CARTESIAN TRACKING OF UNKNOWN TIME-VARYING NUMBER OF SPEAKERS USING DISTRIBUTED MICROPHONE PAIRS

Alireza Masnadi-Shirazi and Bhaskar D. Rao

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

ABSTRACT

This paper considers the challenging problem of Cartesian tracking of multiple sources using multiple distributed microphone arrays when the number of sources is unknown and varies with time due to new sources appearing and existing sources disappearing or undergoing long silence periods. The problem is posed in a bearings-only tracking framework. Frequency-domain independent component analysis (ICA) in conjunction with state coherence transform (SCT) is used as a robust method to extract the bearing information of the speakers. Also, by exploiting the frequency sparsity of the sources, ICA/SCT has proven to be effective even when the number of simultaneous speakers is larger than the number of microphones in an array. Next, the bearing information for each array is fused using a sequential-corrector probability hypothesis density (PHD) filter with a limited field of view (FOV) for each microphone array. The limited FOV is essential for applications like speech in order to account for the more distant sources not registering detections with respect to a sensor array. The promising tracking capability of the proposed method is demonstrated using simulations of multiple speakers in a reverberant environment.

***Index Terms*—** Independent component analysis, source localization and tracking, multi-target multi-source tracking, probability hypothesis density filter

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 a desired speaker from background interference with applications in high-quality hearing aids, speech enhancement and noise reduction for smart phones, among others. 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 Cartesian location information by means of triangulation of the directions of arrival (DOA) calculated from multiple distributed sensor arrays.

Multiple DOA 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 DOAs 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 affected by spatial aliasing. 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]. 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 in an array.

In many real world problems, not only can the sources experience spatial dynamics, they can also experience temporal dynamics where the number of concurrent sources is unknown and varies with time due to new speakers appearing and existing speakers disappearing or undergoing long silence periods. Moreover, the sensors can receive a set of spurious detections (clutter) due to the multi-path propagation caused by reverberation and spatial aliasing. In our previous work, we have proposed an algorithm that can track the DOAs of the sources in such scenarios using a single sensor array by synergistically combining two key ideas, one in the front-end and the other at the back-end. In the front-end, it employs ICA to demix the mixtures and the SCT to represent the sig-

This work was supported in part by a gift from Qualcomm Inc.

nals in a DOA detection framework. 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) [4] is utilized that solves the data association problem intrinsically, hence providing distinct DOA tracks [5, 6]. There, it is shown that the ICA-SCT-PHD filtering approach outperforms other RFS-based methods which use the generalized cross-correlation phase transform (GCC-PHAT) in the front-end to obtain the measurements.

In this paper we extend our previous work to the multiple distributed sensor array case in order to triangulate and track the Cartesian location of the sources by posing it as a bearings-only tracking problem. One typical and practical method that is used to fuse the data obtained from multiple sensor arrays is sequential-sensor updating in the PHD filter update stage [7, 8, 9]. In such problems with mostly application in radar signal processing, it is usually assumed that each sensor array observes the DOAs from all the sources, i.e. the fields of view (FOV) of the sensors is not affected much by the range to the targets/sources, and all that needs to be done is to fuse the DOA observations across the sensor arrays. However, for speech sources in a room environment, such an assumption is not satisfied as a set of sources might not register as DOA observations by a set of sensors due to them being distant, resulting in the derailment of the sequential sensor PHD update. In this paper we propose a state/sensor dependent probability of detection scheme that limits the FOV based for each sensor array based on range, giving more importance to observations coming from closer sources and less importance to observations coming from more distant sources. By doing so we avoid the derailment of sequential PHD update. Computer simulations are carried out that verifies the effectiveness of the proposed method.

2. FRONT-END: ICA/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 assuming that an estimate of the mixing matrices can be obtained through ICA [2] and by neglecting multipath propagations, the time difference of arrival (TDOA) information emerges by taking the ratios of the entries of each column of the mixing matrices and normalizing them with the magnitudes, hence obtaining

$$\bar{r}_1(k) = e^{-j2\pi f_k \hat{\Delta} t_1}, \quad \bar{r}_2(k) = e^{-j2\pi f_k \hat{\Delta} t_2} \tag{4}$$

where $\hat{\Delta} t_i$ are the TDOAs of the sources with respect to the microphones. Such ratios are invariant to the scaling ambiguities of the estimation process [3]. 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 (4) 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 (4) 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

$$c(k,\tau) = e^{-j2\pi f_k \tau} \tag{5}$$

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

$$SCT(\tau) = \sum_k \sum_{m=1}^{2} \left[1 - g\left(\frac{\|c(k,\tau) - \bar{r}_m(k)\|}{2} \right) \right] \tag{6}$$

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 (6) 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 (4) 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

$$\theta = cos^{-1}(c\Delta t/\Delta q) \tag{7}$$

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

3. BACK-END: PHD FILTERING

Let us consider the multi-target scenario of having M 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 a subset of targets are not 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

$$X_t = \{x_t,...,x_{t,M(t)}\} \in \mathcal{F}(\mathcal{X}) \tag{8}$$

$$Z_t = \{z_t,...,z_{t,N(t)}\} \in \mathcal{F}(\mathcal{Z}) \tag{9}$$

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 [10] 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

$$f_{t|t-1}(X_t|Z_{1:t-1}) = \int f_{t|t-1}(X_t|X')f_{t-1|t-1}(X'|Z_{1:t-1})\delta X' \tag{10}$$

$$f_{t|t}(X_t|Z_{1:t}) = \frac{f_{t|t}(Z_t|X_t)f_{t|t-1}(X_t|Z_{1:t-1})}{\int f_{t|t}(Z_t|X'_t)f_{t|t-1}(X'_t|Z_{1:t-1})\delta X'_t} \tag{11}$$

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 (10-11), 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 (10-11) which instead of propagating the full posterior density, it propagates the first moment of multi-target posterior density, known as the posterior intensity [10].

4. MULTI-SENSOR ARRAY GM-PHD FILTER UPDATE WITH LIMITED FOVS

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 to recursively estimate the Cartesian location of sources by posing it as a multi-target bearings-only tracking problem where the respective state dynamics and sensor model for a single source are

$$x_t = x_{t-1} + \gamma_{t-1} \tag{12}$$

$$z_t^{(q)} = \text{arctan}\left(\frac{\xi_t - \xi^{(q)}}{\zeta_t - \zeta^{(q)}}\right) + w_t,\ q = 1,...,Q \tag{13}$$

where $x_t = [\xi_t\ \zeta_t]^T$ is the state vector of Cartesian coordinates, $z_t^{(q)}$ is the DOA observation/detection obtained from the q^{th} sensor array, γ_{t-1} and w_t are independent Gaussian noises and $x^{(q)} = [\xi^{(q)}\ \zeta^{(q)}]^T$ is the location of the q^{th} sensor array. Since the relationship between the observations and states in (13) is non-linear the unscented-Kalman (UK) approximation of the GM-PHD filter needs to be implemented [4]. The sequential-sensor updating in the update stage of the GM-PHD filter works as follows[1].

Assume that the PHD at time $t-1$ is $v_{t-1}(x)$. The PHD is predicted using the state dynamic model to obtain $v_{t|t-1}(x)$. Then the PHD is updated/corrected by the observation set $Z_t^{(1)}$ of sensor array 1 to obtain the update

$$v_t^{(1)}(x) = (1 - p_{D,t}) v_{t|t-1}(x) + \sum_{z \in Z_t^{(1)}} v_{D,t}(x; z) \qquad (14)$$

where $p_{D,t}$ is the probability of detection usually assumed to be independent of the state, and the complete formulation of $v_{D,t}(x; z)$ can be found in [4]. At sensor array 2, $v_{t-1}^{(1)}(x)$ is used as the predicted PHD for sensor array 2 and the PHD is updated similar to (14) as follows

$$v_t^{(2)}(x) = (1 - p_{D,t}) v_t^{(1)}(x) + \sum_{z \in Z_t^{(2)}} v_{D,t}(x; z) \qquad (15)$$

This cycle is repeated for all the Q sensor arrays. From the above formulations, $p_{D,t}$ being state independent implies that each sensor array observes the DOAs from all the sources, i.e. sensors have identical field of views (FOV). However, for speech sources in a room, it could be the case where a sensor does not pick up DOA observations from a set of sources due to them being distant. Hence, updating the sensors using (14-15) would result in the derailment of the sequential-sensor PHD updates. To fix this, we propose the use of a state and sensor location dependent probability of detection $p_{D,t}(x, x^{(q)})$ where a higher probability of detection is assigned to sources that are closer to sensor q and a lower one is assigned to sources that are farther away. This means that the states are less sensitive to PHD sensor updates for the more distant sources as less reliable observations about them are available. We model this behavior by a Gaussian centered at sensor location $x^{(q)}$ as follows

$$p_{D,t}(x, x^{(q)}) = w_{D,t} N(x; x^{(q)}, C_{D,t}) \qquad (16)$$

where $w_{D,t}$ and $C_{D,t}$ are given model parameters such that $p_{D,t}(x, x^{(q)})$ lie between 0 and 1 for all x. With such model for probability of detection the PHD update equation of (14), and similarly, sensor update equation of (15) are modified to

$$v_t^{(1)}(x) = v_{t|t-1}(x) - v_{D,t}(x) + \sum_{z \in Z_t^{(1)}} v_{D,t}(x; z) \qquad (17)$$

$$v_t^{(2)}(x) = v_t^{(1)}(x) - v_{D,t}^{(1)}(x) + \sum_{z \in Z_t^{(2)}} v_{D,t}(x; z) \qquad (18)$$

where complete formulations for $v_{D,t}(x)$ and $v_{D,t}(x; z)$ can be found in Equation (51) of [4]. The UK formulations can also be found in Table V of [4].

5. SIMULATIONS

The proposed method was conducted on simulated data obtained from Lehmann's image method [11]. Signals were sampled at $f_s = 16kHz$ and the STFT frequency-frame segments were obtained using a Hanning window of 2048 taps and 87.5% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being about 0.64 seconds in length. The experiment went on for a total duration of about 17 seconds. The room dimensions were $6m \times 4m \times 2.5m$ with a reverberation time of $T_{60} = 300ms$. Only $L = 2$ microphones being $36cm$ apart were used for each sensor array with a total of 6 sensor arrays distributed across the room. The speakers could appear and disappear at any time and moved in different directions. There were a total of 7 different speakers in which at one point all 7 were active simultaneously. Fig.1 shows the true and estimated source trajectories in the room along with their activity onset and offset marks in seconds. The experiment is repeated for when we assume state independent probability of detections (without any limited FOV) but results are not shown due to the very poor performance. This means that the limited FOV approach has a significant impact on the success of the multi-sensor PHD filter updater in applications like speech where a distant speaker with respect to a sensor array does not register as SCT detections (or if it does the amplitude of the detection is weak and in the level of the amplitude of the clutter). However, for other applications like radar signal processing where most targets are picked up by most sensors a limited FOV approach similar to what has been proposed here, is unnecessary. Finally, we note that because of the limited FOV approach, a source that is farther from most of the sensors, e.g. the source that is in the top left corner of Fig.1, might have the least accurate localization as the effective number of sensors used to carry out the triangulation task is least for that source.

6. CONCLUSIONS

In this paper we present an extension to our previous work that allows for Cartesian tracking of sources for unknown time-varying number of speakers in a reverberant environment using distributed microphone arrays. For each microphone array we utilize a powerful and versatile ICA-based scanning method for multiple DOA estimation and then fuse the DOA detections from multiple sensor arrays using a well known method in radar/sonar multi-sensor multi-target tracking. We limit the FOV for each array allowing for robust triangulation of the source positions. The proposed method

[1] note that full equations of the GM-PHD filter and the UK approximation is not presented in this paper due to space limitations. We refer the reader to [4, 9] for further reading.

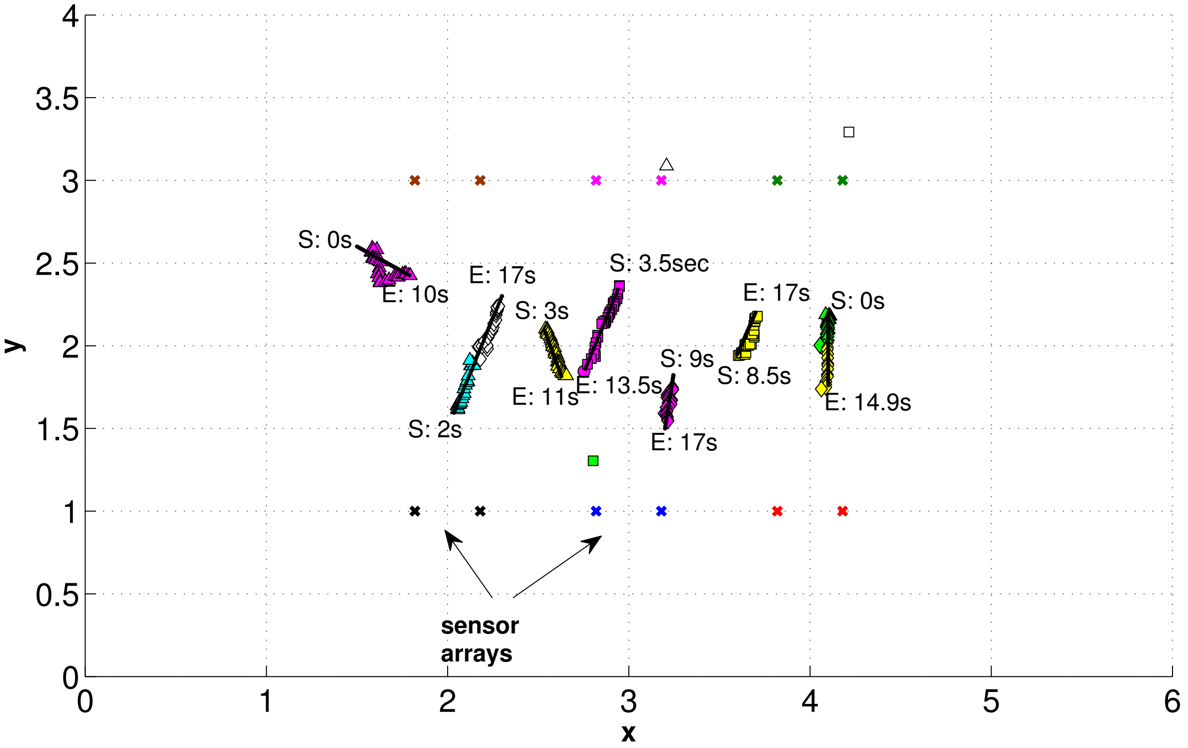

Fig. 1. True trajectories (black line), estimated trajectories (colored shapes) along with start (S) and end (E) mark of each source in seconds.

showed promising results in the tracking task of up to 7 concurrent speakers in reverberant environment using only 2 microphones for each array and a total of 6 arrays.

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

[5] A. Masnadi-Shirazi and B.D. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *to appear in IEEE Trans. on Audio, Speech, and Language Processing*.

[6] A. Masnadi-Shirazi and B. Rao, "An ICA-based RFS approach for DOA tracking of unknown time-varying number of sources," in *Proc. of EUSIPCO*, 2012.

[7] R. Mahler, "Multi-target Bayes filtering via first-order multi-target moments," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, no. 4, 2003.

[8] R. Mahler, "The multisensor PHD filter, I: General solution via multitarget calculus," in *Proc. of SPIE*, 2009.

[9] N. T. Pham, W. Huang, and S.H. Ong, "Multiple sensor multiple object tracking with GMPHD filter," in *Proc. of Int. Conf. on Information Fusion*, 2007.

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

[11] 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.