An ICA-SCT-PHD Filter Approach for Tracking and Separation of Unknown Time-Varying Number of Sources

Alireza Masnadi-Shirazi, *Student Member, IEEE*, and Bhaskar D. Rao, *Fellow, IEEE*

***Abstract*—In this paper we present a solution to the problem of tracking and separation of a mixture of concurrent sources in a reverberant environment where the number of sources is unknown and varies with time: new sources can appear and existing sources can disappear or undergo silence periods. In order to deal with this challenging problem, we synergistically combine two key ideas, one in the front end and the other at the back end. In the front end we employ independent component analysis (ICA) to demix the mixtures and the state coherence transform (SCT) to represent the signals in a direction of arrival (DOA) detection framework. By exploiting the frequency sparsity of the sources, ICA/SCT is even effective when the number of simultaneous sources is greater than the number of sensors therefore allowing for minimal number of sensors to be used. At the back end, the probability hypothesis density (PHD) filter is incorporated in order to track the multiple DOAs and determine the number of sources. The PHD filter is based on random finite sets (RFS) where the multi-target states and the number of targets are integrated to form a set-valued variable with uncertainty in the number of sources. A Gaussian mixture implementation of the PHD filter (GM-PHD) is utilized that solves the data association problem intrinsically, hence providing distinct DOA tracks. The distinct tracks also make the separation task possible by going back and rearranging the outputs of the ICA stage. The tracking and separation capabilities of the proposed method is demonstrated using simulations of multiple sources in reverberant environments.**

***Index Terms*—Blind source separation, independent component analysis, multi-target tracking, probability hypothesis density, random finite sets, source localization.**

I. Introduction

PASSIVE localization and tracking of multiple acoustical sources is of great interest in the field of microphone arrays which is driven by applications such as automatic camera steering for teleconferencing and surveillance. Speaker localization is also very useful in aiding systems achieving the task of separating concurrent speakers or a desired speaker from background interference with applications such as high-quality hearing aids, speech enhancement and noise reduction for smart phones. By localization, one can refer to finding the bearings of the speakers or their Cartesian coordinate. In this paper we are particularly interested in estimating the bearing information of multiple sources or their direction of arrival (DOA) by means of the time difference of arrival (TDOA).

TDOA estimation is the first stage for many speaker localization algorithms involving one or more microphone pairs. In the case of a single speaker, TDOA can be reliably estimated using the generalized cross-correlation phase transform (GCC-PHAT) using one microphone pair [1], [2]. GCC-PHAT is a scanning method that computes the correlation of the microphone pair inputs for a range of TDOAs with an arbitrary resolution, resulting in peaks where the correlation is high. In case of multiple speakers, GCC-PHAT does not always provide reliable TDOA for all the sources since one of the sources can dominate over the others [3]. This means that as the concurrent sources increase in number, multiple TDOA estimation using GCC-PHAT becomes less reliable. Also, multipath propagation due to reverberation can cause additional peaks in the GCC-PHAT. This results in the situation where for example in the case of two sources, the first and second peaks do not always correspond to the first and second sources and sometimes the third or subsequent peaks need to be considered [4]. Extensions of the GCC-PHAT for multiple sources have been proposed [5]–[8]. However, they require microphone pair redundancy and high sampling rates to increase the reliability of the TDOA estimates.

Multiple TDOA estimation using frequency domain independent component analysis (ICA) was first proposed in [9]. In the context of blind source separation (BSS), ICA is a well known tool for the separation of linear and instantaneous mixed signals picked up by multiple sensors [10]. ICA estimates a de-mixing matrix for the separation task and does so by assuming the sources are statistically independent and non-Gaussian distributed. For many real world problems, the signals undergo a convoluted mixing due to reverberation. By transforming the mixture to the frequency domain by applying the short-time Fourier transform (STFT), convolution in the time domain translates to linear mixing in the frequency domain. Subsequently, ICA can be performed on every single frequency bin. Since ICA is indeterminate of source permutation, further post processing methods are necessary to correct for possible permutations of the separated sources in each frequency bin [11], [12]. In [9], multiple TDOAs are calculated directly from the columns of the estimated mixing matrix. However, this method works well only if the possible source permutations in the

Manuscript received May 22, 2012; revised September 28, 2012; accepted November 24, 2012. Date of publication December 24, 2012; date of current version January 18, 2013. This work was supported in part by a grant from Qualcomm Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: amasnadi@ucsd.edu; brao@ece.ucsd.edu).

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/TASL.2012.2236318

frequency bins have been corrected and there are no frequency bins affected by spatial aliasing (hence a minimal microphone spacing). Recently an extension to [9] has been proposed under the name of state coherence transform (SCT) that does not require permutation correction and is insensitive to spatial aliasing [13], [14]. Similar to GCC-PHAT, SCT is a scanning method. However, instead of finding the correlation between the two microphone input signals for TDOA points in the scan, it forms a pseudo-likelihood between a propagation model for the different TDOA scan points and the TDOA observations pertaining to the columns of the mixing matrices, resulting in peaks where the scan points in the model and observations best match. One attractive feature of SCT is that by exploiting the frequency sparsity of the sources, it is effective even when the number of simultaneous sources is larger than the number of sensors. Also, since SCT uses ICA outputs which attempt to separate the sources, it is more suitable for TDOA estimation for multiple sources compared to GCC-PHAT [14].

Assuming that the number of sources is known and fixed in time, some methods exist that track the location information for each source by incorporating a separate tracker for each source [15]. However, in many real world problems, not only do the states of the sources change with time, the number of concurrent sources is unknown and varies with time as new speakers can appear and existing speakers can disappear or undergo long silence periods. Moreover, the measurements can receive a set of spurious peaks (clutter) due to the multi-path propagation caused by reverberation and spatial aliasing, resulting in false alarms. In addition, not all of the sources are detected giving rise to missed detections as well. Therefore, the passive scanning methods discussed earlier result in an assortment of indistinguishable observations where only a subset of them are generated by the sources. Recently, methods based on random finite sets (RFS) have presented promising and mathematically elegant solutions to the problem of multi-target tracking (MTT) for time-varying number of targets [16], [17]. Using RFSs, the collection of indistinguishable observations in the presence of clutter is treated as a set-valued observation while the multi-target states and the number of targets are integrated to form a set-valued state. The goal becomes to estimate the target states and the target number while rejecting clutter and accounting for missed detections. The RFS formulation allows the problem to be posed in an optimal multi-target Bayesian filtering framework, and is an extension of the well known single target Bayes filter. However, the optimal RFS Bayes filter is computationally intractable as it becomes a combinatorial problem on the number of targets involving high dimensional integrals. The probability hypothesis density (PHD) filter is a suboptimal approximation to the RFS Bayes filter which propagates the first moment of multi-target posterior density rather than the full posterior density [16]. This said, the PHD filter still involves multiple integrals with no closed form solution in general. Also, the PHD filter in itself, does not solve the data association problem indicating which estimate belongs to which target. The Gaussian mixture implementation of the PHD filter (GM-PHD) alleviates these two difficulties: It provides a closed form solution of the PHD filter when the target states and observations follow a linear/Gaussian dynamic model (which is a reasonable model for the problem of interest in this paper) [18]. It also solves the data association problem intrinsically and provides track labels which are imperative to the separation task of interest [19].

The problem of extracting location information of unknown time-varying number of speakers using RFSs and PHD filtering has been proposed before. These methods, however, use GCC-PHAT in the front-end to obtain the measurements and bear the inherent limitations of GCC-PHAT for multiple sources including being inherently incapable of source separation [4], [20]. For the same problem, a method exists that uses ICA/SCT in the front-end and uses a naive thresholding approach to estimate the number of targets [21], [22]. This method, however, is sensitive to the selected thresholds and relies solely on the thresholds to reject clutter. Another class of methods uses a steered beamformer for acquiring the measurements and then applies a variable-dimension particle filter or track-before-detect filtering scheme for the tracking and source activity detection [23], [24]. These methods cannot perform the separation task inherently and don't quite estimate the number of targets but estimate the activity pattern of a limited few number of sources. In a previous work, we have proposed an ICA-based approach to separate and track multiple sources for when the sources can experience short silence periods [25]. This method, while being able to separate the sources, only estimates the activity patterns and cannot handle new sources being born or completely dying out. In this paper we propose the use of the GM-PHD to filter the measurements obtained from short time blocks using ICA/SCT. By doing so we are able to track the DOA of multiple time-varying number of sources and from the track labels we are able to go back to the ICA outputs and perform the separation task by associating each separated time-frequency block with its estimated corresponding track. The separation scheme exploits the frequency sparsity of the sources and enables the separation of more concurrent sources than sensors. Computer simulations on the DOA tracking using the proposed method is compared with the first two aforementioned existing approaches and the results are favorable and promising.

Overall, this paper demonstrates how a mixture/superposition model in the framework of BSS can be easily represented as a standard detection model in the framework of multi-target tracking, assuming that the sources have frequency sparsity. Such an idea of transforming a mixture/superposition model to a detection model, was first presented in [26], where the sources were assumed to be narrowband audio tones and the STFT representation was enough to execute such transformation. As it turns out, the approach in [26] is a special case of the proposed method for when the sources have a super-sparse representation to a degree where they will be non-overlapping and occupy a single frequency bin, making the ICA separation scheme unnecessary. The proposed method offers a solution for executing the transformation from the mixture model to the detection model for broadband signals that have some sort of frequency sparsity, such as speech and communication signals. Recently, in the context of multi-target tracking, other methods have been proposed that deal with the mixture/superposition model directly and perform a moment-based RFS filtering [27]–[30]. These methods,

however, are either computationally intractable or do not enjoy the relative simplicity of the PHD filter.

The paper is organized as follows: Section II explains the front end of the system consisting of ICA in junction with SCT where the mixture representation of the sources are transformed to a DOA multi-target detection representation, regardless of permutation, spatial aliasing and the number of sources being more than the sensors. Section III explains the back-end of the system and gives a background theory on multi-target filtering using a RFS framework along with implementations and appropriate extensions of the PHD filter. We present all the formulations of this section in a summarized way while maintaining a consistent context. In Section IV, we present how the front-end and back-end are synergistically combined to perform both the tracking and separation tasks. In Section V, some experimental results are evaluated. Finally, in Section VI, our conclusions are stated and the main contributions of the paper are summarized.

II. Frequency Domain BSS and SCT

Assuming L sensors and M sources, the convolutedly mixed observation at the l^{th} sensor at time u is

$$y_l(u) = \sum_{j=1}^{M} \sum_{r=0}^{R-1} h_{lj}(r) s_j(u-r) \tag{1}$$

where $s_j(u)$ is the j^{th} source in the time domain, h_{lj} is the finite impulse response (FIR) approximation of duration R linking the j^{th} source to the l^{th} sensor. The signals are transformed to the frequency domain using the short time Fourier transform (STFT). The STFT takes the discrete Fourier transform (DFT) of frames of the signal using a sliding window, hence creating a time-frequency representation of the signal, commonly known as the spectrogram. We must note that the window length of the STFT should be sufficiently large, ensuring that the conversion from convolution in the time domain, is approximated fairly by multiplication in the frequency domain. Using STFT, the l^{th} sensor observation at time frame n and frequency bin $k = 1, \ldots, K$ becomes

$$Y_l(n,k) = \sum_{j=1}^{M} H_{lj}(k) S_j(n,k) \tag{2}$$

where $S_j(n,k)$ is the frequency domain representation of the j^{th} source at bin k and frame n. Omitting n for simplicity, we can arrange (2) for frequency bin k in matrix form as

$$Y(k) = H(k)S(k) \tag{3}$$

where $Y(k) = [Y_1(k) \ldots Y_L(k)]^T$, $S(k) = [S_1(k) \ldots S_M(k)]^T$ and $H(k)$ is the $L \times M$ mixing matrix corresponding to the k^{th} frequency bin.

For the case of $L = M$, any complex-valued ICA algorithm [10] can be applied to each frequency bin to estimate the inverse of the mixing matrix $H(k)$. Denoting the estimate of the separated sources at the k^{th} bin as $\hat{S}(k)$, from ICA we get

$$\hat{S}(k) = \hat{W}(k)Y(k) \tag{4}$$

where $\hat{W}(k)$ denotes the estimate of the demixing matrix up to scaling and permutation ambiguities:

$$\hat{W}(k) = \Lambda(k)\Pi(k)\hat{H}^{-1}(k) \tag{5}$$

where $\Lambda(k)$ is a diagonal scaling matrix, $\Pi(k)$ is a permutation matrix and $\hat{H}(k)$ is the estimate of the true mixing matrix $H(k)$.

Without loss of generality, for simplicity, we consider a configuration of two sources and two sensors. In an ideal anechoic setting the true mixing matrix can be modeled as

$$H(k) = \begin{pmatrix} |h_{11}(k)|\, e^{-j2\pi f_k T_{11}} & |h_{12}(k)|\, e^{-j2\pi f_k T_{12}} \\ |h_{21}(k)|\, e^{-j2\pi f_k T_{21}} & |h_{22}(k)|\, e^{-j2\pi f_k T_{22}} \end{pmatrix} \tag{6}$$

where T_{qp} is the propagation time from the p^{th} source to the q^{th} microphone and f_k is the frequency in Hz for the k^{th} frequency bin. By neglecting the permutation problem for now but taking into account the scaling ambiguity, the estimate of the inverse of the demixing matrix becomes

$$\begin{aligned} &\hat{W}^{-1}(k) \\ &= \begin{pmatrix} \left|\hat{h}_{11}(k)\right| e^{-j2\pi f_k \hat{T}_{11}} & \left|\hat{h}_{12}(k)\right| e^{-j2\pi f_k \hat{T}_{12}} \\ \left|\hat{h}_{21}(k)\right| e^{-j2\pi f_k \hat{T}_{21}} & \left|\hat{h}_{22}(k)\right| e^{-j2\pi f_k \hat{T}_{22}} \end{pmatrix} \begin{pmatrix} \frac{1}{\eta_1(k)} & 0 \\ 0 & \frac{1}{\eta_2(k)} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\eta_1(k)} \left|\hat{h}_{11}(k)\right| e^{-j2\pi f_k \hat{T}_{11}} & \frac{1}{\eta_2(k)} \left|\hat{h}_{12}(k)\right| e^{-j2\pi f_k \hat{T}_{12}} \\ \frac{1}{\eta_1(k)} \left|\hat{h}_{21}(k)\right| e^{-j2\pi f_k \hat{T}_{21}} & \frac{1}{\eta_2(k)} \left|\hat{h}_{22}(k)\right| e^{-j2\pi f_k \hat{T}_{22}} \end{pmatrix} \end{aligned} \tag{7}$$

(See equation at bottom of page) where $\eta_i(k)$ represents the diagonal entries of the arbitrary scaling matrix $\Lambda(k)$ in (5). Neglecting reverberation, the TDOA information emerges when taking the ratios of the entries of each column in (7)

$$r_1(k) = \frac{\left|\hat{h}_{11}(k)\right|}{\left|\hat{h}_{21}(k)\right|} e^{-j2\pi f_k \hat{\Delta} t_1}, \quad r_2(k) = \frac{\left|\hat{h}_{12}(k)\right|}{\left|\hat{h}_{22}(k)\right|} e^{-j2\pi f_k \hat{\Delta} t_2} \tag{8}$$

where $\hat{\Delta} t_i$ are the TDOAs of the sources with respect to the microphone pair. As it can be seen from (8), such ratios are invariant to the scaling ambiguities of the estimation process. Since the TDOA information resides only in the phase of the ratios in (8) and is invariant to scaling and magnitude, the ratios can be simplified as

$$\bar{r}_1(k) = \frac{r_1(k)}{|r_1(k)|}, \quad \bar{r}_2(k) = \frac{r_2(k)}{|r_2(k)|} \tag{9}$$

If the permutation of the sources can be somehow corrected and if the mixing does not undergo spatial aliasing, the TDOAs of the sources can be estimated directly from phase information of (9) by exploiting the linear relationship between the TDOAs and the true frequencies along the different bins [9]. However, solving the permutation problem and dealing with spatial aliasing can prove to be difficult in practice. SCT is a method that can sidestep these issues by forming a pseudo-likelihood between the TDOA observations in (8) and a propagation model that can intrinsically account for both permutations and spatial aliasing [13]. The propagation model that results in

TDOA of a source with respect to the microphones, denoted as τ, is assumed to be

$$c(k,\tau) = e^{-j2\pi f_k \tau} \tag{10}$$

The SCT for the configuration of two sources and two microphones is formulated to be

$$SCT(\tau) = \sum_{k} \sum_{m=1}^{2} \left[1 - g\left(\frac{\|c(k,\tau) - \bar{r}_m(k)\|}{2} \right) \right] \tag{11}$$

where the transform is scanned for different values of τ with an arbitrary resolution and $g(.)$ is a function of the Euclidian distance. A good option for $g(.)$ is shown to have a sigmoidal shape such as the following

$$g(\xi) = tanh(\alpha \xi) \tag{12}$$

where α is a real positive constant that defines the inter-source resolution of the spatial likelihood, i.e. the capability of the system to spatially discriminate TDOAs related to different sources and is usually set empirically. The sigmoidal shape gives more emphasis when the observations $\bar{r}_m(k)$ are close to the model $c(k,\tau)$ while ignoring the other values. It can be easily understood from (11) that one could expect to see higher mappings of SCT for values of τ which $\bar{r}_m(k)$ and the model $c(k,\tau)$ are closer in some Euclidian form of distance, thus creating peaks for values of τ matching the TDOAs. One important feature of SCT is that it is invariant to source permutations since it jointly utilizes the TDOA information of all the ratios in (9) across all frequencies. On the other hand, since the model $c(k,\tau)$ incorporates the 2π phase wrap-arounds (i.e. it is periodic for 2π shifts) caused by spatial aliasing, its sensitivity towards spatial aliasing is greatly reduced. Moreover, the most important feature of SCT that makes it an attractive platform for tracking unknown time-varying number of sources is that it is able to map the TDOA peaks for the underdetermined or overcomplete case which involves having more sources than microphones. This is achieved by partitioning the data (STFT frames) into small blocks and performing ICA/SCT on each data block. For example for the case explored so far of two microphones, by exploiting the frequency sparsity of the sources (which is typical of speech) in each data block, and assuming that at each frequency bin and each data block at most two sources are active, a complete TDOA mapping of all the sources (whose number in total can be greater than two) becomes possible. From the far-field assumption, one can convert TDOA detections into DOA using

$$\theta = cos^{-1}\left(\frac{c\Delta t}{\Delta q} \right) \tag{13}$$

where c is the speed of sound and Δq is the distance between the microphone pair.

In case the number of microphones is greater than two, the generalized state coherence transform (GSCT), which is a multi-dimensional extension to SCT, is used. In GSCT each dimension of the domain pertains to a τ_p variable, where p is the index of the microphone pair [13]. In this paper we use two microphones for our experiments, hence the multi-dimensional TDOA mapping using GSCT is not discussed. It is noteworthy to say that even though the SCT propagation model only considers the direct path in an anechoic setting, nonetheless, it is still shown to be effective for multi-path propagation due to reverberation. The reason for this is that in a reverberant environment the direct path between the source and the microphone is usually dominant over other multi-path propagations. As the amount of reverberation increases the chance of multi-paths creating peaks in the SCT increases as well. Consequently, for dealing with unknown time-varying number of sources, as considered in this paper, a suitable filtering technique is needed to reject clutter caused by multi-path propagations.

III. Bayesian Multi-Target Tracking and PHD Filtering

In the previous section we explained how to effectively transform a mixture representation of multiple concurrent sources in the framework of ICA into a detection representation of the source DOAs by extracting significant peaks of the SCT. In the detection framework, hereafter, we will call the SCT peaks that originate from a source as "target", assuming that each source only gives rise to one target. Let's assume that at time t, the sensor makes N_t observations (detections) $z_{t,1}, \ldots, z_{t,N_t}$ each taking values in the state space $\mathcal{Z}$. These detections are ambiguous in the sense that it is not known whether they have originated from targets or are false detections (clutter). Moreover, due to the imperfections in the sensor resolution it is possible that an arbitrary subset of targets do not get detected (missed detections). Our goal is to process such detections in order to reject clutter, account for missed detections, identify the number of sources and track the target states. Now let's consider the multi-target scenario where at time $t-1$ there exist M_{t-1} targets with states $x_{t-1,1}, \ldots, x_{t-1,M_{t-1}}$ taking values in the state space $\mathcal{X}$. At the next instance of time, t, some of the targets can die, some new targets can be born and the surviving targets can evolve according to some dynamic model. This results in M_t targets at time t with states $x_{t,1}, \ldots, x_{t,M_t} \in \mathcal{X}$. Assuming that the respective ordering of the measurements and the state estimates have no significance, the multi-target states and observations can be represented as finite sets such as

$$X_t = \{x_{t,1}, \ldots, x_{t,M_t}\} \in \mathcal{F}(\mathcal{X}) \tag{14}$$

$$Z_t = \{z_{t,1}, \ldots, z_{t,N_t}\} \in \mathcal{F}(\mathcal{Z}) \tag{15}$$

where $\mathcal{F}(\mathcal{X})$ and $\mathcal{F}(\mathcal{Z})$ are finite subsets of the spaces of $\mathcal{X}$ and $\mathcal{Z}$, respectively. By assuming that the multi-target RFS state $X(t)$ is the union of surviving targets, spontaneous births and spawned targets, and the multi-target detection RFS state $Z(t)$ is the union of target-generated detections and clutter, the goal of Mahler's RFS multi-target filtering [16] is to estimate the number of targets and their states while rejecting clutter and accounting for missed detections. The RFS formulation for multi-target Bayesian filtering is the extension of the well known single-target Bayesian filtering which can be computed sequentially via the prediction and update steps as following

$$f_{t|t-1}(X_t|Z_{1:t-1}) = \int f_{t|t-1}(X_t|X_{t-1}) f_{t-1|t-1}(X_{t-1}|Z_{1:t-1}) \delta X_{t-1} \tag{16}$$

$$f_{t|t}(X_t|Z_{1:t}) = \frac{f_{t|t}(Z_t|X_t) f_{t|t-1}(X_t|Z_{1:t-1})}{\int f_{t|t}(Z_t|X'_t) f_{t|t-1}(X'_t|Z_{1:t-1}) \delta X'_t} \quad (17)$$

where $f_{t|t-1}(X_t|Z_{1:t-1})$ is the multi-target predictive density, $f_{t|t}(X_t|Z_{1:t})$ is the multi-target posterior density, $Z_{1:t}$ is the concatenation of all previous measurements up to time t and δ is an appropriate reference measure on $\mathcal{F}(\mathcal{X})$ which indicates that the integrals are set-integrals. A set-integral is a non-trivial extension of a regular integral which is defined as a mixture of regular integrals over all different subsets of the multi-target states. This accounts for the uncertainty in the target number which can vary over time as new targets enter and old ones vanish. The exact definitions of set-integrals and set-derivatives is part of Mahler's Finite Set Statistics (FISST) [16] which provides a systematic calculus-based approach to multi-target filtering using RFSs.

Due to the use of combinatorial set-integrals in the optimal Bayesian recursions of (16), (17), they involve multiple high dimensional integrals on the space $\mathcal{F}(\mathcal{X})$ rendering it computationally intractable. The PHD filter is a suboptimal approximation to the mutli-target Bayesian recursions of (16), (17) which instead of propagating the full posterior density, propagates the FISST-based first moment of multi-target posterior density, known as the posterior intensity [16], [17]. This is analogous to the well known constant-gain Kalman filter in single-target tracking, which also only propagates the first moment (mean) of the target.

Let $D_{t|t-1}(x_t|Z_{1:t-1})$ and $D_{t|t}(x_t|Z_{1:t})$ denote the respective PHD intensities of the multi-target predictive posterior $f_{t|t-1}(X_t|Z_{1:t-1})$ and the multi-target posterior $f_{t|t}(X_t|Z_{1:t})$ of (16), (17). It is worthy to note that due to the first order moment mapping of the PHD filter, the finite set-valued random variable state $X_t \in \mathcal{F}(\mathcal{X})$ of the multi-target posterior is represented by an ordinary random variable $x_t \in \mathcal{X}_0$ with dimensions pertaining to the dimensions of a single target, i.e. $D_{t|t}(x_t|Z_{1:t})$ is an intensity function on the single target space $\mathcal{X}_0$. This PHD intensity function is not in the form of a probability density function (pdf) as its integral does not equate to unity. Under certain assumptions and using FISST [16], [17], the PHD intensities can be recursively estimated as follows

$$D_{t|t-1}(x_t|Z_{1:t-1}) = b_t(x_t) + \int F_{t|t-1}(x_t|x_{t-1}) D_{t-1|t-1}(x_{t-1}|Z_{1:t-1}) dx_{t-1} \quad (18)$$

$$D_{t|t}(x_t|Z_{1:t}) = [1 - p_D(x_t)] D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{z_t \in Z_t} \frac{\psi_{z_t}(x_t) D_{t|t-1}(x_t|Z_{1:t-1})}{\kappa_t(z_t) + \int \psi_{z_t}(\zeta) D_{t|t-1}(\zeta|Z_{1:t-1}) d\zeta} \quad (19)$$

In the prediction (18)

$$F_{t|t-1}(x_t|x_{t-1}) = p_S(x_{t-1}) f_{t|t}(x_t|x_{t-1}) + \beta_{t|t-1}(x_t|x_{t-1}) \quad (20)$$

where $f_{t|t}(x_t|x_{t-1})$ is the single target transition pdf, p_S is the probability of target survival and $\beta_{t|t-1}$ is the intensity of target spawned from targets at time $t-1$. Also in (18), b_t is the intensity of spontaneous new births at time t. In the update (19),

$$\psi_{z_t}(x_t) = p_D(x_t) g(z_t|x_t) \quad (21)$$

where p_D is the probability of detection, $g(z_t|x_t)$ is the single target detection likelihood model (i.e. observation model in the space of $\mathcal{X}_0$) and the intensity of clutter points $\kappa_t(z_t)$ is given as

$$\kappa_t(z_t) = \lambda c_t(z_t) \quad (22)$$

where λ is the average number of Poisson-distributed false alarms and $c_t(z)$ is the spatial distribution of clutter. As we mentioned before the PHD intensity function is not a pdf and in fact it turns out that the integral of the PHD intensity gives the expected number of targets as follows [16]

$$\hat{M}_{t|t} = \int D_{t|t}(x_t|Z_{1:t}) dx_t \quad (23)$$

At the end, the state estimates for each target are extracted by finding the $\hat{M}_{t|t}$ peaks of intensity $D_{t|t}(x_t|Z_{1:t})$. In the case where only a single target is present, the formulations above reduces to the constant-gain Kalman filter.

Even though the PHD filter is much less computationally expensive compared to the multi-target recursions of (16), (17), due to the fact that it operates in the space of a single target $\mathcal{X}_0$, the integrals present in the PHD recursions of (18), (19) result in it not having a closed form solution in general. Therefore, Sequential Monte Carlo (SMC) methods are usually used to approximate the integrals in general [31]. However, for the special case where the target dynamics follow a linear Gaussian Model, a Gaussian mixture (GM) implementation can provide a closed form solution to the PHD filter [18]. The GM-PHD does not suffer from the complexities of sampling and resampling in SMC methods and due to its closed form solution, it is more accurate. In this paper, since it is reasonable to assume that our measurements and target state dynamics follow a linear/Gaussian model, GM-PHD is used for the multi-source filtering.

A. GM-PHD Implementation

Assuming that the target dynamics and sensor model follows a linear/Gaussian form, we have

$$f_{t|t-1}(x_t|x_{t-1}) = \mathcal{N}(x_t; A_{t-1}x_{t-1}, Q_{t-1}) \quad (24)$$

$$g(z_t|x_t) = \mathcal{N}(z_t; B_t x_t, R_t) \quad (25)$$

where $\mathcal{N}(.; a, C)$ denotes a Gaussian pdf with mean a and covariance C, A_{t-1} is the state transition matrix, B_t is the observation matrix, Q_{t-1} is the transition process noise covariance and R_t is the observation noise covariance. The GM-PHD requires that the survival and detection probabilities be state independent, therefore $p_S(x_t) = p_S$ and $p_D(x_t) = p_D$. Another assumption is that the birth and spawn intensities are Gaussian mixtures [18]. For simplicity we neglect the spawning of new targets from previous targets and just rely on spontaneous births to model new targets. Therefore we have

$$b_t(x_t) = \sum_{i=1}^{J_{b,t}} \omega_{b,t}^{(i)} \mathcal{N}\left(x_t; m_{b,t}^{(i)}, P_{b,t}^{(i)}\right) \quad (26)$$

where $J_{b,t}$, $\omega_{b,t}^{(i)}$, $m_{b,t}^{(i)}$, $P_{b,t}^{(i)}$, $i = 1, \ldots, J_{b,t}$, are given model parameters that determine the shape of the birth intensity. Usually one adapts these parameters to model regions in the state space which correspond to detection persistences. Again, we make note that (26) is not a pdf, in general. That is because there is no restriction on the coefficients $\omega_{b,t}^{(i)}$, $i = 1, \ldots, J_{b,t}$, adding to unity.

Assuming that the posterior PHD at time $t-1$ is a Gaussian mixture of the form

$$D_{t-1|t-1}(x_{t-1}|Z_{1:t-1}) = \sum_{i=1}^{J_{t-1}} \omega_{t-1}^{(i)} \mathcal{N}\left(x_{t-1}; m_{t-1}^{(i)}, P_{t-1}^{(i)}\right), \tag{27}$$

then the predicted intensity at time t is a Gaussian mixture given by

$$D_{t|t-1}(x_t|Z_{1:t-1}) = b_t(x_t) + p_{S,t} \sum_{j=1}^{J_{t-1}} \omega_{t-1}^{(j)} \mathcal{N}\left(x_t; m_{S,t|t-1}^{(j)}, P_{S,t|t-1}^{(j)}\right) \tag{28}$$

where

$$m_{S,t|t-1}^{(j)} = A_{t-1} m_{t-1}^{(j)} \tag{29}$$

$$P_{S,t|t-1}^{(j)} = Q_{t-1} + A_{t-1} P_{t-1}^{(j)} A_{t-1}^T \tag{30}$$

As the predicted intensity for time t can be rearranged to have a Gaussian mixture of the form

$$D_{t|t-1}(x_t|Z_{1:t-1}) = \sum_{i=1}^{J_{t|t-1}} \omega_{t|t-1}^{(i)} \mathcal{N}\left(x_t; m_{t|t-1}^{(i)}, P_{t|t-1}^{(i)}\right), \tag{31}$$

the posterior intensity at time t also becomes a Gaussian mixture as follows

$$D_{t|t}(x_t|Z_{1:t}) = (1 - p_{D,t}) D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{z_t \in Z_t} D_{D,t}(x_t; z_t) \tag{32}$$

where

$$D_{D,t}(x_t; z_t) = \sum_{j=1}^{J_{t|t-1}} \omega_t^{(j)}(z_t) \mathcal{N}\left(x_t; m_{t|t}^{(j)}(z_t), P_{t|t}^{(j)}(z_t)\right) \tag{33}$$

$$\omega_t^{(j)}(z_t) = \frac{p_{D,t} \omega_{t|t-1}^{(j)} q_t^{(j)}(z_t)}{\kappa_t(z_t) + p_{D,t} \sum_{l=1}^{J_{t|t-1}} \omega_{t|t-1}^{(l)} q_t^{(l)}(z_t)} \tag{34}$$

$$q_t^{(j)}(z_t) = \mathcal{N}\left(z_t; B_t m_{t|t-1}^{(j)}, R_t + B_t P_{t|t-1}^{(j)} B_t^T\right) \tag{35}$$

$$m_{t|t}^{(j)}(z_t) = m_{t|t-1}^{(j)} + K_t^{(j)}\left(z_t - B_t m_{t|t-1}^{(j)}\right) \tag{36}$$

$$P_{t|t}^{(j)}(z_t) = \left[I - K_t^{(j)} B_t\right] P_{t|t-1}^{(j)} \tag{37}$$

$$K_t^{(j)} = P_{t|t-1}^{(j)} B_t^T \left(B_t P_{t|t-1}^{(j)} B_t^T + R_t\right)^{-1} \tag{38}$$

Similar to the Gaussian sum filter in single target tracking [32], as time progresses, the number of Gaussian components in GM-PHD increases without bound. To fix this, a simple pruning and merging technique can be used to limit the growth of number of Gaussians [18]. It works by discarding Gaussians whose weight $\omega_t^{(i)}$, $i = 1, \ldots, J_t$ falls below some threshold and then normalizes the weights of the surviving Gaussians so that the sum of the weights, which from (23) is the expected number of targets, remains the same. Then it uses a Mahalanobis distance measure to merge Gaussians that are close to each other.

Once the expected number of targets $\hat{M}_{t|t}$ is found, estimating the multi-target states at first glance appears to be straightforward since the peaks in the posterior intensity $D_{t|t}(x_t|Z_{1:t})$ correspond to the means of the Gaussians, given that they are well-separated. However, since the height of peaks in the posterior intensity depends on both weight and covariance, selecting the $\hat{M}_{t|t}$ highest peaks may result in state estimates that correspond to Gaussians with weak weights. This is not desirable since the expected number of targets due to these peaks is small even though the magnitudes of the peaks are large. A better alternative is to select the means of the Gaussians with weights greater than some threshold, say 0.5 [18].

B. Data Association Using the GM-PHD

The PHD filter, in itself, does not solve the data association problem, therefore one cannot tell which state estimates belong to which target. However, by associating tags to the mixture components of the GM-PHD filter, a data association scheme can be utilized providing us with distinct tracks on the sources. The tag labeling steps for GM-PHD filter with track management is as follows [19]:

1) Initialization: At time $t = 0$, J_0 Gaussians are distributed across the state space to form the intensity

$$D_0(x_t) = \sum_{j=1}^{J_0} \omega_0^{(j)} \mathcal{N}\left(x_t; m_0^{(j)}, P_0^{(j)}\right) \tag{39}$$

A unique tag is assigned to each Gaussian to form the set

$$\mathcal{T}_0 = \left\{\Upsilon_0^{(1)}, \ldots, \Upsilon_0^{(J_0)}\right\} \tag{40}$$

2) Prediction: After predicting forward the PHD intensity, Gaussians associated with new births receive new tags and Gaussians that are associated with surviving ones retain previous tags, i.e. the set of tags is as follows

$$\mathcal{T}_{t|t-1} = \mathcal{T}_{t-1|t-1} \cup \left\{\Upsilon_{b,t}^{(1)}, \ldots, \Upsilon_{b,t}^{(J_{b,t})}\right\} \tag{41}$$

where $\Upsilon_{b,t}^{(j)}$ is the j^{th} new tag associated with the spontaneous birth intensity in (26) and $\mathcal{T}_{t-1|t-1}$ contains the tags of targets at time $t-1$ such that the predicted Gaussian with mean $m_{S,t|t-1}^{(j)}$ in (28), (29) retains the tag of the Gaussian with mean $m_{t-1|t-1}^{(j)}$.

3) Update: The predicted intensity is updated according to (32). Hence, each Gaussian component of the predicted intensity gives rise to $1 + |Z_t|$ components in the updated intensity, where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. Hence, the tags of the predicted Gaussian components get propagated to the updated Gaussian components, i.e. the Gaussian with mean $m_{t|t}^{(j)}(z_t)$ in (33) gets the same tag that the Gaussian with mean $m_{t|t-1}^{(j)}$ had in (31).

4) Pruning and Merging: At this step the tags of the Gaussians that get pruned vanish. For the Gaussian that are merged, the tag of the one with the largest weight is retained.

5) Multi-Target State Estimation: At this step the means and tags of the Gaussians with weights higher than the aforementioned threshold (see end of Section III-A) are reported as state targets and track labels, respectively. Hence, there is an identifying tag associated with each estimate. If the target is a new born one, it has a new tag and if it is a surviving target it retains its previous track label.

C. Incorporating Amplitude Information in the PHD Likelihood

In target tracking applications, the detection step consists of extracting local peaks in the observations that are higher than some certain threshold. These detection points either come from targets or from clutter. The standard PHD filter discussed treats all detections equally and relies on the track continuity of the targets to reject clutter. However, in most cases the amplitude of detections generated from targets are higher than clutter and carry reliability information. This information about the amplitudes can be incorporated in the PHD tracking algorithm to further assist the discrimination of targets from clutter [33]. This is done by introducing an augmented measurement vector $\bar{z}_t = [z_t^T a]^T$, where $a \geq 0$ is the detection amplitude. Assuming that the amplitudes are independent of the target states, the respective target and clutter likelihood functions ψ_{z_t} and $\kappa_t(z_t)$ in (19) are modified to become

$$\bar{\psi}_{\bar{z}_t}(x_t) = g(z_t|x_t)g_a(a|d) \tag{42}$$

$$\bar{\kappa}_t(\bar{z}_t) = \lambda c_t(z_t)c_a(a) \tag{43}$$

where d is related to the signal-to-noise ratio (SNR), i.e. the ratio between target amplitude and clutter amplitude. SNR is defined in the log scale as

$$\text{SNR(dB)} = 10log_{10}(1+d) \tag{44}$$

and is assumed to be the same for all targets. For the case of $d = 0$ the amplitude of the targets and clutter become the same, hence it is reduced to the standard PHD filter. In (42), (43), $g_a(a|d)$ and $c_a(a)$ are the amplitude likelihood densities for targets and clutter, respectively. Assuming that the detection threshold is τ_o in which all peaks above τ_o are reported, the amplitude likelihoods for measurements that exceed τ_o are denoted as $g_a^{\tau_o}(a|d)$ and $c_a^{\tau_o}(a)$. Hence, due to normalization we have

$$g_a(a|d) = g_a^{\tau_o}(a|d)p_D^{\tau_o}(d) \tag{45}$$

$$c_a(a) = c_a^{\tau_o}(a)p_{FA}^{\tau_o} \tag{46}$$

where

$$p_D^{\tau_o}(d) = \int_{\tau_o}^{\infty} g_a(a|d)da \tag{47}$$

$$p_{FA}^{\tau_o} = \int_{\tau_o}^{\infty} c_a(a)da \tag{48}$$

are the probability of detection and probability of false alarm, respectively. By incorporating the amplitude likelihoods, the PHD update of (19) becomes

$$D_{t|t}(x_t|Z_{1:t}) = [1 - p_D^{\tau_o}(d)]\, D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{\bar{z}_t \in \bar{Z}_t} \frac{\bar{\psi}_{\bar{z}_t}(x_t)D_{t|t-1}(x_t|Z_{1:t-1})}{\bar{\kappa}_t(\bar{z}_t) + \int \bar{\psi}_{\bar{z}_t}(\zeta)D_{t|t-1}(\zeta|Z_{1:t-1})d\zeta} \tag{49}$$

For the case of known d, it is common to model the amplitude likelihoods with Rayleigh distributions

$$g_a^{\tau_o}(a|d) = \frac{a}{1+d}\exp\left(\frac{\tau_o^2 - a^2}{2(1+d)}\right),$$

$$p_D^{\tau_o}(d) = \exp\left(\frac{-\tau_o^2}{2(1+d)}\right) \tag{50}$$

$$c_a^{\tau_o}(a) = a\exp\left(\frac{\tau_o^2 - a^2}{2}\right),$$

$$p_{FA}^{\tau_o} = \exp\left(\frac{-\tau_o^2}{2}\right) \tag{51}$$

Note that the Rayleigh parameter for the clutter model in (51) is assumed to be unity which might not be true in general. However, given that the clutter level is known, the amplitudes of the detections can be scaled so that the parameter for the clutter Rayleigh distribution becomes unity while the parameter for the target Rayleigh distribution conforms with the SNR level corresponding to $(1+d)$. On the other hand, for the case of unknown d, one can marginalize (50) over a range of possible values $[d_1\ d_2]$ and find a distribution for g_a that is not conditional on d, hence

$$g_a(a) = \int_{d_1}^{d_2} p(\gamma)g_a(a|\gamma)d\gamma \tag{52}$$

$$p_D^{\tau_o} = \int_{d_1}^{d_2} p(\gamma)p_D^{\tau_o}(\gamma)d\gamma \tag{53}$$

By picking a suitable prior distribution $p(d)$ and assuming $g_a(a|d)$ is Rayleigh distributed with parameter $(d+1)$, one can obtain a closed form solution to (52). The probability of detection $p_D^{\tau_o}$ in (53) can then be found using numerical integration offline since it does not need to be computed for every iteration [33].

IV. System Integration

A. Tracking Task

In the previous two sections we described the front-end (ICA/SCT) and the back-end (PHD filtering) of our system model, respectively. The front-end uses the output of ICA to perform

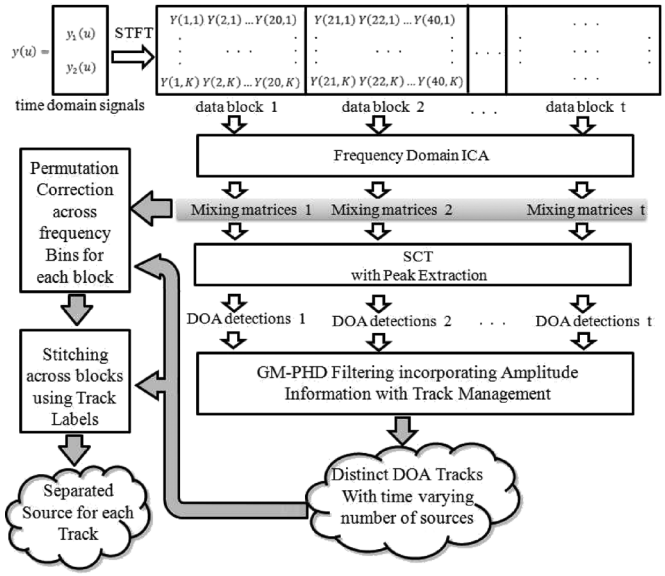

Fig. 1. Block diagram of proposed method: STFT, ICA and SCT segments form the front-end and the PHD filtering segment form the back-end. The feedback from the back-end to the front-end describes the separation task which uses the distinct estimated tracks to perform permutation correction across frequencies for each block and to stitch together the separated components from one block to another.

the SCT mapping where peaks that are above some detection threshold are selected. These peaks are declared as DOA measurements or detections and are fed into the PHD filter. The PHD filter then filters the measurements and estimates the DOA and number of targets using the GM-PHD filter assuming that the state dynamics and sensor model for a single source follow a linear/Gaussian model according to (24), (25) so that

$$x_t = A_{t-1}x_{t-1} + \nu_{t-1} \tag{54}$$

$$z_t = B_t x_t + \vartheta_t \tag{55}$$

where $\nu_t \sim \mathcal{N}(0, Q_t)$, $\vartheta_t \sim \mathcal{N}(0, R_t)$ and x_t is the vector of states at time t for a single target model. The dimensions of x_t depends on the number of microphone pairs since each microphone pair has a separate TDOA. Also information about the velocity information of the DOA (denoted as $\dot{\theta}_t$) can be incorporated in the state to represent a constant velocity model. In the most simple case where only one microphone pair is present $(L = 2)$ and the velocity information is not considered, the state reduces to $x_t = \theta_t$ with a dimension of unity and the model parameters in (54), (55) reduce to $A_{t-1} = 1$ and $B_t = 1$.

Fig. 1 illustrates the system model incorporating ICA/SCT with PHD filtering. As depicted in Fig. 1, ICA is performed on blocks of data in which each block is a collection of a certain number of STFT frames. Note that the time index of the sensor raw data is u, the frame index after converting to the frequency domain using STFT is n and the block index for a collection of frames is t. Any complex-valued ICA algorithm can be used on the blocks. An important note is that the initialization of the ICA iterations for each block should be done from scratch and not based on the previous block converged values. This is to encourage diversity in the ICA estimates so that if a source dies out or a new source is born, such dynamics can be picked up by ICA and translated to meaningful location information via SCT. To better distinguish between clutter and targets, the GM-PHD filter incorporates the detection amplitudes as described in Section III-C. Also, the GM-PHD tracker enables data association and track labels as described in Section III-B. The track labeling is crucial for the separation task, since the track labels will be used to stitch together the ICA outputs from the blocks enabling a separated source for each track.

B. Separation Task

The key notion that allows us to perform the separation task even for the overcomplete/underdetermined case is assuming block-frequency sparsity of the sources in which the number of source components at each frequency-block segment does not exceed the number of sensors even though the overall number of sources can exceed the number of sensors. However, the estimated mixing matrices, i.e. the immediate output of the ICA stage, contain no valuable separated information of the mixed sources. This is due to the fact that at that stage no inference on number of sources is achieved and the ordering of the columns of the mixing matrices across the frequency bins and time blocks are indeterminate. Since, SCT is invariant to such mismatch in ordering, it is able to translate the mixture model of ICA into a detection model similar to that commonly used in radar/sonar (hence the use of the term 'sources' in the mixture model and 'targets' in the detection model) and from there the GM-PHD filter determines the expected number of sources and provides distinct tracks on the DOA of the sources. Now one can use this information obtained from the output of the PHD filter and feed it back to the output of the ICA stage to effectively carry out the separation task in the following two steps:

1) Permutation Correction in Each Block Across Frequencies: The expected number of sources and the estimates of TDOAs obtained from the PHD filter is used to correct for possible permutations. Let's consider the case where we have L sensors and at time block t the PHD filter has declared $\hat{M}_{t|t} \geq L$ sources to be active with corresponding TDOAs $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{M}_{t,t}}$. The mixing matrix in each bin for block t has dimensions $L \times L$. From the sparsity assumption, this means that at block t, each frequency bin has at most L columns of its mixing matrix that are active, however at this stage, the ordering of which L out of $\hat{M}_{t|t}$ sources being active is not known. We introduce $\hat{M}_{t|t} - L$ virtual null columns of zeros to represent inactivity. Therefore, for each frequency bin there are a total of $\hat{M}_{t|t}!/(\hat{M}_{t|t} - L)!$ possible permutations of the mixing matrix columns with augmented null columns. We can now use the PHD filter estimates of the TDOA as reference to align the columns of the null augmented mixing matrix for frequency bins $k = 1, \ldots, K$ as following (similar to [34])

$$\bar{\Pi}(k) = \arg\min_{\Pi} \sum_{m=1}^{\hat{M}_{t|t}} \|c(k, \hat{\tau}_m) - \bar{r}_{\Pi_m}(k)\| \tag{56}$$

where Π is a permutation of the mixing matrix as described in (5) and $\bar{r}_{\Pi_m}(k)$ is the normalized ratio of the m^{th} column of the matrix affected by permutation Π as described in (9). We make the note that for the null columns, the value of such ratios are not defined and one can replace them with a single constant as it does not effect the minimization of (56). Once the ordering

of the sources is estimated from (56), one can perform the separation in the time block for each frequency bin by rearranging the rows of the demixing matrix (inverse of the aforementioned $L \times L$ mixing matrix) to align with their corresponding L active source components and forcing the remainder of inactive source components to zero. Next step would be to determine whether the separated sources at the current block t are newborn sources or surviving ones, and if surviving, to stitch it to the corresponding segments of the same source from the previous block $t-1$.

2) Stitching Segments Across Blocks: In the previous step we explained how for each time block, the mixing matrix for each frequency bin can be aligned so that each column is linked to a single DOA obtained from the PHD filter. In this step we explain how the components from one time block are stitched to the components from the previous block. If the DOAs of the sources do not undergo any dynamics, then one can use the DOAs themselves to link the ICA components of one block to the previous block. In the case where the DOAs undergo dynamics in terms of both values and birth/death occurrences, then some kind of data association scheme is required to link the DOAs of surviving sources and initiate a new track for newborn sources. The track labeling algorithm described in Section III-B using the GM-PHD implementation effectively accomplishes the task of data association, therefore enabling the stitching of sources from one block to another. We note that in such a separation scheme, any newborn source is declared as new source even though, for example, it might be coming from a previous source that underwent a silence period. The feedback arrows in Fig. 1 illustrate the separation task where the PHD tracks are used to go back and perform the alignment of the mixing matrices across the frequencies and the stitching of the source components across the blocks. At the end, in order to regularize the scaling ambiguity of the ICA outputs we use the well-known minimal distortion principle [35] for each block and frequency bin. Once the stitching and scaling of the ICA outputs are performed, the inverse Fourier transform using the overlap add method is used to reconstruct the time domain signals.

V. Experimental Results

In this section we present some experiments on simulated data for both tasks of interest: Online tracking and separation of multiple moving sources with birth/death dynamics. The simulated data was obtained using Lehmann's image method [36] which simulates the impulse response between a source and a sensor for a rectangular room environment. For each task we use a different experimental set-up since each task has a different level of difficulty with the separation task being more difficult in general than the tracking task. Thus we try to introduce experiments so that it would push the complexity envelope for each task in its own context independently.

A. Tracking Results

We evaluate the performance for the DOA tracking task and compare the proposed method mainly with the two alternative approaches discussed earlier. One method uses the GCC-PHAT at the front-end to acquire detection measurements and the same PHD filter as the proposed back-end for the filtering. The other

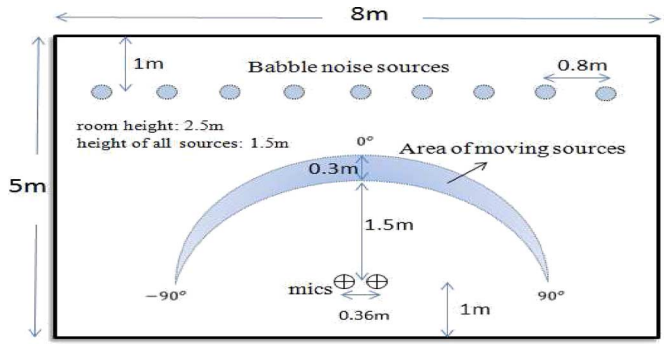

Fig. 2. Room set-up (not drawn to scale). Note that the source trajectories are not shown but rather the area of motion is illustrated. The reason for this is that their activities are time-varying. Refer to Fig. 3 for their true activities and trajectories in terms of DOA.

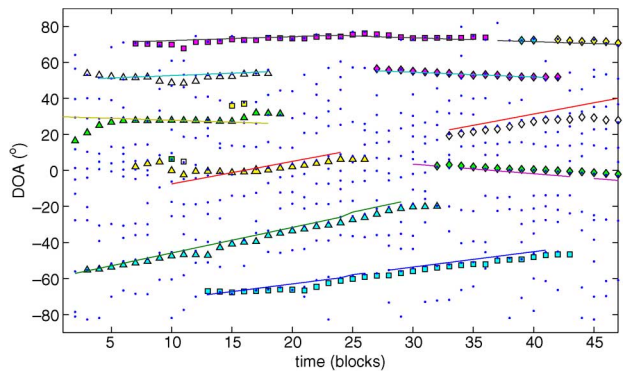

Fig. 3. Proposed method: front-end (ICA/SCT) + back-end (GM-PHD with amplitude information). True DOA (colored lines), SCT peaks (dots) and estimated DOA tracks (colored shapes).

method uses the same ICA/SCT of the proposed front-end and a naive thresholding method to post-process the detections in the back-end. For the first experiment, the room dimensions used in the simulation are 8 m $\times$ 5 m $\times$ 2.5 m with a reverberation time of $T_{60} = 600$ ms. Signals were sampled at $f_s = 16$ kHz and the STFT frequency-frame segments were obtained using a Hanning window of size 2048 samples with 87.5% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being 0.64 seconds (40 frames) in length. The experiment lasted for a total duration of 15.04 seconds. Only $L = 2$ microphones were used which were placed 36 cm apart. The speakers could appear and disappear at any time. There were a total of 7 different speakers with the maximum number of 6 concurrent speakers in this experiment. The speakers all moved along a semi-circular path about 1.5 m from the microphone pair as depicted in Fig. 2. The ICA algorithm carried out for each block was a standard complex valued maximum likelihood Infomax algorithm [10]. Fig. 3 shows the DOA detections and the true source DOAs along with their estimated tracks using the proposed method: front-end (ICA/SCT) + back-end (GM-PHD with amplitude information). The tag or label for each track is represented using a unique colored shape. Fig. 4 illustrates the results using the "GCC-PHAT + proposed back-end" approach (similar to [4], [20]) while Figs. 5 and 6 show the results using two variations of the "proposed front-end + naive thresholding" method ([21], [22]). The first variation uses a higher detection threshold compared to the second variation, hence resulting in fewer clutter but bearing the risk of more

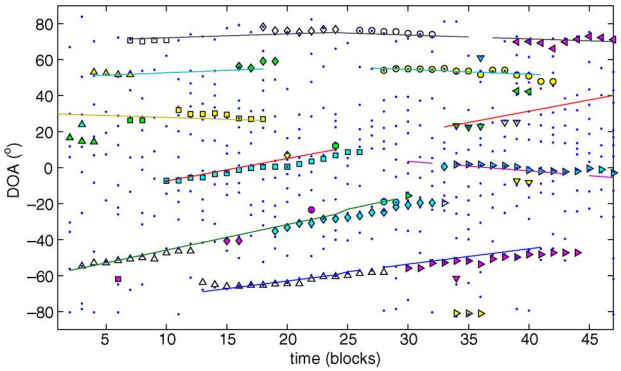

Fig. 4. GCC-PHAT + proposed back-end: True DOA (colored lines), GCC-PHAT peaks (dots) and estimated DOA tracks (colored shapes).

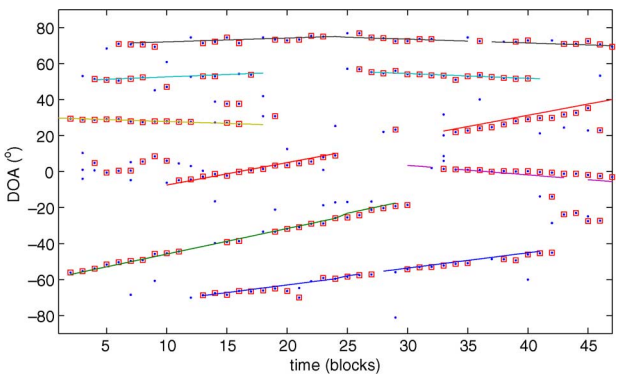

Fig. 5. Proposed front-end + naive thresholding (high): True DOA (colored lines), SCT peaks after thresholding (dots) and selected peaks (squares).

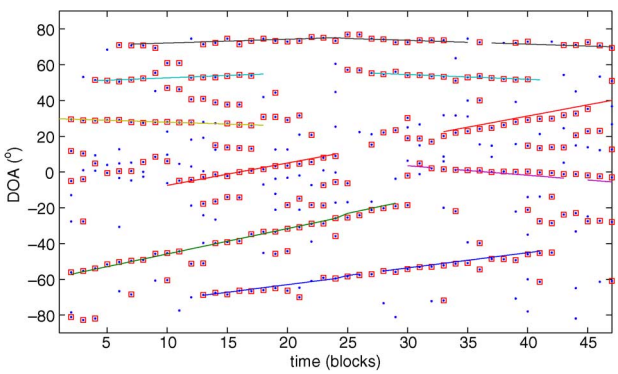

Fig. 6. Proposed front-end + naive thresholding (low): True DOA (colored lines), SCT peaks after thresholding (dots) and selected peaks (squares).

missed detections. In contrast the second variation results in more clutter but with less missed detections. In addition to [21], [22], we also disregard non-persistent peaks using some distance measure in order to reject isolated clutter. We note that the "proposed front-end + naive thresholding" method is not a tracking technique (but a peak selection scheme), and thus does not solve the data association problem inherently and requires an additional module to do so. That is why the estimates in Figs. 5 and Fig. 6 are not color-shape coded. In order to highlight the importance of incorporating amplitude information in our method, we also run our proposed method without considering any amplitude information: "proposed front-end + GM-PHD without considering amplitude information" and show the result in Fig. 7.

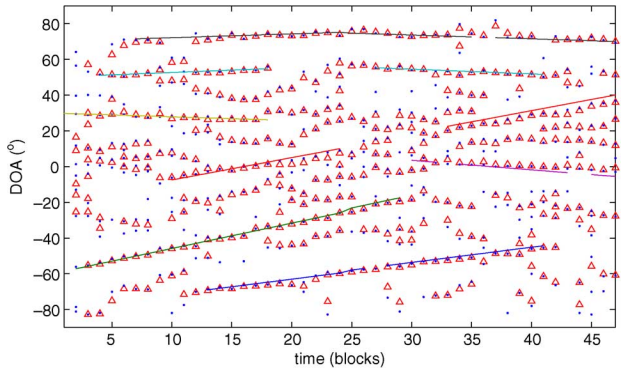

Fig. 7. Proposed front-end + GM-PHD filtering without considering amplitude information: True DOA (colored lines), SCT peaks (dots) and DOA estimates (triangles).

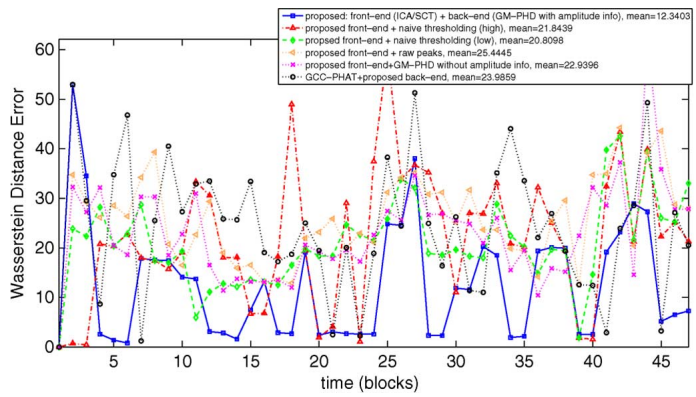

Fig. 8. Wasserstein miss distance for different methods of Figs. 3–7, maximum 6 concurrent sources, $T_{60} = 600$ ms.

Wasserstein miss distance which is an optimal multi-target error metric for time-varying number of targets is used to evaluate the performances of such experiments [37]. Wasserstein miss distance is optimal in the sense that it intrinsically considers the mismatch in target number and state values. Assuming that $X_t = \{x_1, \ldots, x_n\}$ is the subset of true states at time t and $\hat{X}_t = \{\hat{x}_1, \ldots, \hat{x}_m\}$ is the estimated subset of states, the Wasserstein miss distance is defined as

$$d(X_t, \hat{X}_t) = \min_{C} \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} C_{i,j}\|x_i - \hat{x}_j\|^2}, \qquad (57)$$

where the minimum is taken over all $n \times m$ transportation matrices $C = \{C_{i,j}\}$. An $n \times m$ matrix C is a transportation matrix if for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$

$$C_{i,j} \geq 0, \quad \sum_{i=1}^{n} C_{i,j} = \frac{1}{m}, \quad \sum_{j=1}^{m} C_{i,j} = \frac{1}{n}. \qquad (58)$$

The minimization in (57) means that it gives the distance for the best association between true and estimated set of states, and can be done using standard linear programming algorithms. For the aforementioned experiment with results illustrated using the different methods in Figs. 3–7, we present the point-wise and mean-valued Wasserstein distance in Fig. 8. In addition to the methods discussed earlier, Fig. 8 also shows the Wasserstein distance using the raw peaks of the proposed front-end as the

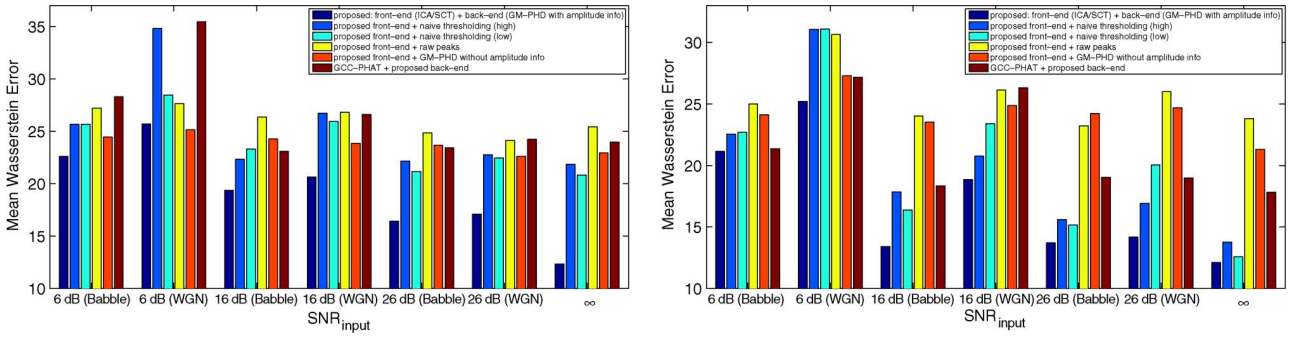

Fig. 9. Performance evaluation for different noise values/types for maximum 6 concurrent sources. Left: $T_{60} = 600$ ms, Right: $T_{60} = 300$ ms.

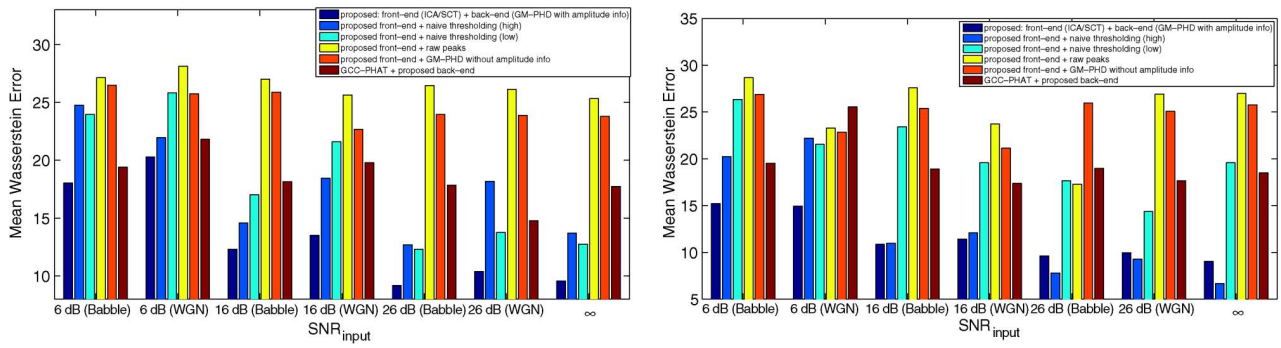

Fig. 10. Performance evaluation for different noise values/types for maximum 4 concurrent sources. Left: $T_{60} = 600$ ms, Right: $T_{60} = 300$ ms.

DOA estimates. For this experiment, Fig. 8 shows that the proposed algorithm outperforms the other methods. In order to get a better quantification of the robustness and versatility of the proposed method compared to the other methods, the mean Wasserstein distance error is computed for other experiments varying the input signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{input}}$[1]), T_{60} and maximum number of concurrent speakers. Two different noise types were considered. One being additive white Gaussian noise (WGN) and the other being Babble noise. Babble noise was simulated using 9 speakers speaking from 9 different locations distributed across the room, at least 3 m away from microphone pair, as depicted in Fig. 2. The probability of activity of each one of these babble sources was 80%. The $\mathrm{SNR}_{\mathrm{input}}$ was computed as follows

$$\mathrm{SNR}_{\mathrm{input}}(\mathrm{dB}) = 10log_{10}\left(\frac{\sum_{i=1}^{2}\sum_{u}|y_{i,target}(u)|^2}{\sum_{i=1}^{2}\sum_{u}|y_{i,noise}(u)|^2}\right) \quad (59)$$

where $y_{i,target}(u)$ and $y_{i,noise}(u)$ are the microphone inputs due to the target sources and the noise (additive WGN or babble sources), respectively. Two different reverberation times $T_{60} = 600$ ms and 300 ms were considered along with two different scenarios with maximum number of concurrent sources being 6 and 4. The trajectories for the maximum 6 concurrent sources being the same as those depicted in Fig. 3 and the trajectories for the maximum 4 concurrent sources being the "blue", "dark green", "green", "magenta" and "cyan" solid lines of Fig. 3. The results of the experiments for all such different variations in input noise values/types, T_{60} and maximum number of concurrent speakers are presented in Figs. 9 and 10. These figures show that the proposed method outperforms the other methods for most scenarios. As the problem becomes least challenging, for example when $T_{60} = 300$ ms with maximum 4 concurrent speakers and $\mathrm{SNR}_{\mathrm{input}} \geq 26$ dB (Fig. 10-Right), the method that uses the "proposed front-end + naive thresholding (high)" performs slightly better compared to the proposed method. In such scenarios the amplitudes of the SCT peaks originating from clutter and targets are more discriminative/separable compared to the more challenging scenarios. Thus by choosing the right threshold, one could effectively reject clutter while keeping the peaks belonging to the targets. As a result, simple peak selection methods, like that of naive thresholding (high), can perform pretty well in such cases and sometimes even perform better compared to multi-target tracking methods. The reason the better performance in such scenarios is that multi-target tracking methods can lose some accuracy in the track initiation and termination (can lag behind 1–3 time updates when initiating and terminating a track), while peak selection methods can promptly identify a target's initiation and termination given that the threshold is correct and the amplitudes of targets and clutter are separable.

We note that the parameters d_1 and d_2 in (52) which characterize the amplitude information of the targets in the proposed back-end and the thresholding parameters for the naive thresholding methods were fixed and obtained by inspection based on the physical properties such as the room's T_{60}, STFT frame window size and percentage overlap. Also, as explained in Section III-C, it is assumed that the clutter level is known and amplitudes of the detections are scaled so that the parameter for the clutter Rayleigh distribution in (51) becomes unity. The assumed clutter level effects the sensitivity of the algorithm: the lower it is, the more likely the algorithm is in declaring a detection as a target and vice versa. Moreover, for all experiments the

[1]note that $\mathrm{SNR}_{\mathrm{input}}$ is different than the SNR in (44) which defines the detection amplitude of the targets compared to clutter

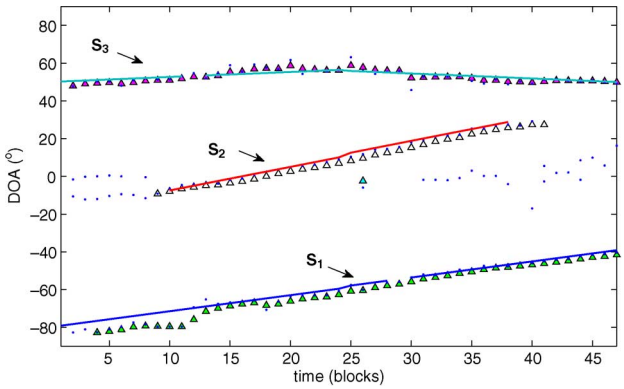

Fig. 11. Separation experiment with 3 unknown/time-varying sources and 2 microphones, $T_{60} = 200$ ms: True DOA (colored lines), SCT peaks (dots) and estimated tracks (colored shapes).

average number of Poisson-distributed false alarm was $\lambda = 10$, the probability of detection was $p_D = 0.5$, the probability of target survival was $p_S = 0.99$, birth rate was $\sum_i \omega_{b,t}^{(i)} = 0.1$, the process and observation noise covariances Q_t and R_t were both set to $10I$, where I is the identity matrix.

B. Separation Results

In this section the separation capabilities of the proposed method is investigated. The room dimensions used in all the simulations in this section were 6 m × 4 m × 2.5 m and the signals were sampled at $f_s = 16$ kHz. We experimented with a total of four scenarios and increased the level of difficulty with each scenario. For the first scenario a reverberation time of $T_{60} = 200$ ms was considered. The STFT frequency-frame segments were obtained using a Hanning window of 1024 taps with 75% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being 0.64 seconds (40 frames) in length. Again, only $L = 2$ microphones were used which were placed 6 cm apart. Since separating the sources is our focus here, we use the recursively regularized ICA algorithm [38] for our ICA module as it has been shown to achieve better separation results for short duration mixtures when compared to regular infomax ICA. A total of 3 speakers were involved in this experiment. Speaker 1 and 3 are active for the total experiment which lasted for 15.04 seconds. Speaker 2 enters the conversation at around the 3 second mark and leaves the conversation at around the 11 second mark. All three speakers moved along a semi-circular path 1 m from the microphone pair, i.e. with radii $r_i = 1$ m, $i = 1, \ldots, 3$. Fig. 11 illustrates the DOA detections and the true source DOAs along with the estimated tracks for all three speakers. The PEASS toolbox which is a perceptual BSS evaluation software, was used for the performance evaluation of the separated sources [39]. The signal to disturbance ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifact ratio (SAR) metrics with the decomposition estimated by the PEASS tool was used for the evaluation of our separated sources. At first no assumption on the number of sources was made. As seen in Fig. 11, the proposed method was able to identify the correct number of sources at the appropriate intervals. Next, in order to create a benchmark for comparison of the separation results of our proposed algorithm we conduct another experiment where we mute speaker 2 for the entire duration and assume the number of sources is known and equal to two, while all the other settings remain the same. Since now the number of sources is assumed to be known, we use a separate gated Kalman filter for each source to track the DOAs (similar to [15]) which is necessary for permutation correction and stitching across blocks in the separation task. Table I shows the separation performances for the controlled benchmark experiment and the main experiment. The performances indicate that even though the main experiment compared to the benchmark one had an uncertainty factor in the number of sources and also had to deal with more sources than sensors for most of the duration of the experiment, the performance only degraded marginally, suggesting robustness and versatility of the algorithm in coping with dynamic scenarios. For our second scenario, the same is repeated for a set-up where the radii of the sources increase with the source number e.g. $r_1 = 1$ m, $r_2 = 1.7$ m, $r_3 = 2$ m while all the other settings remain the same, with results shown in Table II. In this scenario the separation of the more distant sources becomes more challenging as the tracking algorithm assumes all sources have roughly the same detection amplitude compared to clutter. Table II shows the farthest source experiences the largest drop in performance from the benchmark experiment as expected, nevertheless it does not fail in carrying out the separation task.

Next we consider four sources using two microphones 8 cm apart and the total experiment lasting for 16.32 seconds. We start with a reverberation time of $T_{60} = 200$ ms for this scenario and jump to $T_{60} = 300$ ms for the last scenario. The radii of the sources were $r_i = 1$ m, $i = 1, \ldots, 4$ with the DOA trajectories

TABLE I
SEPARATION RESULTS FOR $r_i = 1$ m, $i = 1, \ldots, 3$, USING 2 MICROPHONES, $T_{60} = 200$ ms

2 Sources, Known

	S_1	N/A	S_3	mean
SDR (dB)	5.0	–	5.6	5.3
ISR (dB)	5.9	–	6.2	6.05
SIR (dB)	13.8	–	17.4	15.6
SAR (dB)	17.1	–	18.6	17.85

3 Sources, Unknown/Time-varying

	S_1	S_2	S_3	mean
SDR (dB)	4.6	5.4	4.3	4.77
ISR (dB)	5.8	8.8	5.9	6.83
SIR (dB)	12.2	7.8	10.5	10.17
SAR (dB)	17.1	16.8	17.1	17.0

TABLE II
SEPARATION RESULTS FOR $r_1 = 1$ m, $r_2 = 1.7$ m, $r_3 = 2$ m, USING 2 MICROPHONES, $T_{60} = 200$ ms

2 Sources, Known

	S_1	N/A	S_3	mean
SDR (dB)	5.9	–	3.5	4.7
ISR (dB)	6.8	–	5.3	6.05
SIR (dB)	15.8	–	8.9	12.35
SAR (dB)	17.3	–	14.7	16.0

3 Sources, Unknown/Time-varying

	S_1	S_2	S_3	mean
SDR (dB)	5.5	4.3	2.2	4.0
ISR (dB)	6.7	6.2	4.6	5.83
SIR (dB)	13.2	8.3	5.3	8.93
SAR (dB)	18.3	15.7	15.2	16.4

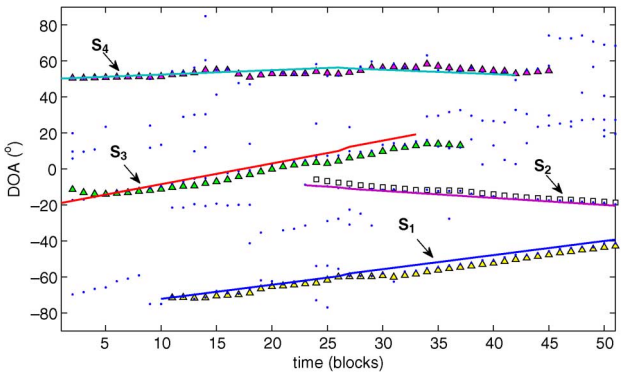

Fig. 12. Separation experiment with 4 unknown/time-varying sources and 2 microphones, $T_{60} = 200$ ms: True DOA (colored lines), SCT peaks (dots) and estimated tracks (colored shapes).

TABLE III
SEPARATION RESULTS FOR $r_i = 1$ m, $i = 1, \ldots, 4$ USING 2 MICROPHONES, $T_{60} = 200$ ms

2 Sources, Known

	S_1	N/A	N/A	S_4	mean
SDR (dB)	3.6	–	–	3.9	3.75
ISR (dB)	4.8	–	–	4.5	4.65
SIR (dB)	11.7	–	–	13.6	12.65
SAR (dB)	14.6	–	–	16.5	15.55

3 Sources, Unknown/Time-varying

	S_1	N/A	S_3	S_4	mean
SDR (dB)	3.4	–	4.0	3.5	3.63
ISR (dB)	4.7	–	7.4	4.5	5.53
SIR (dB)	11.1	–	6.4	10.7	9.4
SAR (dB)	15.1	–	14.7	15.6	15.13

4 Sources, Unknown/Time-varying

	S_1	S_2	S_3	S_4	mean
SDR (dB)	2.8	4.4	3.3	3.4	3.46
ISR (dB)	4.6	7.3	6.6	4.4	5.73
SIR (dB)	8.4	7.3	5.4	10.4	7.86
SAR (dB)	15.4	15.2	14.4	15.9	15.23

depicted in Fig. 12. All the other elements in this scenario were the same as the previous scenarios. Similar to the previous cases we start with the benchmark case of known 2 sources with S_1 and S_4 being only active, then progress upwards to a setting with unknown/time-varying number of sources with S_1, S_3 and S_4 active and finally to a case with unknown/time-varying number of sources with all 4 sources S_i, $i = 1, \ldots, 4$ active. Fig. 12 shows the estimated DOA tracks for the latter case using the proposed method. Table III presents the performance evaluation results for this scenario. In the last scenario the same is repeated but for a more challenging scenario with $T_{60} = 300$ ms. In order to better cope with the increase in reverberation, a larger STFT window size of 2046 samples with 87.5% overlap was utilized. The results of the last scenario are shown in Table IV. Tables III and IV demonstrate the flexibility of the proposed algorithm in separating unknown time-varying number of moving sources for overcomplete situations with twice as many concurrent sources as sensors.

VI. Discussion and Conclusions

In this paper we present a novel framework to solve the problem of tracking and separation of unknown time-varying number of speakers using minimal number of microphones in a reverberant environment. We proposed the integration of a

TABLE IV
SEPARATION RESULTS FOR $r_i = 1$ m, $i = 1, \ldots, 4$ USING 2 MICROPHONES, $T_{60} = 300$ ms

2 Sources, Known

	S_1	N/A	N/A	S_4	mean
SDR (dB)	2.4	–	–	3.7	3.05
ISR (dB)	3.8	–	–	4.9	4.35
SIR (dB)	9.2	–	–	11.5	10.35
SAR (dB)	10.2	–	–	12.1	11.15

3 Sources, Unknown/Time-varying

	S_1	N/A	S_3	S_4	mean
SDR (dB)	1.6	–	2.5	2.6	2.23
ISR (dB)	3.6	–	5.0	4.2	4.27
SIR (dB)	7.0	–	4.8	7.6	6.47
SAR (dB)	9.8	–	10.9	10.8	10.5

4 Sources, Unknown/Time-varying

	S_1	S_2	S_3	S_4	mean
SDR (dB)	1.0	1.1	1.9	2.3	1.58
ISR (dB)	3.5	3.3	4.2	4.0	3.75
SIR (dB)	4.3	2.0	3.4	7.1	4.2
SAR (dB)	10.1	10.2	11.2	11.7	10.8

powerful and versatile ICA-based scanning method for multiple DOA estimation with a well known method in multi-target tracking. Such combination showed promising results in both the tracking and separation tasks using only two microphones for relatively high reverberant environments and in challenging dynamic scenarios involving moving sources and spontaneous births/deaths.

This paper demonstrates how a mixture/superposition model in the framework of BSS can be easily represented as a standard detection model in the framework of multi-target tracking, assuming that the sources have block-frequency sparsities. The solution involves, first, performing frequency-domain ICA on the sensor measurements, then utilizing a permutation-invariant TDOA scanning method such as SCT on the ICA outputs, therefore enabling the mixture model observations to be represented as source location observations in a detection model. The PHD filter, which has proven to be a highly effective method for multi-target tracking when observations are posed in a detection model, is then used for the tracking of the location detections. The post-filtered DOAs are then used to align and stitch the ICA outputs across frequencies and blocks, respectively.

As part of our future work we would like to extend our work performing tracking and separation using multiple sensor pairs hence representing the TDOA measurements in a multidimensional framework using GSCT. One advantage of incorporating multiple dimensional TDOAs is that one can provide more detailed location information and possibly extract the Cartesian location information of the sources. Also the extra dimensions in the measurement model can provide better discrimination in track labeling when sources cross over or get very close. The other advantage would be in the separation task since the extra sensors used can allow for the improvement in the separation performance.

References

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug.. 1976.
[2] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum based technique," in *Proc. ICASSP*, 1994, pp. 273–276.

[3] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with microphone array," in *Proc. Int. Workshop Acoustic Echo and Noise Control*, 2003, pp. 315–318.
[4] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddeley, "Tracking and 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.
[5] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech, Music Process.*, pp. 11:1–11:17, 2010.
[6] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst. J. (Elsevier)*, vol. 55, no. 3, pp. 216–228, 2007.
[7] G. Lathoud and J. Odobez, "Short-term spatio-temporal clustering applied to multiple moving speakers," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 5, pp. 1696–1710, Jul. 2007.
[8] J. A. Wnd, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP*, 2004, pp. 605–608.
[9] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. ISSPA*, 2003.
[10] A. Hyvarinen, J. Karhunen, and E. Oja, *Indepedent Component Analysis*. New York: Wiley Interscience, 2001.
[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[12] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency domain BSS," in *Proc. ISCAS*, 2007, pp. 3247–3250.
[13] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 246–260, Jan. 2012.
[14] F. Nesta, P. Svaizer, and M. Omologo, "Robust two-channel TDOA estimation for multiple speaker localization by using recursive ICA and a state coherence transform," in *Proc. ICASSP*, 2009, pp. 4597–4600.
[15] A. Brutti and F. Nesta, "Multiple source tracking by sequential posterior kernel density estimation through GSCT," in *Proc. EUSIPCO*, 2011, pp. 259–263.
[16] R. Mahler, *Statistical Multisource Multitarget Information Fusion*. Norwood, MA: Artech House, 2007.
[17] R. Mahler, "Multi-target Bayes filtering via first-order multi-target moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.
[18] B.-N. Vo and W. K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.
[19] K. Panta, D. Clark, and B.-N. Vo, "Data association and track management for the Gaussian mixture probability hypothesis density filter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 3, pp. 1003–1016, Jul. 2009.
[20] S. S. B.-N. Vo and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. ICASSP*, 2004, pp. 357–360.
[21] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Proc. LVA/ICA*, 2010, pp. 1–8.
[22] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *Proc. LVA/ICA*, 2010, pp. 41–48.
[23] M. Fallon and S. Godsill, "Multi target acoustic source tracking with an unknown and time varying number of targets," in *Proc. HSCMA*, 2008, pp. 77–80.
[24] M. Fallon and S. Godsill, "Multi target acoustic source tracking using track before detect," in *Proc. WASPAA*, 2007, pp. 102–105.
[25] A. Masnadi-Shirazi and B. Rao, "Separation and tracking of multiple speakers in a reverberant environment using a multiple model particle filter glimpsing method," in *Proc. ICASSP*, 2011, pp. 2516–2519.
[26] B. Balakumar, A. Sinha, T. Kirubarajan, and J. P. Reilly, "PHD filtering for tracking an unknown number of sources using an array of sensors," in *Proc. IEEE Workshop Statist. Signal Process.*, 2005, pp. 43–48.
[27] R. Mahler, "CPHD filters for superpositional sensors," in *Signal and Data Process. Small Targets 2009, SPIE Proc.*, O. E. Drummond, Ed., 2009, vol. 7445.
[28] F. Thouin, S. Nannuru, and M. Coates, "Multi-target tracking for measurement models with additive contributions," in *Proc. Int. Conf. Inf. Fusion*, 2011.
[29] E. Biglieri and M. Lops, "Multiuser detection in a dynamic environment part I: User identification and data detection," *CoRR*, 2007.
[30] D. Angelosante and M. L. E. Biglieri, "Multiuser detection in a dynamic environment: Part II: Joint user identification and parameter estimation," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2365–2374, May 2009.
[31] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multi-target filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 4, pp. 1224–1245, Oct. 2005.
[32] B. Ristik, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Norwood, MA: Artech House, 2004.
[33] D. Clark, B. Ristic, B.-N. Vo, and B.-T. Vo, "Bayesian multi-object filtering with amplitude feature likelihood for unknown object SNR," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 26–37, Jan. 2010.
[34] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in *Proc. MLSP*, 2008.
[35] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. ICA*, 2001, pp. 803–806.
[36] E. Lehmann and A. 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.
[37] J. Hoffman and R. Mahler, "Multi-target miss distance via optimal assignment," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, no. 3, pp. 327–336, May 2004.
[38] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 624–639, Mar. 2011.
[39] 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.

Alireza Masnadi-Shirazi (S'09) received the B.S. degree (summa cum laude) in electrical engineering from the University of Texas at Arlington in 2005 and the Ph.D. degree in electrical and computer engineering from University of California, San Diego, in 2012.

His main research interests are in the areas of estimation theory, blind source separation, multi-target tracking and speech signal processing.

Bhaskar D. Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1981 and 1983, respectively. Since 1983, he has been with the University of California at San Diego, La Jolla, where he is currently a Professor with the Electrical and Computer Engineering Department. His interests are in the areas of digital signal processing, estimation theory, and optimization theory, with applications to digital communications, speech signal processing, and human-computer interactions.

He is the holder of the Ericsson endowed chair in Wireless Access Networks and was the Director of the Center for Wireless Communications (2008–2011). His research group has received several paper awards. Recently, a paper he co-authored with B. Song and R. Cruz received the 2008 Stephen O. Rice Prize Paper Award in the Field of Communications Systems and a paper he co-authored with S. Shivappa and M. Trivedi received the best paper award at AVSS 2008. He was elected to the fellow grade in 2000 for his contributions in high resolution spectral estimation. He has been a Member of the Statistical Signal and Array Processing technical committee, the Signal Processing Theory and Methods technical committee, and the Communications technical committee of the IEEE Signal Processing Society. He has also served on the editorial board of the EURASIP Signal Processing Journal.