner [37], and $f(X_{k+1}|X_k)$ is known as the *multi-source* transition density which gives the probability density that multisource state X_k at frame k transitions to X_{k+1} at the next frame k + 1. The *multi-source transition density* is formulated based on a stochastic model that encapsulates all possible source births, deaths and motions, i.e. the time permutation aspect. The details of this transition model are further discussed in Section IV-B. Consequently, the time-updated density (13) characterizes the transition of X_k to X_{k+1} , given all multi-array measurements $Z_{1:k}$ up to the current time frame, and addresses the time permutation part of the data association problem. The second step is the data update of $\pi_{k+1|k}(X_{k+1}|Z_{1:k})$ with the multi-array measurements Z_{k+1} obtained at frame k + 1 via [26]:

$$\pi_{k+1|k+1}(\boldsymbol{X}_{k+1}|Z_{1:k+1}) = \frac{g(Z_{k+1}|\boldsymbol{X}_{k+1})\pi_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k})}{[g(Z_{k+1}|\boldsymbol{X}_{k+1})\pi_{k+1|k}(\boldsymbol{X}_{k+1}|Z_{1:k})\delta\boldsymbol{X}_{k+1}]}, \quad (11)$$

where $g(Z_{k+1}|X_{k+1})$ is known as the *multi-array measurement likelihood* which gives the probability density of the multiarray measurements Z_{k+1} , given the multi-source state X_{k+1} . The *multi-array measurement likelihood* is formulated based on a stochastic model that encapsulates noise, detections, missed detections, false detections and association uncertainty, i.e. the space permutation aspect, in the obtained multi-array measurements. The details of this multi-array measurement model are given in Section IV-C. The data-updated density (14) contains all information about the number of sources and their states (with labels) at the next time frame k + 1, conditioned on the multi-array measurements up to that frame. This step consequently addresses the space permutation part of the data association problem.

In summary, the combination of both time-update and dataupdate steps in the propagation of the filtering density solves the space-time permutation problem. To obtain a multi-source state estimate at each frame, which contains the estimated number of sources, their positions and labels, a conventional Bayesian multi-source estimator is applied to the filtering density at each frame. The closed-form representation of the filtering density and the implementation of the filter, i.e. the tractable (recursive) propagation of the filtering density, are discussed in Section IV-D.

B. The Multi-Source Transition Model

The function $f(\cdot|\cdot)$ is a probability density function characterizing all possible source births, deaths and motions that take place in the transition of a multi-source state from one frame to the next [26]. The function $f(\cdot|\cdot)$ is parameterized as per Table II, and explanations of these parameters are given as follows.

Given the multi-source state X_k , each state $x_k \triangleq (x_k, \ell) \in X_k$ either survives with probability P_S and transition to a new

 TABLE I

 PARAMETERS FOR THE MULTI-SOURCE TRANSITION DENSITY (15)

Probability of survival	P_S
Single-source transition density	$f_S(\cdot \cdot)$
Probability of birth	$r_B(\cdot)$
Birth density	$p_B(\cdot)$

 TABLE II

 PARAMETERS FOR THE MULTI-ARRAY MEASUREMENT LIKELIHOOD (17)

Probability of detection	$P_D^{(1)},, P_D^{(Q)}$
Single-source likelihood	$g^{(1)}(\cdot \cdot),, g^{(Q)}(\cdot \cdot)$
False detection intensity	$\kappa^{(1)}(\cdot),, \kappa^{(Q)}(\cdot)$

state (x_{k+1}, ℓ) that inherits the same label whose uncertainty is captured by the transition density $f_S(x_{k+1}, \ell | x_k, \ell)$, or dies with probability $1-P_S$. At this next time, a set of new sources denoted by B_{k+1} with labels $\{\ell : (x_{k+1}, \ell) \in B_{k+1}\}$ can be born or appear individually with probability $r_B(\ell)$ and distributed according to the birth density $p_B(\cdot, \ell)$. Recall that labels of a multi-source state are distinct/unique for all frames, hence a label is defined as $\ell = (\varsigma, \iota) \in \mathbb{L}_k$, where $\varsigma \in \{k\}$ denotes the time frame of birth and $\iota \in \mathbb{N}$ denotes the index of source born at the same time [26] (see Fig. 3 (b) for illustration). Consequently, the label space for sources at frame k is constructed recursively by $\mathbb{L}_{0:k} = \mathbb{L}_{0:k-1} \cup \mathbb{L}_k$.

The multi-source state X_{k+1} is the superposition of the surviving sources W_{k+1} and the new born sources B_{k+1} , which are assumed to be statistically independent. Let $f_S(W_{k+1}|X_k)$ and $f_B(B_{k+1})$ be the probability densities of the survivability of X_k to W_{k+1} , and the new born sources B_{k+1} respectively, then the multi-source transition density is given by [26]:

$$\boldsymbol{f}(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k) = \boldsymbol{f}_S(\boldsymbol{W}_{k+1}|\boldsymbol{X}_k) \boldsymbol{f}_B(\boldsymbol{B}_{k+1}). \quad (12)$$

The product in (15) presents a model for addressing the time permutation problem. In particular, source appearance, disappearance and motion are considered to be statistically independent. However, labels are kept the same for sources that move and continue to be active, and appearing active sources are assigned a new distinct label, while deactivated sources are removed. The derivation of (15) is beyond the scope of this paper, but interested readers are referred to [26].

C. The Multi-Array Measurement Likelihood Model

The function $g(\cdot|\cdot)$ is a probability density function characterizing noise, missed detections, false detections and association uncertainty in the multi-array measurements. The function $g(\cdot|\cdot)$ is parameterized as per Table III, and explanations of these parameters are given as follows.

Given the multi-source state X_k , each $x_k = (x_k, \ell_k) \in X_k$ is either detected at array q with probability $P_D^{(q)}$ and generates a detection $z_k^{(q)} \in Z_k^{(q)}$ with likelihood $g^{(q)}(z_k^{(q)}|x_k, \ell_k)$, or missed detected with probability $1 - P_D^{(q)}$. The detection process also generates false detections at array q, conventionally characterized by an intensity function $\kappa^{(q)}(\cdot) \triangleq \lambda_{FD}^{(q)} \mathcal{U}(\cdot)$ on the measurement space [17], [26]. The number of false detections is modeled by a Poisson distribution with mean $\lambda_{FD}^{(q)}$, and the false detections themselves are uniformly distributed in the measurement space according to $\mathcal{U}(\cdot)$. In standard multi-source tracking, it is standard to assume that the detections are statistically independent from the false detections [26].

A single-array association $\theta_k^{(q)} \in \Theta_k^{(q)}$ is defined as a mapping from the source labels to the measurement indexes, i.e. $\theta_k^{(q)} : \{\ell_k : (x_k, \ell_k) \in \mathbf{X}_k\} \rightarrow \{0 : |Z_k^{(q)}|\}$. Note that $\Theta_k^{(q)}$ is the space of all mappings, such that *no two distinct arguments are mapped to the same positive value* [26]. This property ensures each detection comes from at most one source. For example, $\theta_k^{(q)}(\ell_k) > 0$ corresponds to source ℓ_k generating detection $z_{k,\theta_k^{(q)}(\ell_k)}^{(q)}$ at array q, while $\theta_k^{(q)}(\ell_k) = 0$ means source ℓ_k is misdetected at array q. For multiple arrays, a multi-array association is the vector $\theta_k \triangleq (\theta_k^{(1)}, \dots, \theta_k^{(Q)}) \in \Theta_k$ of all single-array associations having the same aforementioned positive one-to-one property, where $\Theta_k \triangleq \Theta_k^{(1)} \times \ldots \times \Theta_k^{(Q)}$ is the space of all possible multiarray associations [29].

Under the assumption that the set $Z_k^{(q)}$ at array q is conditionally independent from those at other arrays, the *multi-array* measurement likelihood is given by [29]:

$$g(Z_{k}|\mathbf{X}_{k}) \propto \sum_{\theta_{k}^{(1)} \in \Theta_{k}^{(1)}} \dots \sum_{\theta_{k}^{(Q)} \in \Theta_{k}^{(Q)}} \prod_{\substack{(x_{k},\ell_{k}) \\ \in \mathbf{X}_{k}}} \prod_{q=1}^{Q} \psi_{Z_{k}^{(q)}}^{(q,\theta_{k}^{(q)}(\ell_{k}))}(x_{k},\ell_{k}),$$
(13)

where

$$\psi_{Z_{k}^{(q,j)}}^{(q,j)}(x_{k},\ell_{k}) = \begin{cases} \frac{P_{D}^{(q)}g^{(q)}\left(z_{k,j}^{(q)}|x_{k},\ell_{k}\right)}{\kappa^{(q)}\left(z_{k,j}^{(q)}\right)}, & j > 0\\ 1 - P_{D}^{(q)}, & j = 0 \end{cases}$$
(14)

It is important to note that the nested sum in (17) indicates the enumeration of all possible multi-array associations, thereby taking into account all possible combinations of missed detections, false detections and the source detections. In other words, the nested sum in (17) presents a model for addressing the space permutation problem by considering all possible mappings of position candidates to source labels. The derivation for (17) is beyond the scope of this paper, but interested readers are referred to [17], [26], [29].

D. The Multi-Sensor Generalized Labeled Multi-Bernoulli (MS-GLMB)

Under the transition and measurement models as described above, the time-updated and data-updated (filtering) densities admit a closed-form solution via the Generalized Labeled Multi-Bernoulli (GLMB) density [26], [27], [29]:

$$\boldsymbol{\pi}(\boldsymbol{X}_k) = \Delta(\boldsymbol{X}_k) \sum_{\theta_{1:k} \in \Theta_{1:k}} \omega^{(\theta_{1:k})}(\mathcal{L}(\boldsymbol{X}_k)) \prod_{\boldsymbol{x}_k \in \boldsymbol{X}_k} p^{(\theta_{1:k})}(\boldsymbol{x}_k),$$
(15)

where $\mathcal{L}(\mathbf{X}_k) \triangleq \{\ell: (x_k, \ell) \in \mathbf{X}_k\}, \Delta(\cdot)$ is a distinct label indicator, i.e. $\Delta(\mathbf{X}_k) = 1$ if and only if the cardinality $|\mathcal{L}(\mathbf{X}_k)| = |\mathbf{X}_k|, \theta_{1:k} \in \Theta_{1:k}$ is the history of multi-array association mappings up to frame k, i.e. $\theta_{1:k} \triangleq (\theta_1, \dots, \theta_k)$. Each $\omega^{(\theta_{1:k})}(\mathcal{L}(\boldsymbol{X}_{k}))$ is a non-negative weight such that

$$\sum_{L \subseteq \mathbb{L}_{0:k}} \sum_{\theta_{1:k} \in \Theta_{1:k}} \omega^{(\theta_{1:k})}(L) = 1,$$
(16)

and can be interpreted as the probability of sources with label set $\mathcal{L}(\mathbf{X}_k)$ being active, as well as being associated with the detections given by the association history $\theta_{1:k}$. Each $p^{(\theta_{1:k})}(\cdot, \ell)$ is a probability density of the source state with label ℓ and association history $\theta_{1:k}$, where $p^{(\theta_{1:k})}(x_k, \ell)$ is the probability density of the source with label ℓ being located at state $x_k = (\alpha_k, \dot{\alpha}_k)$.

In plain terms, the GLMB (20) can be interpreted as a mixture model, i.e. a weighted sum of the products of single-source probability densities, where each weight is a function of the labels in the multi-source state. From an implementation standpoint, the number of terms in the mixture grows exponentially over time, partly due to the enumeration of all possible multiarray associations at each time frame. To maintain tractability, pruning of the terms with low weights is required, and has been shown to minimize the L_1 approximation error [29]. The Multi-Sensor GLMB (MS-GLMB) filter offers a polynomial time implementation mechanism that generates highly weighted components without exhaustive enumeration of the sum in (20), which has a linear complexity in the sum of the total number of measurements across all arrays [29]. A multi-source state estimate can be obtained from the GLMB posterior density via a simple GLMB estimator [27], [29]. Since we only require the position component of the single-source state, the estimated multi-source state \hat{X}_k at frame k is:

$$\hat{\boldsymbol{X}}_{k} = \{ (\hat{\alpha}_{k,1}, \hat{\ell}_{1}), \dots, (\hat{\alpha}_{k,|\hat{\boldsymbol{X}}_{k}|}, \hat{\ell}_{|\hat{\boldsymbol{X}}_{k}|}) \},$$
(17)

where $\hat{N}_k = |\hat{X}_k|$ is the estimated number of sources.

V. SOURCE SEPARATION

This section describes the use of the multi-source state estimate from the MS-GLMB filter to construct a Generalized Sidelobe Canceler (GSC) for source separation. For post-processing, we adopt a time-frequency masking step to further suppress interfering speech.

A. Spatial Filtering

At each frame k, the tracking filter provides the multi-source state estimate \hat{X}_k , which contains the estimated source positions and labels from the available data. The combination of source positions and labels constitutes the estimated source tracks, thereby solving the space-time permutation problem that arises from the multi-array measurements as depicted in Fig. 3. With this information, we design a set of spatial filters that is changing at each frame depending on \hat{X}_k , based on a free space near-field room model. We adopt a variant of the linearly constrained minimum variance beamformer called the Generalized Sidelobe Canceler (GSC). A GSC is a constrained beamformer that has been converted to a non-constrained design by means of a blocking matrix [30]. The GSC contains two parts: a beamformer that determines the response of the source of interest (SOI), and



Fig. 4. Spatial Filtering via Generalized Side-lobe Canceler (GSC).

a mechanism that blocks the SOI from entering the canceler. Fig. 4 shows a block diagram of the GSC.

In the first part, we use a beamformer that emphasizes the direction of the SOI specified by label $\hat{\ell}_i$ with position $\hat{\alpha}_{k,i}$, while nulling other interfering sources specified by $\{(\hat{\alpha}_{k,j}, \hat{\ell}_j) \in \hat{X}_k\}_{j=1}^{\hat{N}_k}$ for $i \neq j$. For each TF point (λ, k) , the weight of the beamformer $\hat{W}_{k,\hat{\ell}_i}^{(q)}(\lambda)$ is given by:

$$\begin{pmatrix} \mathbf{D}_{k,\hat{\mathbf{X}}_{k}}^{(q)}(\lambda) \end{pmatrix}^{H} \hat{W}_{k,\hat{\ell}_{i}}^{(q)}(\lambda) = r_{\hat{N}_{k}}(\hat{\ell}_{i})$$

$$\hat{W}_{k,\hat{\ell}_{i}}^{(q)}(\lambda) = \left(\left(\mathbf{D}_{k,\hat{\mathbf{X}}_{k}}^{(q)}(\lambda) \right)^{H} \right)^{\dagger} r_{\hat{N}_{k}}(\hat{\ell}_{i}), (18)$$

where the operator H is the Hermitian transpose, the dagger \dagger denotes the Moore-Penrose pseudo-inverse, $r_{\hat{N}_{k}}$ is a selection vector whose dimension varies depending on the estimated number of sources \hat{N}_{k} , i.e. $r_{\hat{N}_{k}}(\hat{\ell}_{i}) = [\delta_{\hat{\ell}_{1}}(\hat{\ell}_{i}), \ldots, \delta_{\hat{\ell}_{\hat{N}_{k}}}(\hat{\ell}_{i})]^{T}$ such that $\delta_{i}(j) = 1$ if i = j and zero otherwise, and

$$\mathbf{D}_{k,\hat{\mathbf{X}}_{k}}^{(q)}(\boldsymbol{\lambda}) = \begin{bmatrix} e^{j\omega_{\lambda}\left(\tau(\hat{\alpha}_{k,1}, u^{(q,1)})\right)} \cdots e^{j\omega_{\lambda}\left(\tau(\hat{\alpha}_{k,\hat{N}_{k}}, u^{(q,1)})\right)} \\ \vdots & \ddots & \vdots \\ e^{j\omega_{\lambda}\left(\tau(\hat{\alpha}_{k,1}, u^{(q,M_{q})})\right)} \cdot e^{j\omega_{\lambda}\left(\tau(\hat{\alpha}_{k,\hat{N}_{k}}, u^{(q,M_{q})})\right)} \end{bmatrix}, \quad (19)$$

is a matrix with columns representing the steering vectors for each estimated source. The number of columns depends on the estimated number of sources \hat{N}_k . Note that if $\hat{N}_k = 1$, (23) reduces to the classical delay-and-sum beamformer.

The second part involves a blocking matrix that is defined to be the orthogonal complement to $(\hat{W}_{k,\hat{\ell}_i}^{(q)}(\lambda))^H$ [30]:

where I is an identity matrix. Subsequently, the weight vector of the GSC is defined by:

$$G_{k,\hat{\ell}_i}^{(q)}(\lambda) = \hat{W}_{k,\hat{\ell}_i}^{(q)}(\lambda) - \mathcal{B}_{k,\hat{\ell}_i}^{(q)}(\lambda)V_k(\lambda),$$
(21)

where

$$V_{k,opt}(\lambda) = \arg\min_{V} \sum_{\eta=1}^{k} \gamma^{k-\eta} \left| \left(\hat{W}_{\eta,\hat{\ell}_{i}}^{(q)}(\lambda) - \mathcal{B}_{\eta,\hat{\ell}_{i}}^{(q)}(\lambda) V \right)^{H} Y_{\eta}^{(q)}(\lambda) \right|^{2}, \quad (22)$$

 $\gamma \in [0, 1]$ is a positive constant. Eq. (27) can be solved recursively using recursive least squares [38].

The output of the GSC for estimated source label $\hat{\ell}_i$ at each TF point (λ, k) and array q is given by:

$$S_{k,\hat{\ell}_i}^{(\text{GSC},q)}(\lambda) = \left(G_{k,\hat{\ell}_i}^{(q)}(\lambda)\right)^H Y_k^{(q)}(\lambda).$$
(23)

B. Post-Processing: Time-Frequency Masking

To improve the quality of the separated source signals, we exploit the spatial-spectral content of the GSC signals to construct a time-frequency (TF) mask following the approach in [39]. The construction of the TF mask relies on the assumption that the power spectrum of $S_{k,\hat{\ell}_i}^{(GSC,q)}(\lambda)$ is dominated by its corresponding source $\hat{\ell}_i$. For each source $\hat{\ell}_i$, a TF binary mask $\mathcal{M}_{k,\hat{\ell}_i}^{(q)}$ is constructed by comparing the relative power of the SOI to each of the interfering sources, with the intention of suppressing the interference. The estimated source is given by $\hat{S}_{k,\hat{\ell}_i}^{(q)}(\lambda) = \mathcal{M}_{k,\hat{\ell}_i}(\lambda) \cdot S_{k,\hat{\ell}_i}^{(GSC,q)}(\lambda)$, and the time-domain signal $\hat{s}_{\hat{\ell}_i}^{(q)}$ is given by the inverse STFT. In separating the source, we simply select the closest array to the estimated source position at each frame.

VI. EXPERIMENTS

In this section, we present the evaluations of the obtained multi-array measurements, the tracking filter performance, and the source separation performance on real data recorded in a physical room. Based on the same setting, we go further in evaluating the tracking and separation performance on simulated data with different reverberation times. The experimental setup is summarized in Section VI-A. The parameters used for the proposed method are explained in Section VI-B. Subsequently, we evaluate the quality of the SRP-PHAT multi-array measurements in Section VI-C, followed by the tracking performance of the multi-source Bayesian filter in Section VI-D, and the separation performance in Section VI-E.

A. Experimental Setup

The experiment is conducted in a $7.67 \text{m} \times 3.41 \text{m} \times 2.7 \text{m}$ enclosed room with reverberation measured at $T_{60} \approx 0.25 \text{s}$ using 4 linear arrays of 6 microphones (total of 24 mics), where all microphones are calibrated to the same gain/sensitivity. These microphones are connected into 3 *RME-OctaMic 8-channel* pre-amps. Each pre-amp is daisy-chained via MADI cables into the computer. All 4 microphone arrays are placed at the sides of the room as shown in Fig. 5.

As our proposed method is capable of handling an unknown number of moving sources, we design the experiment such that an active source (female speech) first appears in the scene and starts moving, followed later by another 2 active sources (male and female speech). It is also important to point out that the times at which these sources appear and disappear from the scene are unknown. The movement of each individual source is annotated by hand and the trajectories of the sources are illustrated in Fig. 5. In recording the source signals, we traverse each source according on the indicated path so that we can evaluate the



Fig. 5. Experimental Room Setup.

tracking results. Note that the sources are continuously active with typical short pauses in speech.

To evaluate the performance of the proposed method with different reverberation times, i.e. $T_{60} = 0.05s, 0.25s, 0.55s$, we use the Image Source Model (ISM) [34], [35] to simulate the acoustic room response for these reverberation times. The movements of the sources are the same as the annotated (ground-truth) trajectories in Fig. 5, and the source signals are convolved with simulated room impulse responses using a 512-sample block length.

B. Parameters Breakdown

1) Multi-Array Measurements: The microphone signals are sampled at $F_s=16$ kHz and subjected to high-pass filtering with 1 kHz cutoff to minimize the impact of reverberation on the multi-array measurements. The STFT of the raw signals is performed with a Hann window of frame length T=512, where each frame increment corresponds to a 32 ms time block. The multi-array position measurements are obtained via peakpicking with an empirically selected threshold.

2) Multi-Source Bayesian Tracking Filter: Recall that the parameters of the multi-source transition density are shown in Table II of Section IV-B. In audio speaker tracking where speech typically has short pauses, the Langevin model [11], [18], [40] is an appropriate choice for acoustic speaker tracking [12]. The motion model has the following state space equations [40]: $\alpha_{k+1} = \alpha_k + \phi \dot{\alpha}_k, \ \dot{\alpha}_{k+1} = e^{-\beta \phi} \dot{\alpha}_k + \phi \dot{\alpha}_k$ $\nu\sqrt{1-e^{-2\beta\phi}\Xi_k}$, where α_k and $\dot{\alpha}_k$ are the 3D position and velocity vectors respectively, β is the rate constant that controls the rate at which the velocity decays, ν is the steady-state root-mean-square velocity constant, ϕ is the discretization time step interval and Ξ_k is the process noise. The process noise Ξ_k models random disturbances in the state transition, and Ξ_k is a 3-dimensional Gaussian random vector with zero mean and covariance $\sigma_{\Xi}\sigma_{\Xi}^{T}$, where σ_{Ξ} is a column vector of the component standard deviations. Note that each component of Ξ_k is a Gaussian random variable that is statistically independent of one another and across time.

Based on this motion model, we specify the single-source state transition density as $f_S(x_{k+1}|x_k) = \mathcal{N}(x_{k+1}; Fx_k, RR^T)$, where $x_k \triangleq (\alpha_k, \dot{\alpha}_k)$, $\mathcal{N}(\cdot; Fx_k, RR^T)$ is a Gaussian *pdf* with mean Fx_k and covariance RR^T , $F = F_{\text{pseudo}} \otimes I_3$, $R = R_{\text{pseudo}} \otimes I_3$, I_3 an identity matrix of 3 dimensions, \otimes is the Kronecker product, and

$$\mathbf{F}_{\text{pseudo}} = \begin{bmatrix} 1 & \phi \\ 0 & e^{-\beta\phi} \end{bmatrix} \mathbf{R}_{\text{pseudo}} = \sigma_{\Xi} \begin{bmatrix} 0 \\ \nu\sqrt{1 - e^{-2\beta\phi}} \end{bmatrix}$$

In the experiment, the values of the Langevin model parameters are set to $\beta = 10s^{-1}$, $\nu = 1ms^{-1}$, and $\phi = 32ms$. The noise standard deviation is $\sigma_{\Xi} = [4.7, 4.7, 0.7]^T ms^{-1}$, where the z-component standard deviation is lower than that of the other components because movements in the z-axis are small. A high probability of survival $P_S = 0.999$ is selected as existing sources are likely to be persist.

The birth parameters are given by $\{r_B(\ell_i), p_B(\cdot, \ell_i) \triangleq \mathcal{N}(\cdot; \mu_B^{(i)}, P_B^{(i)})\}_{i=1}^3$, where $r_B(\ell_i)$ is the birth probability of a source with label ℓ_i and $p_B(\cdot, \ell_i)$ is the birth probability density which is a Gaussian with mean $\mu_B^{(i)}$ and covariance $P_B^{(i)}$. The Gaussian mean is a vector containing the expected location of source birth while the associated covariance specifies its spatial uncertainty. In the experiment, the values of these parameters are: $r_B(\ell_1) = r_B(\ell_2) = r_B(\ell_3) = 0.005, \mu_B^{(1)} = [5.0 \ 1.0 \ 1.8 \ 0 \ 0]^T, \mu_B^{(2)} = [4.0 \ 3.0 \ 1.5 \ 0 \ 0]^T, \mu_B^{(3)} = [2.5 \ 0.5 \ 1.5 \ 0 \ 0]^T, P_B^{(1)} = P_B^{(2)} = P_B^{(3)} = \text{diag}([0.15; 0.15; 0.15; 0.15; 0.15; 0.15]^T)^2$. Note that the Gaussian means have units of m for the 3D position components and ms⁻¹ for the 3D velocity components.

Subsequently, recall that the parameters of the *multi-array* measurement likelihood are shown in Table III of Section IV-C. The obtained array measurements are noisy in nature. Hence, each measurement from each array $z_k^{(q)}$ is related to the source state x_k via the measurement equation: $z_k^{(q)} = Hx_k + \zeta_k^{(q)}$ where q = 1, ..., Q, $\mathbf{H} = [\mathbf{I}_3, 0]$, and $\zeta_k^{(q)}$ is an additive Gaussian random vector that is used to model noise in the measurement. Similar to the process noise, $\zeta_k^{(q)}$ is a 3-dimensional Gaussian random vector with zero mean and covariance $\sigma_{\zeta^{(q)}} \sigma_{\zeta^{(q)}}^T$, where $\sigma_{\zeta(q)}$ is a column vector of the component standard deviations. Note that each component is a Gaussian random variable that is statistically independent of one another and across time. Based on this measurement model, the singlesource likelihood for each array q is given as: $g^{(q)}(z_k^{(q)}|x_k) =$ $\mathcal{N}(z_k^{(q)}; \mathbf{H}x_k, \sigma_{\zeta^{(q)}}\sigma_{\zeta^{(q)}}^T)$. In the experiment, the noise standard deviation vector is set to $\sigma_{\zeta^{(q)}} = [0.1, 0.1, 0.1]^T m$ for $q=1,\ldots,Q$. The probability of detection $P_D^{(q)}=0.6$ for q= $1, \ldots, Q$ is chosen to reflect the quality to the obtained measurements. The intensity function $\kappa^{(q)}(\cdot) = \mathcal{U}(\cdot)$ for $q = 1, \ldots, Q$ denotes an average of 10 false detections per frame where each individual false detection is uniformly distributed in its space.

3) Source Separation: In the separation module, the STFT of the raw microphone signals is performed with a 1024-sample Hann window with 50% overlap to reduce the effect of windowing [41]. Since STFT from a 1024-sample with 50% overlapping window corresponds to the same number of frames as STFT from a 512-sample window with no overlapping, the frames are synchronized from the tracking module to the separation module, so that tracking estimates obtained at each frame are used for the separation accordingly.



Fig. 6. Observed measurements projected onto 2D ground plane as represented by black crosses at frames k=120, 121 and 122 for Array 2. The true positions for the active sources at the relevant times are denoted by colored asterisks.

C. Evaluation of SRP-PHAT Multi-Array Measurements

Due to space constraints, we only present the evaluation on real data. Fig. 6 (a) shows the real measurements obtained from an array compared with the ground-truth source positions. Notice that there is noise, missed detections (false negatives) and false detections (false positives) as expected across time frames. To evaluate the quality of the obtained multi-array measurements, we need a distance function between two sets of points, i.e. the set which contains the array measurements and the set which contains the ground-truth source positions. This distance function must be able to capture the accuracy of the individual points and the mismatch in number of points. Conceptually, the distance function must satisfy the three axioms of a metric: *identity*, symmetry and triangle inequality. While the first two axioms are often easily met, the triangle inequality is equally important. Conformity with the triangle inequality ensures the metric to be consistent with geometric interpretation, i.e. the shortest distance between two points is a straight line.

To this end we employ the Optimal Sub-Pattern Assignment (OSPA) distance which is an established mathematically consistent and physically meaningful metric between two finite sets of points [31]. The OSPA distance captures both localization and cardinality errors between two finite sets with a suitable base-distance between the points. The Euclidean distance (2norm) is often used as the base-distance, and the resulting OSPA distance captures the perturbation error (localization) in the measurements caused by noise, and the error in the number of measurements (cardinality) caused by potential missed detections and false detections. Base-distances between two points that exceed the cutoff are capped at the cutoff value. The cutoff value is effectively the threshold at which a localization error is deemed as a cardinality error. A higher cutoff value brings more emphasis on the cardinality errors, and vice versa. The OSPA distance between the set containing the array measurements and the set containing true source positions is interpreted as a per-point error that ranges from zero to the cutoff value with units in meters. Interested readers can refer to [31] for full details.

For this evaluation, we compute the OSPA distance between the set of measurements obtained from each array and the set of source ground-truth trajectories with cutoff of 1 m as shown in Fig. 7. It is observed that the OSPA distance for each array has a



Fig. 7. OSPA distance between the obtained source measurements and true source positions (lower is better) for each microphone array (q=1,2,3,4).

TABLE III AVERAGE OSPA DISTANCE ON THE OBTAINED SOURCE MEASUREMENTS

Array	Average OSPA Components (m)			
	Localization	Cardinality	OSPA	
1	0.32	0.51	0.83	
2	0.31	0.54	0.85	
3	0.33	0.53	0.86	
4	0.33	0.51	0.84	

time average of about 0.8 m. This is supported by Table V, which shows the time average OSPA distance for each array along with its localization and cardinality components. The table indicates that the average localization error for each array is about 0.3 m, while the average cardinality error for each array is about 0.5 m. From these values, we observe that the OSPA distances have noticeable localization errors but are still dominated by cardinality errors. Consequently, when measurements corresponding to the direct path are obtained, they are somewhat noisy, while it it also clear that there is a high number of missed detections and false detections. Combined with the fact that the measurements have no identities or labels, and that the number of sources are unknown and time-varying, it is clear that source separation via spatial filtering using the multi-array measurements is not viable.

D. Evaluation of Multi-Source Tracking Filter

The multi-array measurements are fed into the multi-source Bayesian tracking filter (MS-GLMB filter) at each frame, which outputs the filtering density. This output is fed back into the filter to process multi-array measurements at the next frame, and into the estimator to generate the multi-source state estimate which contains the estimated source tracks (positions and labels).

A track is defined when the source position estimates across frames are associated with a common label. Specifically, the mathematical definition of a track is a function whose domain is the set of time instants at which the source exists. In online tracking, a track can be fragmented or "broken" when the estimated source labels are not matching across time frames. Another common error is track switching which occurs when the label of a track switches to another. While the OSPA distance provides an indication of the acoustic measurement performance, it does not account for labeling errors between the estimated and true sets of tracks. As a result, it does not penalize track switching and fragmentation. In order to evaluate the estimated source tracks against the ground-truth source trajectories, we need a distance function to characterize the error between tracks over a time window.

To achieve this, we use the $OSPA^{(2)}$ metric which is defined for two sets of tracks, i.e. the set of estimated source tracks and



Fig. 8. 3D estimated source tracks (colored dots) vs the true source trajectories (colored lines) plotted against time.

the set of true source tracks. The construction of the $OSPA^{(2)}$ metric is based on the OSPA metric. In particular, $OSPA^{(2)}$ uses a time-averaged OSPA distance (over the common track times, with an appropriate cutoff) between a pair of tracks as the base-distance. The $OSPA^{(2)}$ distance treats the individual tracks as individual points in a larger space of tracks. The $OSPA^{(2)}$ distance is constructed as the OSPA distance between the two sets of tracks where the base distance is defined directly above [32]. Hence, the name $OSPA^{(2)}$ reflects the OSPA-on-OSPA nature in its construction. The $OSPA^{(2)}$ distance is capable of penalizing track switches (label changes) and fragmentations ("broken" tracks). The $OSPA^{(2)}$ is also parameterized by the cutoff value, which provides a sensitivity tradeoff between localization and cardinality errors between the tracks. The interpretation of the OSPA⁽²⁾ distance evaluated over a fixed time window is consequently a time-averaged per-track error. The complete breakdown of the $OSPA^{(2)}$ metric can be found in [32].

For online tracking, it is desirable to have the tracking performance as a function of time. This can be achieved by computing the $OSPA^{(2)}$ distance over a sliding window instead of a fixed time window. This means that the $OSPA^{(2)}$ distance is plotted against time as the sliding window moves forward. Tracks whose domains lie outside the window are disregarded. This is useful for "forgetting" errors that were made further in the past. For this evaluation, a cutoff of 1 m and a window length of 30 frames are used.

1) Real Data: The 3D estimated tracks (colored dots) from the MS-GLMB tracking filter are compared with the source ground-truth trajectories (colored lines) in Fig. 8, where the color of a dot represents the label of a particular track. While the estimated tracks for Source 1 (red), 2 (green) and 3 (blue) at frame 1, 11 and 61 respectively have slight delays in the initiations, we observe that the tracking filter manages to initiate and maintain all 3 estimated tracks consistently across frames with respect to the ground-truth trajectories.



Fig. 9. $OSPA^{(2)}$ distance between estimated and true source trajectories (lower is better).



Fig. 10. OSPA⁽²⁾ distance between estimated and true source trajectories (lower is better).

Fig. 9 shows the OSPA⁽²⁾ distance between the estimated tracks and the ground-truth trajectories plotted against time. Notice that the spikes of the curve correspond to the errors caused by the late track initiations and terminations of Source 1, 2 and 3 as depicted Fig. 8. Despite noise, false detections (false positives) and missed detections (false negatives) in the obtained multi-array measurements, the result validates the proposed tracking filter for solving the space-time permutation problem, and producing tracks for each source with reasonable accuracy as corroborated by both Fig. 8 and Fig. 9.

2) Simulated Data: Due to space constraints, we omit the 3D-track plots for simulated data and only present the OSPA⁽²⁾ distances for the tracking estimates generated at reverberation times $T_{60} = 0.05$ s, 0.25s, and 0.55s in Fig. 10.

At $T_{60} = 0.05$ s (in black), the MS-GLMB tracking filter achieves the lowest OSPA⁽²⁾ distance compared to the other 2 curves, indicating that the tracking result is the best out of the other 2 examples. This is expected as the multi-array measurements capture the direct path.

At $T_{60} = 0.25$ s (in blue), we see that the error curve is similar to the OSPA⁽²⁾ error curve on real data, where the spikes are caused by the delays in track initiations and terminations. This indicates an agreement between the simulation and the real measurements.

At $T_{60} = 0.55$ s (in red), we observe that the error curve is higher than that of the previous two curves, indicating a poorer tracking result. This increase in error is caused by late track initiations and terminations, and larger localization error due to higher reverberation.

E. Evaluation of Source Separation

For moving sources, the delay of the source signal with respect to any microphone array is changing over time. In our proposed method, the selection of the array for source separation depends on the source position at each frame. Therefore, perceptual measures such as PESQ [42], STOI [43] and PEASS [44] that rely on delay-compensation, are not directly applicable. One

TABLE IV Scales of SIG, BAK AND OVRL IN THE SUBJECTIVE LISTENING TEST

	SIC		
	010		
Rating	Description		
5	Very natural, no degradation		
4	Fairly natural, little degradation		
3	Somewhat natural, somewhat degraded		
2	Fairly unnatural, fairly degraded		
1	Very unnatural, very degraded		
BAK			
Rating	Description		
5	Not noticeable		
4	Somewhat noticeable		
3	Noticeable but not intrusive		
2	Fairly conspicuous, somewhat intrusive		
1	Very conspicuous, very intrusive		
	OVRL		
Rating	Description		
5	Excellent		
4	Good		
3	Fair		
2	Poor		
1	Bad		

possibility for using these measures is to consider time blocks where the sources are almost stationary. However, this is outside the scope of this paper as there may not be enough signal information in those frames, and a very complex study is needed with the development of suitable measures. Conventional BSS performance measures that are based on signal (energy) ratios, i.e. the BSSEval [45], require an exact time-alignment between the estimated and true signals to work [45]. As our experiment involves sources that are moving, and the exact times at which the sources appear in the scene are unknown, BSSEval is also not suitable for evaluating the source separation performance.

To evaluate the separation performance, we administered a subjective listening test on all scenarios based on the ITU-T P.835 methodology specifically designed to evaluate the distortions and overall quality of noise suppression algorithms [33]. In the test, each participant is instructed to listen to the clean speech signal (upper anchor reference), the separated speech signal (to be evaluated) and the mixture signal (lower anchor reference), then rate them on:

- The speech signal alone using a five-point scale of signal distortion (SIG);
- The background interfering noise alone using a five-point scale of background intrusiveness (BAK);
- The overall quality using the scale of mean opinion score (OVRL).

The scales of SIG, BAK and OVRL are described in Table VI. The listening tests are carried out on the separated signals both before and after the post-processing step. This form of ablation study is undertaken with the intention of understanding the tradeoff between additional speech suppression and signal distortion due to the optional post-processing.

In this evaluation, 17 people (11 males, 6 females) of ages from 20 to 40 are recruited to partake in the listening test. To assess the overarching discrepancies between the test ratings on the separated speech signal and the unprocessed mixture signal, a statistical analysis of variance (ANOVA) is adopted to present



Fig. 11. Mean scores for SIG, BAK, and OVRL for the estimated source signals, ablation study (estimation without post-processing), and original mixture signals evaluated on real data.

TABLE V ONE-WAY ANOVA TEST BETWEEN THE ESTIMATED SOURCE SIGNALS AND ORIGINAL MIXTURE SIGNALS ON REAL DATA, AND CORRESPONDING ANOVA TEST FOR THE ABLATION STUDY (ESTIMATION WITHOUT POST-PROCESSING)

Source			p-value	
Source		SIG \uparrow	BAK \downarrow	$OVRL\downarrow$
1	Proposed	0.0791*	0.0001	0.0001
	Ablation	0.9247*	0.0058	0.0053
n	Proposed	0.1122*	0.0001	0.0001
2	Ablation	0.9349*	0.0059	0.0051
2	Proposed	0.1494*	0.0001	0.0001
3	Ablation	0.8694*	0.0054	0.0052

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (\uparrow means higher is better while \downarrow means lower is better.)

the significant statistical difference between the quality of the separated speech signal and the unprocessed mixture based on a 0.05 significance level.

1) Real Data: For the subjective listening test, the mean scores over all 3 aspects, i.e. SIG, BAK and OVRL, of the separated/estimated source signals and the unprocessed mixture signals are presented in Fig. 12. We observe that the BAK and OVRL mean scores of all three estimated source signals from the proposed method (the blue bars) are relatively high as compared to the mean scores of the mixture signals, while the SIG mean scores of all estimated and mixture signals are relatively close. This indicates that the source signals are well separated with minimal signal distortions.

The*p*-values of the one-way ANOVA test between the estimated source signals and the unprocessed mixture signals are tabulated in Table VIII. In terms of SIG, the table shows that all values of the proposed method are higher than 0.05, which means that there is no statistically significant difference in signal distortion between the estimated source signals and the mixture signals. In terms of BAK and OVRL, the table shows that all values of the proposed method are less than 0.05, indicating a statistically significant difference in speech intrusiveness and overall quality respectively.

From the results of the ablation study in Fig. 12 (the green bars) and Table VIII, it can be seen that the BAK and OVRL means scores are slightly poorer than that of the proposed method, but the SIG mean scores are better than that of the proposed method across the board. Subsequently, the BAK and OVRL *p*-values indicate that there is a statistically significant difference in speech intrusiveness and overall quality, whereas

TABLE VI ONE-WAY ANOVA TEST BETWEEN THE ESTIMATED SOURCE SIGNALS AND ORIGINAL MIXTURE SIGNALS ON SIMULATED DATA, AND CORRESPONDING ANOVA TEST FOR THE ABLATION STUDY (ESTIMATION WITHOUT POST-PROCESSING)

$T_{\rm ex}(a)$	Source		p-value		
160(8)			SIG ↑	BAK \downarrow	OVRL ↓
	1	Proposed	0.1903*	0.0001	0.0001
		Ablation	0.7393*	0.0031	0.0034
0.05	2	Proposed	0.2279*	0.0001	0.0001
0.03	2	Ablation	0.7618*	0.0031	0.0033
	3	Proposed	0.2169*	0.0001	0.0001
		Ablation	0.6511*	0.0031	0.0032
0.25	1	Proposed	0.1802*	0.0001	0.0001
	1	Ablation	0.8421*	0.0051	0.0048
	2	Proposed	0.1885*	0.0001	0.0001
		Ablation	0.7978*	0.0051	0.0047
	3	Proposed	0.1868*	0.0001	0.0001
		Ablation	0.7296*	0.0048	0.0049
0.55	1	Proposed	0.1629*	0.0355	0.1051*
	1	Ablation	0.8397*	0.0411	0.2481*
	2	Proposed	0.1741*	0.0396	0.3791*
		Ablation	0.9711*	0.0443	0.4521*
	3	Proposed	0.0791*	0.0343	0.2041*
		Ablation	0.8883*	0.0403	0.3451*

The asterisk (*) denotes values that are above the selected significance level, i.e. 0.05. (\uparrow means higher is better while \downarrow means lower is better.)

the SIG p-values indicate that there is no statistically significant difference in signal distortion. These observations indicate that the proposed method minus post-processing achieves noticeable speech suppression with negligible signal distortion. The addition of the post-processing does indeed further enhance interference suppression, but at the cost of some signal distortion which manifests as musical noise in the estimated signals.

In summary, the proposed method achieves source separation with good noise (interfering speech) suppression, which is corroborated by both the mean scores and the ANOVA test in Fig. 12 and Table VIII respectively. The audio files for this experiment are available in Supplementary Materials.

2) Simulated Data: The mean scores for the subjective listening test on the estimated source signals and the unprocessed mixture signals, obtained under reverberation times $T_{60} = 0.05$ s, 0.25s and 0.55s, are shown in Fig. 13. Based on the relative differences for SIG, BAK and OVRL between all estimated and mixture signals at $T_{60} = 0.05$ s and 0.25s, we observe a similar pattern as for the real data, which shows that the proposed algorithm is capable of separating the sources reasonably well. However, at $T_{60} = 0.55$ s, the separation performance degrades as the mean scores between all estimated and mixture signals are relatively close. Overall, we observe a downward trend in mean scores of the estimated source signals from low T_{60} to high T_{60} . This degradation is expected because both tracking and separation performance degrades with increasing reverberation.

The *p*-values of the one-way ANOVA test between the estimated source signals and the unprocessed mixture signals are tabulated in Table IX. In terms of SIG, the *p*-values for all three sources at all reverberation times are above 0.05. This indicates that signal distortions between the estimated source signals and mixture signals are very similar. In terms of BAK and OVRL, the computed *p*-values for all three sources at $T_{60} = 0.05$ s and



Fig. 12. Mean scores for SIG, BAK, and OVRL for the estimated source signals, ablation study (estimation without post-processing), and original mixture signals evaluated on simulated data.

0.25s are below 0.05. This suggests that speech intrusiveness and the overall quality of the estimated source signals are statistically different from the mixture signals, thus indicating good separation.

At $T_{60} = 0.55$ s however, we see that the *p*-values on OVRL for all three sources are higher than 0.05, suggesting that there is no statistically significant difference between the overall quality of the estimated source signals and the mixture signals. This, combined with the fact that the BAK values are fairly close to 0.05, suggests an overall poorer separation performance as interfering speech is not well suppressed. This decrease in performance is most likely due to two main reasons. The first being as reverberation time increases, the quality of tracking deteriorates, resulting in more localization errors. The second being the failure in the signal sparsity assumption, which results in leakage in the TF masking.

Examination of the ablation results in Fig. 13 and Table IX reveals similar trends to those observed in the real-data experiments. For each of the three reverberation levels, the proposed method minus post-processing achieves noticeable suppression with negligible distortion, but the addition of the post-processing involves a trade-off between further suppression and audible distortion.

It can be seen that the separation performance on simulated data at $T_{60} = 0.25$ s matches the results on real data, and that the separation performance generally degrades as reverberation time increases. This is corroborated by both the mean scores and the ANOVA test in Fig. 13 and Table IX. We also note the presence of more perceptible signal distortion in real data compared to simulated data. This is likely due to the mismatch between the real room environment and the simulated room model, which leads to additional spectral leakage in the time-frequency masking of the post-processing. The audio files for the experiments on simulated data are also provided in Supplementary Material.

VII. CONCLUSION

This paper proposes a block-wise or online solution for blind source separation with multiple microphone arrays, which can accommodate an unknown time-varying number of acoustic moving sources in mild reverberation. The proposed solution is based on first obtaining source position measurements, then estimating the trajectories of the sources, and finally separating the mixed signal with corresponding spatial filtering. In real acoustic recordings measured at $T_{60} \approx 0.25$ s, it is observed that

the SRP-PHAT source measurements are relatively noisy, and contain significant false and missed detections. In addition, the measurements are unlabeled, and coupled with the unknown appearance, disappearance and movement of sources, it is not known which source generated which measurement at the current time, nor which measurements are connected to the same source across time. These observations verify the extent of the inherent space-time permutation problem, which is then addressed with the application of a labeled RFS based MS-GLMB tracking filter. Results indicate that the tracking filter is able to recover the source trajectories (i.e. the positions and identities) from the imperfect source measurements with some delay in initiation and termination. Separation is carried out via a corresponding set of time-varying generalized side-lobe cancellers. Evaluations with subjective listening tests confirm acceptable performance in mild reverberation. Additional experiments via acoustic room simulations with the ISM method indicate clear separation performance at lower reverberation $T_{60} = 0.05$ s, matching performance in mild reverberation $T_{60} = 0.25$ s, and noticeable deterioration at higher reverberation $T_{60} = 0.55$ s. Future works will investigate the impacts of array configuration and placement which are beyond the scope of the current paper.

REFERENCES

- X. Yu, D. Hu, and J. Xu, Blind Source Separation: Theory and Applications. Hoboken, NJ, USA: Wiley, 2013.
- [2] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
- [4] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and-norm minimization," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, 2006, Art. no. 024717.
- [5] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.
- [6] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, Aug. 2016.
- [7] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 281–295, Feb. 2018.
- [8] A. Masnadi-Shirazi and B. D. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 828–841, Apr. 2013.

- [9] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 657–670, Mar. 2018.
- [10] K. Weisberg, B. Laufer-Goldshtein, and S. Gannot, "Simultaneous tracking and separation of multiple sources using factor graph model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2848–2864, Jan. 2020.
- [11] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.
- [12] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, 2006, Art. no. 017021.
- [13] C. Evers, E. A. Habets, S. Gannot, and P. A. Naylor, "DoA reliability for distributed acoustic tracking," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1320–1324, Sep. 2018.
- [14] C. Evers *et al.*, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.
- [15] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying rao-blackwellised monte carlo data association," in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, 2009, pp. 253–256.
- [16] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Proc. Int. Eval. Workshop Classification Events, Activities Relationships.*, 2006, pp. 137–150.
- [17] R. P. Mahler, Advances in Statistical Multisource-Multitarget Information Fusion. Norwood, MA, USA: Artech House, 2014.
- [18] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.
- [19] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearingonly acoustic tracking of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2015, pp. 1206–1210.
- [20] H. Pessentheiner, "Localization, characterization, and tracking of harmonic sources with applications to speech signal processing," Ph.D. dissertation, Graz University of Technology, 2017.
- [21] C. Evers and P. A. Naylor, "Acoustic SLAM," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 9, pp. 1484–1498, Sep. 2018.
- [22] A. Plinge, D. Hauschildt, M. H. Hennecke, and G. A. Fink, "Multiple speaker tracking using a microphone array by combining auditory processing and a Gaussian mixture cardinalized probability hypothesis density filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 2476–2479.
- [23] N. Chong, S. Nordholm, B. T. Vo, and I. Murray, "Tracking and separation of multiple moving speech sources via cardinality balanced multi-target multi-bernoulli (CBMeMBer) filter and time frequency masking," in *Proc. Int. Conf. Control, Automat. Inf. Sci..*, 2016, pp. 88–93.
- [24] X. Zhong and J. R. Hopgood, "A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2356–2370, Dec. 2015.
- [25] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1, 2007, pp. I– 121.
- [26] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3460–3475, Jul. 2013.
- [27] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled random finite sets and the bayes multi-target tracking filter," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6554–6567, Dec. 2014.
- [28] B.-N. Vo, B.-T. Vo, and H. G. Hoang, "An efficient implementation of the generalized labeled multi-bernoulli filter," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1975–1987, Apr. 2017.
- [29] B.-N. Vo, B.-T. Vo, and M. Beard, "Multi-sensor multi-object tracking with the generalized labeled multi-bernoulli filter," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 5952–5967, Dec. 2019.

- [30] S. E. Nordholm, H. H. Dam, C. C. Lai, and E. A. Lehmann, "Broadband beamforming and optimization," in *Academic Press Library Signal Process.*. Elsevier, 2014, vol. 3, pp. 553–598.
- [31] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.
- [32] M. Beard, B. T. Vo, and B.-N. Vo, "A solution for large-scale multi-object tracking," *IEEE Trans. Signal Process.*, vol. 68, pp. 2754–2769, 2020.
- [33] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IIEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
- [34] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the imagesource model," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 159–162.
- [35] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.
- [36] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.
- [37] R. P. Mahler, Statistical Multisource-Multitarget Information Fusion. Norwood, MA, USA: Artech House, Inc., 2007.
- [38] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "Recursive leastsquares algorithms," in *Perspective Stereophonic Acoustic Echo Cancellation.* Springer, 2011, pp. 63–69.
- [39] J. P. Morgan, "Time-frequency masking performance for improved intelligibility with microphone arrays," Ph.D. dissertation, Master Thesis in the College of Engineering at the University of Kentucky, 2017.
- [40] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, 2001, pp. 3021–3024.
- [41] I. Cohen, J. Benesty, and S. Gannot, Speech Processing in Modern Communication: Challenges and Perspectives, vol. 3, Springer, 2009.
- [42] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (Cat. No. 01CH37221)*, vol. 2, 2001, pp. 749–752.
- [43] 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.
- [44] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
- [45] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.



Jonah Ong received the B.E. degree in 2018 in electrical and power engineering (with first-class honors) from Curtin University, Perth, WA, Australia, where he is currently working toward the Ph.D. degree in electrical engineering. His research interests include statistical signal processing, Bayesian filtering and estimation, random sets, and multitarget tracking.



Ba Tuong Vo received the B.Sc. degree in applied mathematics and the B.E. degree in electrical and electronic engineering (with first-class honors) in 2004 and the Ph.D. degree in 2008 in engineering (with Distinction) from The University of Western Australia, Perth, WA, Australia. He is currently a Professor of signal processing with Curtin University, Perth, WA, Australia. His primary research interests include random sets, filtering and estimation, multiple object systems.



Sven Nordholm received the M.Sc.-EE (Civilingenjör) degree, the Licentiate degree in engineering, and the Ph.D. degree in signal processing from Lund University, Lund, Sweden, in 1983, 1989, and 1992, respectively. Since 1999, he has been a Professor of signal processing with the School of Electrical and Computer Engineering, Curtin University, Perth, WA, Australia. He is the Co-Founder of two start-up companies, which include Sensear, providing voice communication in extreme noise conditions and Nuheara a hearables company. He has authored or coauthored

more than 200 papers in refereed journals and conference proceedings. He frequently contributes to book chapters and encyclopedia articles. He is holding seven patents in the area of speech enhancement and microphone arrays. His primary research interests include speech enhancement, adaptive and optimum microphone arrays, audio signal processing, and wireless communication. He was a Lead Editor for a SPECIAL ISSUE ON ASSISTIVE LISTING TECHNIQUES IN IEEE SIGNAL PROCESSING MAGAZINE and several other EURASIP special issues. He is a former Associate editor for the *Eurasip Advances in Signal Processing* and *Journal of Franklin Institute*, and is currently an Associate Editor for the IEEE/ACM TRANSACTION ON AUDIO, SPEECH AND LANGUAGE PROCESSING.