AN ICA-BASED RFS APPROACH FOR DOA TRACKING OF UNKNOWN TIME-VARYING NUMBER OF SOURCES

Alireza Masnadi-Shirazi and Bhaskar D. Rao

Dept. of Electrical and Computer Engineering, University of California, San Diego
{amasnadi, brao}@ucsd.edu

ABSTRACT

Methods based on frequency-domain independent component analysis (ICA) in junction with state coherence transform (SCT) have been shown to be robust for extracting source location information like direction of Arrival (DOA) in highly reverberant environments and in the presence of spatial aliasing. Also, by exploiting the frequency sparsity of the sources, such methods have proven to be effective when the number of simultaneous sources is larger than the number of microphones. In many real world problems the number of concurrent speakers is unknown and varies with time as new speakers can appear and existing speakers can disappear or undergo silence periods. In order to deal with this challenging scenario of unknown time-varying number of speakers, we propose the use of the probability hypothesis density (PHD) filter which 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. The tracking capabilities of the proposed method is demonstrated using simulations of multiple sources in reverberant environments.

Index Terms— Blind source separation, independent component analysis, source localization, multi-target tracking, probability hypothesis density

1. 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 desired speaker from background interference which can be used in 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 multiple bearing information or the direction of arrival (DOA) by means of the time difference of arrival (TDOA).

Multiple TDOA estimation using frequency domain independent component analysis (ICA) was first proposed in [1]. 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 [2]. ICA estimates a de-mixing matrix for the separation task. For many real world problems, the signals undergo a convoluted mixing due to reverberation. By transforming the mixture to the frequency domain by use of 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. In [1], 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 frequency bins have been corrected and there are no frequency bins effected by spatial aliasing. Recently an extension to [1] has been proposed under the name of state coherence transform (SCT) that does not require permutation correction and is insensitive to spatial aliasing [3]. SCT is a scanning method which forms a pseudo-likelihood between the propagation model of the TDOA scan points and the propagation pertaining to the columns of the estimated mixing matrices obtained from ICA, therefore exposing their phase coherence. 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.

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 [4]. 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 sensors can receive a set of spurious detections (clutter) due to the multipath propagation caused by reverberation and spatial aliasing. Recently, random finite sets (RFS) have allowed the problem of multi-target tracking with uncertainty in target number

to be posed in an optimal Bayesian filtering framework, one that is the extension of the well known single target Bayesian framework [5]. Using RFSs, the multi-target states and the number of targets are integrated to form a set-valued variable. However, the optimal RFS Bayes filter is computationally intractable as it becomes a combinatorial problem on the number of targets. 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 [5]. 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 for when the targets follow a linear/Gaussian dynamic model [6, 7].

The problem of extracting location information of unknown time-varying number of speakers using RFSs and PHD filtering have been proposed before. These methods, however, use the generalized cross-correlation phase transform (GCC-PHAT) in the front-end to obtain the measurements, hence bearing the inherent limitations of GCC-PHAT for multiple sources. One limitation is that it is not able to obtain reliable detections as the number of concurrent speakers increases [8]. Other methods exit that use ICA/SCT in the front-end but use a naive thresholding approach to estimate the number of targets [9]. In this paper we propose the use of the best of both methods, that is the use of ICA/SCT in the front-end and the use of the GM-PHD in the back-end.

2. FREQUENCY DOMAIN BSS AND SCT

We assume there are L microphones in the array and M sources. After taking the short time Fourier transform (STFT) of the convolutedly mixed (due to reverberation) signals, the observations would end up having a linear mixture representation in each frequency bin k and frame n:

$$Y(k,n) = H(k)S(k,n) \tag{1}$$

where sensors $Y(k,n) = [Y_1(k,n)...Y_L(k,n)]^T$, sources $S(k,n) = [S_1(k,n)...S_M(k,n)]^T$ and $H(k)$ is the mixing matrix corresponding to the k^{th} frequency bin. From hereafter, we will omit the index n for brevity. For the case of $L = M$, complex-valued ICA 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

$$\hat{S}(k) = \hat{W}(k)Y(k) \tag{2}$$

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

$$H(k) = \begin{pmatrix} |h_{11}(k)|e^{-j2\pi f_k T_{11}} & |h_{12}(k)|e^{-j2\pi f_k T_{12}} \\ |h_{21}(k)|e^{-j2\pi f_k T_{21}} & |h_{22}(k)|e^{-j2\pi f_k T_{22}} \end{pmatrix} \tag{3}$$

where T_{qp} is the propagation time from the p^{th} source and 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

$$\hat{W}^{-1}(k) = \begin{pmatrix} \eta_1|\hat{h}_{11}(k)|e^{-j2\pi f_k \hat{T}_{11}} & \eta_2|\hat{h}_{12}(k)|e^{-j2\pi f_k \hat{T}_{12}} \\ \eta_1|\hat{h}_{21}(k)|e^{-j2\pi f_k \hat{T}_{21}} & \eta_2|\hat{h}_{22}(k)|e^{-j2\pi f_k \hat{T}_{22}} \end{pmatrix} \tag{4}$$

where η_i represents the diagonal entries of the arbitrary scaling matrix. By making a farfield assumption for the sources and neglecting reverberation, the TDOA information emerges by taking the ratios of the entries of each column in (4)

$$r_1(k) = \frac{|\hat{h}_{11}(k)|}{|\hat{h}_{21}(k)|}e^{-j2\pi f_k \hat{\Delta}t_1}, \quad r_2(k) = \frac{|\hat{h}_{12}(k)|}{|\hat{h}_{22}(k)|}e^{-j2\pi f_k \hat{\Delta}t_2} \tag{5}$$

where $\hat{\Delta}t_i$ are the TDOAs of the sources with respect to the microphones. As it can be seen from (5), 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 (5) and is invariant to scaling and magnitude, the ratios can be simplified as

$$\bar{r}_1(k) = \frac{r_1(k)}{|r_1(k)|}, \quad \bar{r}_2(k) = \frac{r_2(k)}{|r_2(k)|} \tag{6}$$

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 (6) by exploiting the linear relationship between the TDOAs and the true frequencies along the different bins [1]. 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 (5) and a propagation model that can intrinsically account for both permutations and spatial aliasing [3]. The propagation model that results in TDOA of a source with respect to the microphones, denoted as τ, is assumed to be

$$c(k,\tau) = e^{-j2\pi f_k \tau} \tag{7}$$

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

$$SCT(\tau) = \sum_k \sum_{m=1}^{2} \left[1 - g\left(\frac{\|c(k,\tau) - \bar{r}_m(k)\|}{2}\right)\right] \tag{8}$$

where the transform is scanned for different values of τ and $g(.)$ is a function of the Euclidian distance. A good option for $g(.)$ is shown to have a sigmoidal shape such as $g(x) = tanh(\alpha x)$, where α is a real positive constant that defines the TDOA sensitivity and is usually set empirically. It can be easily understood from (8) that one could expect to see higher mappings of SCT for values of τ which $r_m(k)$ 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 (6) 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 in τ) caused by spatial aliasing it greatly reduces its sensitivity towards spatial aliasing. Moreover, one 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 of 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, by exploiting the frequency sparsity of the sources (which is typical of speech) in each data block, and assuming that at each frequency-block segment at most two sources are active, a complete TDOA mapping with peaks pertaining to the possible sources becomes possible. From the far-field assumption, one can convert TDOA detections into DOA using

$$\theta = cos^{-1}(c\Delta t/\Delta q) \tag{9}$$

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

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. This is why for our problem of tracking the DOA of unknown time-varying number of sources, a suitable filtering technique is needed to reject clutter caused by multi-path propagations.

3. BAYESIAN MULTI-TARGET TRACKING AND PHD FILTERING

Let us consider the multi-target scenario of having $M(t-1)$ targets at time $t-1$ with states $x_{t-1,1}, ..., x_{t-1,M(t-1)}$ 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 $x_t, ..., x_{t,M(t)} \in \mathcal{X}$. On the other hand, let's assume that at time t, the sensor makes $N(t)$ observations (detections) $z_{t,1}, ..., z_{t,N(t)}$ 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 any subset of targets not get detected (missed detections). Assuming that the ordering and association of the measurements and the state estimates has no significance, the multi-target states and observations can be represented as finite sets such as

$$X_t = \{x_t, ..., x_{t,M(t)}\} \in \mathcal{F}(\mathcal{X}) \tag{10}$$

$$Z_t = \{z_t, ..., z_{t,N(t)}\} \in \mathcal{F}(\mathcal{Z}) \tag{11}$$

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 muti-target detection RFS state $Z(t)$ is the union of target generated detections and clutter, the goal of Mahler's RFS mutli-target filtering [5] is to estimate the number of targets and their states while rejecting clutter and accounting for missed detections. With the RFS formulation, the multi-target Bayesian filter can be computed sequentially via the prediction and update steps as following

$$f_{t|t-1}(X_t|Z_{1:t-1}) = \int f_{t|t-1}(X_t|X')f_{t-1|t-1}(X'|Z_{1:t-1})\delta X' \tag{12}$$

$$f_{t|t}(X_t|Z_{1:t}) = \frac{f_{t|t}(Z_t|X_t)f_{t|t-1}(X_t|Z_{1:t-1})}{\int f_{t|t}(Z_t|X'_t)f_{t|t-1}(X'_t|Z_{1:t-1})\delta X'_t} \tag{13}$$

where $Z_{1:t}$ is the series 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. Due to the use of combinatorial set-integrals in the optimal Bayesian recursions of (12-13), 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 multi-target Bayesian recursions of (12-13) which instead of propagating the full posterior density, it propagates the first moment of multi-target posterior density, known as the posterior intensity [5] .

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 equations (12-13). It is worthy to note that due to the first order moment mapping of the PHD filter, $D_{t|t}(x_t|Z_{1:t})$ 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 [5], the PHD intensities can be recursively estimated as follows

$$D_{t|t-1}(x_t|Z_{1:t-1}) = \int F_{t|t-1}(x'|x_{t-1}) D_{t-1|t-1}(x'|Z_{1:t-1}) dx' + b_t(x_t) \quad (14)$$

$$D_{t|t}(x_t|Z_{1:t}) = [1 - p_D(x_t)]\, D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{z_t \in Z_t} \frac{\psi_{z_t}(x_t) D_{t|t-1}(x_t|Z_{1:t-1})}{\kappa_t(z_t) + \int \psi_{z_t}(\zeta) D_{t|t-1}(\zeta|Z_{1:t-1}) d\zeta} \quad (15)$$

In the prediction equation (14)

$$F_{t|t-1}(x_t|x_{t-1}) = p_S(x_{t-1}) f_{k|k}(x_t|x_{t-1}) + \beta_{t|t-1}(x_t|x_{t-1}) \quad (16)$$

where $f_{k|k}(x_t|x_{t-1})$ 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 (14), b_t is the intensity of spontaneous new births at time t. In the update equation (15),

$$\psi_{z_t}(x_t) = p_D(x_t) g(z_t|x_t) \quad (17)$$

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

$$\kappa_t(z_t) = \lambda c_t(z_t) \quad (18)$$

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 fact it turns out that the integral of the PHD intensity gives the expected number of targets [5]

$$\hat{M}_{t|t} = \int D_{t|t}(x_t|Z_{1:t}) dx_t \quad (19)$$

At the end, the multi-target state estimates are extracted by finding the $\hat{M}_{t|t}$ peaks of $D_{t|t}(x_t|Z_{1:t})$.

Even though the PHD filter is much less computationally expensive compared to the multi-target recursions of (12-13), the integrals present in the PHD recursions of (14-15) 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 the PHD filter. 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 [6]. 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 our measurements and target state dynamic follow a linear/Gaussian model, GM-PHD is used for the multi-source filtering.

4. SYSTEM INTEGRATION

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 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 follow a linear/Gaussian model such as

$$f_{t|t-1}(\theta_t|\theta_{t-1}) = \mathcal{N}(\theta_t; \theta_{t-1}, Q_{t-1}) \quad (20)$$

$$g(z_t|\theta_t) = \mathcal{N}(z_t; \theta_t, R_t) \quad (21)$$

Fig.1 illustrates the system model incorporating ICA/SCT with PHD filtering. From Fig.1 it can be seen that ICA is performed in 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. In this case we use the complex-valued maximum-likelihood Infomax ICA [2]. The initialization of the ICA stage performed on each block is done from scratch and not based on the previous block convergences. This is to encourage diversity in the ICA estimates so 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.

5. SIMULATIONS

The proposed method was conducted on simulated data obtained from Lehmann's image method [10]. Signals were sampled at $f_s = 16kHz$ and the STFT frequency-frame segments were obtained using a Hanning window of 2048 taps and 75% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being about 0.4 seconds in length. The experiment went on for a total duration of about 18 seconds. The room dimensions were $8m \times 5m \times 3.5m$ with a reverberation time of $T_{60} = 600ms$. Only $L = 2$ microphones were used which were placed $36cm$ 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. The speakers all moved along a half-circular path about $1.5m$ from the microphone pair. Fig.2 shows the DOA detections and the true source DOAs along with their estimated tracks . The tracks labels were calculated using an extension to the GM-PHD filter [7]. The performance of the proposed method is compared with two alternative approaches discussed earlier. One method uses the GCC-PHAT at the front-end to acquire

detection measurements and uses the same PHD filter at the back-end for the filtering, similar to [8]. The other method uses ICA/SCT at front end and a naive thresholding method for selection of the peaks, similar to [9]. The Wasserstein distance which is a multi-target error metric for time-varying number of targets is used to evaluate the performances [5]. The proposed method had a mean Wasserstein error of 12.34 while the "GCC-PHAT+PHD filtering" method had a mean error of 23.99 and the "ICA/SCT+naive thresholding" had a mean error of 20.85. It is worthy to note that the proposed method shares the back-end (GM-PHD filtering) with the first method and the front-end (ICA/SCT) with the second method, therefore a better performance of the proposed compared to the other two methods implies it is incorporating the best of both approaches.

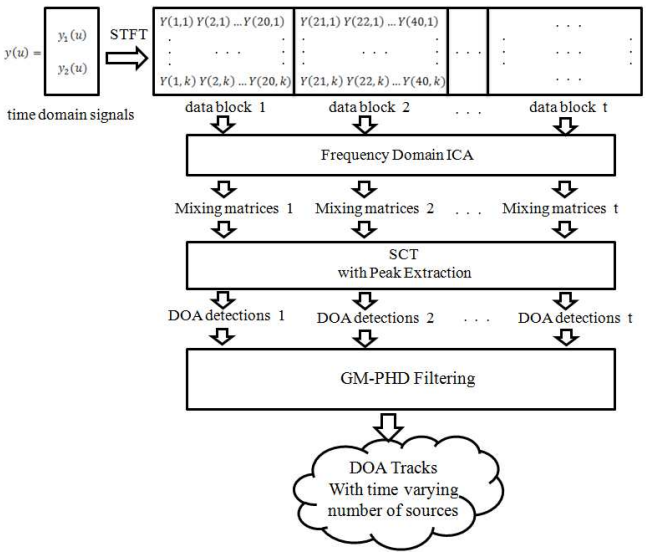

Fig. 1. Block diagram of proposed method: STFT, ICA and SCT segments form the front-end and the PHD filtering segment form the back-end.

6. CONCLUSIONS

In this paper we present a novel framework to solve the problem of tracking the DOAs of unknown time-varying number of speakers using minimal number of microphones in a reverberant environment. We proposed the integration of a powerful and versatile scanning method for multiple DOA estimation with a well known method in radar/sonar multi-target tracking. Such a combination showed promising results in the DOA estimation task of up to 6 concurrent speakers in relatively high reverberant environment using only 2 microphones. Future investigation will address the separation problem of unknown time-varying number of sources using a similar framework.

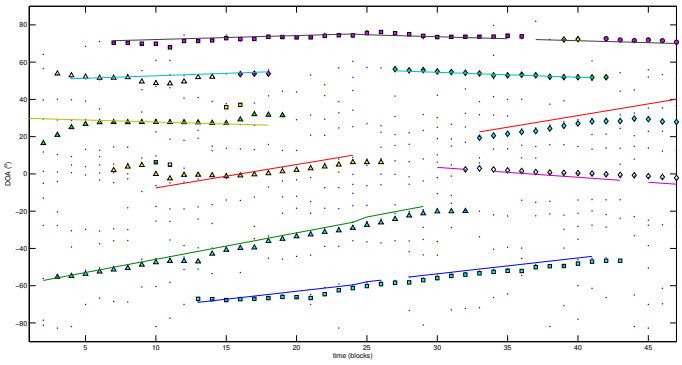

Fig. 2. True (colored lines), SCT peaks or detections (dots) and estiamted tracks (colored shapes)

7. REFERENCES

[1] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. of ISSPA*, 2003.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Indepedent Component Analysis*, New York, Wiley Interscience, 2001.

[3] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional tdoa estimation of multiple sources," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, 2012.

[4] A. Brutti and F. Nesta, "Multiple source tracking by sequential posterior kernel density estimation through gsct," in *Proc. of EUSIPCO*, 2011, pp. 259–263.

[5] R. Mahler, *Statistical multisource multitarget information fusion*, Norwood, MA, Artech House, 2007.

[6] B.-N. Vo and W. K. Ma, "The gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, 2006.

[7] K. Panta, D. Clark, and B.-N. Vo, "Data association and track management for the gaussian mixture probability hypothesis density filter," *IEEE Trans. on Aerospace and elec. systems*, vol. 45, no. 3, 2009.

[8] 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, 2006.

[9] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Proc. of LVA/ICA*, 2010, pp. 1–8.

[10] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image source model," *J. of the Acoustical Soc. of America*, vol. 124, no. 1, 2008.