Bearing-only Acoustic Tracking of Moving Speakers for Robot Audition

Christine Evers, Alastair H. Moore and Patrick A. Naylor
Department of Electrical & Electronic Engineering
Imperial College London
London, SW7 2AZ, United Kingdom
Email: c.evers@imperial.ac.uk

Jonathan Sheaffer and Boaz Rafaely
Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva 84105, Israel
Email: sheaffer@ee.bgu.ac.il

***Abstract*—This paper focuses on speaker tracking in robot audition for human-robot interaction. Using only acoustic signals, speaker tracking in enclosed spaces is subject to missing detections and spurious clutter measurements due to speech inactivity, reverberation and interference. Furthermore, many acoustic localization approaches estimate speaker direction, hence providing bearing-only measurements without range information. This paper presents a probability hypothesis density (PHD) tracker that augments the bearing-only speaker directions of arrival with a cloud of range hypotheses at speaker initiation and propagates the random variates through time. Furthermore, due to their formulation PHD filters explicitly model, and hence provide robustness against, clutter and missing detections. The approach is verified using experimental results.**

***Index Terms*—Acoustic signal processing; Speaker tracking; Acoustic tracking; Bearing-only tracking; Clutter; Missing detections.**

I. Introduction

The increasing availability of humanoid robots, such as NAO by Aldebaran Robotics, is driving demand for robots capable of interacting with humans in intuitive ways. However, Human-Robot Interaction (HRI) is a complex and largely unsolved problem, particularly when faced with multiple people who may simultaneously require the robots attention. Whilst vision-based HRI has received widespread attention in the literature, *robot audition* [1] is mainly used to facilitate a natural communication channel between the human user and robot, with a focus on speech recognition and enhancement. Nonetheless, beyond the user request dialogue, speech signals contain information that should be exploited constructively to infer additional information about the environment. Furthermore, audio signals often facilitate detectability of events that may be occluded for other sensors. For example, humans outside of the Field of View (FoV) of cameras or occluded by other objects produce speech signals that can be used for acoustic scene analysis. Thus, robust audio processing systems are required for meaningful audio-visual fusion.

In order to adjust its position and sensor to interact with humans, the robot needs to localize and track humans within the surrounding environment. Localization can be used to estimate instantaneous source directions from the speech signals [2], [3], [4], [5], [6]. However, localization does not account for the evolution of directions over time. Furthermore, localization estimates are subject to estimation errors due to reverberation and noise. Speaker tracking is used to smooth the noisy measurements and estimate speaker trajectories. Traditional tracking systems [7] often utilise variants of the Kalman filter in order to estimate the unknown speaker positions [8]. A particle filter extension was presented in [9]. A track-before-detect system was proposed in [10], estimating the tracks directly form the microphone measurements without the need for localization. Multi-speaker tracking using Probability Hypothesis Density (PHD) filters and Time-Difference-of-Arrival measurements was presented in [11].

PHD filters [12], [13] track multiple objects without the need for data association by explicitly accounting for missing measurements, clutter, track initiation and termination. PHD filters are therefore particularly attractive for speaker tracking in the presence of clutter and missing measurements and are hence utilized in this paper.

This paper focuses on three challenges affecting speaker tracking, namely 1) bearing-only measurements, 2) missing detections, and 3) clutter. Speech radiated in enclosed environments is subject to reverberant reflections from surrounding walls and objects. Whilst localization often reduces the measurements to dominant path signals [14], [6], strong reflections can lead to spurious detections. Robustness against clutter measurements is therefore required for tracking. Furthermore, speakers cannot be localized during periods of speech inactivity. Tracking hence needs to ensure track propagation through silent periods to avoid false track termination.

Furthermore, localization algorithms for acoustic data often estimate Directions-of-Arrival (DoAs), such that the measurements are bearing-only. However, in order to map from the angular directions to Cartesian positions, range estimates are required to solve a system of determined equations. Range estimates can be periodically obtained either by means of fusion with other sensors or by kinematic ranging where robot maneuvers are used to triangulate a speaker position [15]. Nonetheless, confirming range on every estimated DoA of a possible sound source is computationally unfeasible, especially in the presence of clutter. In order to initiate and propagate speaker tracks with missing range measurements, this paper proposes to initiate speakers with an appended range estimate which is extrapolated in time whilst the angular component is updated by the measurements. A similar approach was proposed in [16], [17], where the angular components of newborn states are sampled over the entire surveillance volume independent of the measurements.

A more efficient speaker initiation scheme is a measurement-driven birth process as proposed in [18]. This paper proposes a modified measurement-driven birth process for bearing-only measurements. Rather than sampling uniformly over the state space, the angular birth components are sampled from the DoA measurements. Range estimates can be augmented in the birth states by drawing random variates along the birth components' directions.

This paper proposes a new approach to bearing-only acoustic tracking using a Gaussian Mixture PHD (GM-PHD) filter. Section II summarizes the signal model. PHD filters for tracking in the presence of clutter and missing detections are discussed in Section III. Section IV proposes the measurement-driven birth process for bearing-only measurements. Experiments are presented in Section V and conclusions drawn in Section VI.

II. Signal model

Signal models for the single speaker are now presented and extended to the multiple speaker case.

A. Single speaker

Consider a speaker n at time t with state $\mathbf{x}_{t,n} \triangleq \begin{bmatrix}\mathbf{p}_{t,n}^T & \mathbf{v}_{t,n}^T\end{bmatrix}^T$, where the speaker's Cartesian coordinates in the room are given by $\mathbf{p}_{t,n} \triangleq \begin{bmatrix}x_{t,n} & y_{t,n} & z_{t,n}\end{bmatrix}^T$ and its velocity is denoted as $\mathbf{v}_{t,n} \triangleq \begin{bmatrix}\dot{x}_{t,n} & \dot{y}_{t,n} & \dot{z}_{t,n}\end{bmatrix}^T$. The state can be modelled as

$$\mathbf{x}_{t,n} = \boldsymbol{F}_t\, \mathbf{x}_{t-1,n} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}_{q\times 1},\, \boldsymbol{Q}_t\right) \tag{1}$$

where $\boldsymbol{F}_t$ models the speaker dynamics and the process noise term, $\mathbf{v}_t$, captures subtle deviations of the speaker position from the motion model with covariance, $\boldsymbol{Q}_t$ (see [19] for a review of model choices).

Assuming M_t DoA measurements are obtained from a localization algorithm at time t, the m^{th} DoA measurement, $\mathbf{z}_{t,m} \triangleq \begin{bmatrix}\theta_{t,m} & \phi_{t,m}\end{bmatrix}^T$, $\forall m = 1, \ldots, M_t$, can be modelled as

$$\mathbf{z}_{t,m} = g(\mathbf{x}_{t,n}) + \mathbf{w}_{t,m}, \qquad \mathbf{w}_{t,m} \sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{R}\right), \tag{2}$$

where $\mathbf{w}_t$ is the measurement noise with covariance $\boldsymbol{R}$, and $g(\mathbf{x}) = \begin{bmatrix}\theta & \phi\end{bmatrix}^T$ is the Cartesian to spherical tranformation, with inclination $\theta = \arccos\left(z/\sqrt{x^2+y^2+z^2}\right)$ and azimuth $\phi = \arctan\left(y/x\right)$.

B. Multiple speakers

The set of states of N_t speakers in the acoustic scene is given by $\boldsymbol{X}_t \triangleq \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N_t}\}$. In multi-speaker tracking, both the number of speakers as well as their positions are unknown. The multi-speaker states are hence defined as a point process [20] with realizations

$$\boldsymbol{\xi}_t \triangleq (N_t, \boldsymbol{X}_t)\,. \tag{3}$$

Similar to (3), the DoA measurement state, $\boldsymbol{\upsilon}_t$, at the tracker input can be expressed as a point process with realizations

$$\boldsymbol{\upsilon}_t = (M_t, \boldsymbol{Z}_t), \tag{4}$$

where M_t, is the number of measurements and $\boldsymbol{Z}_t \triangleq \{\mathbf{z}_{t,1}, \ldots, \mathbf{z}_{t,M_t}\}$ is the set of measurements.

III. Tracking framework

In a Bayesian framework the multi-speaker states, $\boldsymbol{\xi}_t$, can be fully described by the joint posterior Probability Density Function (pdf), $p(\boldsymbol{\xi}_t|\boldsymbol{Z}_{1:t})$, of the number of speakers and their positions. It is convenient to assume that N_t is Poisson distributed and the speaker locations are Independent and Identically Distributed (i.i.d.). Under these assumptions, (3) is a Poisson Point Processs (PPPs) [21], and the joint posterior pdf of $\boldsymbol{\xi}_t$ is fully described by the density of $\boldsymbol{X}_t$, also known as the *intensity function* [21]. Therefore, it is sufficient to derive an estimator of the intensity function in order to estimate the pdf. The PHD filter [12], summarized in this section, recursively estimates the first-order moment of the intensity.

A. Speaker prediction

The predicted intensity, $\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1})$, of a PPP is given by [12]

$$\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) = \int_{\mathcal{X}^+} f_t(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})\, \lambda(\boldsymbol{X}_{t-1}|\boldsymbol{Z}_{1:t-1}) d\boldsymbol{X}_{t-1}. \tag{5}$$

where the region of support is defined as $\mathcal{X}^+ = \mathcal{X} \cup \phi$. Here, $\mathcal{X} = \mathbb{R}^6$ defines the region of support of the 6-dimensional states of real numbers, and ϕ denotes the empty space for which $\boldsymbol{\xi}_t = (0, \emptyset)$, where $\emptyset$ is the empty set. The definition of $\mathcal{X}^+$ is crucial for the developments in this Section, as the inclusion of ϕ facilitates explicit modelling of 1) the probability of speaker termination, $f(\emptyset|\mathbf{x}_{t-1,n})$, 2) the likelihood of speaker initialization, $f(\mathbf{x}_{t,n}|\emptyset)$, and 3) clutter measurements with likelihood $p(\boldsymbol{Z}_t|\emptyset)$.

Extending the integral over the region of support, $\mathcal{X}^+$, defined in Section II-B, into the space over real numbers, $\mathcal{X}$, and ϕ, and defining $p_s = (1-p_b)$ as the probability of survival and p_b as the probability of speaker initiation, the predicted intensity is given by

$$\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) = \lambda^b(\boldsymbol{X}_t|\emptyset) + \lambda^s(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}), \tag{6}$$

where

$$\lambda^b(\boldsymbol{X}_t|\emptyset) \triangleq p_b\, f_t(\boldsymbol{X}_t|\emptyset)\, \lambda(\boldsymbol{X}_t|\emptyset)$$

$$\lambda^s(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \triangleq p_s \int_{\mathcal{X}} f_t(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})\, \lambda(\boldsymbol{X}_{t-1}|\boldsymbol{Z}_{1:t-1}) d\boldsymbol{X}_{t-1}.$$

where $\lambda^s(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1})$ is the predicted intensity of surviving speakers. The birth process, $\lambda^b(\boldsymbol{X}_t|\emptyset)$, proposed in this paper facilitates bearing-only tracking and is discussed in Section IV.

Vo *et al.* demonstrated in [13] that the PHD for Gaussian state spaces such as (1) and (2), the PHD filter takes a closed form solution as a Gaussian Mixture Model (GMM). The predicted intensity of surviving speakers is hence given

$$\lambda^s(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) = \sum_{j=1}^{J_{t-1}} w_{t|t-1}^{(j)} \mathcal{N}(\mathbf{x}|\mathbf{m}_{t|t-1}^{(j)}, \boldsymbol{\Sigma}_{t|t-1}^{(j)}), \tag{7}$$

where J_{t-1} is the udpated number of Gaussian Mixture (GM) components at time $t-1$, the predicted weights are $w_{t|t-1}^{(j)} = p_s\, w_{t-1}^{(j)}$, and where the mean and covariance terms are given by the Extended Kalman Filter (EKF) prediction:

$$\mathbf{m}_{t|t-1}^{(j)} = \boldsymbol{F}_t^{(j)}\, \mathbf{m}_{t-1|t-1}^{(j)}, \tag{8a}$$

$$\boldsymbol{\Sigma}_{t|t-1}^{(j)} = \boldsymbol{F}_t^{(j)}\, \boldsymbol{\Sigma}_{t-1|t-1}^{(j)} [\boldsymbol{F}_t^{(j)}]^T + \boldsymbol{Q}, \tag{8b}$$

where $\boldsymbol{F}_t$ and $\boldsymbol{Q}$ were defined in Section II-A.

B. Speaker detection and missing detections

To account for missing detections of speakers, a Bernoulli thinning process [21] is applied to the predictions that accepts $\boldsymbol{X}_t$ with probability of detection, p_d, and rejects the state with probability, $(1-p_d)$, such that

$$\lambda^s(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) = \lambda^d(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) + \lambda^u(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \tag{9}$$

with intensities of detection, $\lambda^d(\cdot)$, and missed detection, $\lambda^u(\cdot)$:

$$\lambda^d(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \triangleq p_d\, \lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \tag{10}$$

$$\lambda^u(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \triangleq (1-p_d)\, \lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \tag{11}$$

The predicted intensity, $\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1})$, split into four branches as in (6) and (9), is now updated using knowledge inferred from the DoA measurements. Note that undetected speakers are not subject to the information update as $\lambda^u(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t}) = \lambda^u(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1})$.

C. Speaker update and clutter

Information can be inferred from the measurements by application of Bayes's theorem to the predicted intesnity. However, the resulting process is not a PPP, such that the intensity recursion is not analytically tractable [21]. Nonetheless, the process can be approximated by its first-order moment, resulting in the PHD filter with prediction in (5) and update [12], [18]:

$$\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t}) = \lambda^u(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) + \sum_{m=1}^{M_t} \frac{p(\mathbf{z}_{t,m}|\boldsymbol{X}_t)\, \lambda^{(b\cup d)}(\boldsymbol{X}_t|\mathbf{z}_{1:t-1,m})}{\lambda^{(b\cup d)}(\mathbf{z}_{t,m})} \tag{12}$$

where $\lambda^{(b\cup d)}(\cdot) \triangleq \lambda^{b}(\cdot) + \lambda^{d}(\cdot)$ assuming a measurement-driven speaker birth process [18], and with partition function, $\lambda(\mathbf{z}_{t,m})$:

$$\lambda(\mathbf{z}_{t,m}) = \lambda^{c} + \int_{\mathcal{X}} p(\mathbf{z}_{t,m}|\boldsymbol{X}_t)\lambda^{(b\cup d)}(\boldsymbol{X}_t|\mathbf{z}_{1:t-1,m})d\boldsymbol{X}_t \quad (13)$$

where $\lambda^{c} \triangleq p(\mathbf{z}_{t,m}|\emptyset)\,\lambda(\emptyset)$ is the clutter intensity. The GM-PHD update is given by [13], [18]:

$$\begin{aligned}\lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t}) &= (1-p_d)\lambda^{s}(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t-1}) \quad (14)\\ &+ \sum_{m=1}^{M_t}\left[\sum_{j=1}^{J_{b,t}} w_{t,b,m}^{(j)}\lambda^{b}(\boldsymbol{X}_t|\emptyset) + \sum_{j=1}^{J_{t-1}} w_{t,m}^{(j)}\,\mathcal{N}(\mathbf{x}|\mathbf{m}_{t,m}^{(j)}, \boldsymbol{\Sigma}_{t,m}^{(j)})\right]\end{aligned}$$

where $J_{b,t}$ is the number of birth components and the weights are

$$w_{t,b,m}^{(j)} = \frac{w_{t|t-1,b}^{(j)}}{\mathcal{L}(\mathbf{z}_{t,m})} \quad \text{and} \quad w_{t,m}^{(j)} = p_d\,\frac{p(\mathbf{z}_{t,m}|\mathbf{x}_t^{(j)})w_{t|t-1}^{(j)}}{\mathcal{L}(\mathbf{z}_{t,m})}$$

with partition function, $\mathcal{L}(\mathbf{z})$, defined as

$$\mathcal{L}(\mathbf{z}) = \kappa_t(\mathbf{z}) + \sum_{j=1}^{J_{b,t}} w_{t|t-1,b}^{(j)} + p_d \sum_{j=1}^{J_{t-1}} p(\mathbf{z}|\mathbf{x}_t^{(j)})w_{t|t-1}^{(j)}. \quad (15)$$

The GM mean and covariance are expressed as

$$\mathbf{m}_{t|t,m}^{(j)} = \mathbf{m}_{t|t-1}^{(j)} + \boldsymbol{K}_{t,m}^{(j)}\left(\mathbf{z}_{t,m} - g(\mathbf{m}_{t|t-1}^{(j)})\right) \quad (16)$$

$$\boldsymbol{\Sigma}_{t|t,m}^{(j)} = \left(\boldsymbol{I} - \boldsymbol{K}_{t,m}^{(j)}\,\boldsymbol{G}_{t,m}^{(j)}\right)\boldsymbol{\Sigma}_{t|t-1}^{(j)}. \quad (17)$$

The gain, $\boldsymbol{K}_t$, and innovation covariance, $\boldsymbol{S}_t$, are

$$\boldsymbol{K}_{t,m}^{(j)} = \boldsymbol{\Sigma}_{t|t-1}^{(j)}\,[\boldsymbol{G}_{t,m}^{(j)}]^{T}\,[\boldsymbol{S}_{t,m}^{(j)}]^{-1} \quad (18)$$

$$\boldsymbol{S}_{t,m}^{(j)} = \boldsymbol{G}_{t,m}^{(j)}\,\boldsymbol{\Sigma}_{t|t-1}^{(j)}[\boldsymbol{G}_{t,m}^{(j)}]^{T} + \boldsymbol{R}. \quad (19)$$

where $\boldsymbol{G}_t$ is the Jacobian matrix of the Cartesian to spherical transformation, and $\boldsymbol{R}$ is the measurement noise covariance in (2). After the update, the GMM consists of $J_t = J_{t-1}(M_t+1) + J_{b,t}\,M_t$ components. In order to reduce the exponential growth of components, pruning as proposed in [13] is applied for GM reduction.

D. Estimated number of speakers

The partition function ensures that the GM components are weighted correctly such that the expected value of the number of speakers, $\mathbb{E}[N_t]$, is given by [21]

$$\mathbb{E}[N_t] = \int_{\mathcal{X}^{+}} \lambda(\boldsymbol{X}_t|\boldsymbol{Z}_{1:t})d\boldsymbol{X}_t \quad (20)$$

For the GM-PHD filter, the number of speakers can be estimated similarly to the states as per Sections III-A to III-C, such that

$$N_t = \left\lfloor (1-p_d)N_{t|t-1} + \sum_{m=1}^{M_t}\left[\sum_{j=1}^{J_t} w_{t,m}^{(j)} + \sum_{i=1}^{J_{t,b}} w_{t,b,m}^{(i)}\right]\right\rfloor, \quad (21)$$

where $N_{t|t-1} = p_s\,N_{t-1}$.

E. Point estimate extraction

GM reduction techniques are applied in order to extract extract N_t point estimates from the GM components in (14). Statistically unlikely components are first truncated. Clustering as in [22] is applied to the remaining components with a suitably aggressive merging threshold [23]. If the number of merged GMs exceeds the estimated number of speakers, the N_t components with highest weight are chosen to extract the point estimates, $\hat{\mathbf{x}}_t$, of the speaker state. Note that the state extraction in this Section does not affect the GMs in (14) but merely reduces a copy of the GMM.

IV. Proposed birth process for bearing-only tracking

The measurements at the tracker input are assumed in this paper to be DoAs and are hence bearing-only. However, for a commensurate mapping from spherical to Cartesian coordinates, range is required.

Beard *et al.* [16], [17] modified the GM-PHD filter for the bearings-only problem by introducing a diffuse birth model accounting for the unmeasured range component. The angular component of target birth states is drawn from a uniform distribution over the surveillance region, whilst the range component is drawn from a prior.

In practical speech applications, new speakers can start dialogues at any point within the room. Therefore, sufficiently many birth components need to be sampled in order to cover the state space $\mathcal{X}$ at a sufficient resolution to capture initiation of new speakers. This paper proposes to adapt the more efficient measurement-driven birth process in [18] to the bearing-only problem. Rather than sampling uniformly over the state space, new speaker processes are drawn from areas within the room where the likelihood corresponds to high values.

A. Measurement-driven bearing-only birth process

In [18], the positions of birth components are sampled from the observations, $\{\mathbf{z}_{t,m}\}_{m=1}^{M_t}$. In this paper, in order to introduce the missing range measurements, the observations are augmented with a range prior, $\hat{r}$, with variance $\sigma_{\hat{r}}^2$, such that $\tilde{\mathbf{z}}_{t,m} \triangleq \begin{bmatrix}\hat{r} & \mathbf{z}_{t,m}^{T}\end{bmatrix}^{T}$ and $\tilde{\boldsymbol{R}} \triangleq \text{diag}\begin{bmatrix}\sigma_r^2 & \boldsymbol{R}\end{bmatrix}$. Furthermore, the speaker velocities are sampled from a velocity prior, $\mathbf{v}$, with covariance $\boldsymbol{\Sigma}_{\mathbf{v}}$. Hence for each observation, N_r positions are sampled form $\tilde{\mathbf{z}}_{t,m}$, such that

$$\mathbf{x}_{b,m}^{(i)} \sim \mathcal{N}\left(h(\tilde{\mathbf{z}}_{t,m}),\, \boldsymbol{H}_{t,m}\,\tilde{\boldsymbol{R}}\,\boldsymbol{H}_{t,m}^{T}\right), \quad (22)$$

for each $m = 1, \ldots, M_t$ and $i = 1, \ldots, N_r$, where h is the spherical to Cartesian transformation with Jacobian $\boldsymbol{H}_{t,m}$. Furthermore, N_v velocity vectors, $\dot{\mathbf{x}}_b^{(j)}$, are sampled from

$$\dot{\mathbf{x}}_b^{(j)} \sim \mathcal{N}\left(\mathbf{v}^{(j)},\, \boldsymbol{\Sigma}_{\mathbf{v}}\right). \quad (23)$$

for all $j = 1, \ldots, N_v$. The resulting $J_{b,t} = M_t\,N_r\,N_v$ birth states are constructed from the sampled positions and velocities as the vector $\mathbf{m}_{t,b,m}^{(i,j)} \triangleq \left[\begin{bmatrix}\mathbf{x}_{b,m}^{(i)}\end{bmatrix}^{T} \quad \begin{bmatrix}\dot{\mathbf{x}}_b^{(j)}\end{bmatrix}^{T}\right]^{T}$. The predicted birth weights are givenas

$$w_{t|t-1,b}^{(j)} = \frac{N_b}{M_t \cdot N_v \cdot N_r}, \quad (24)$$

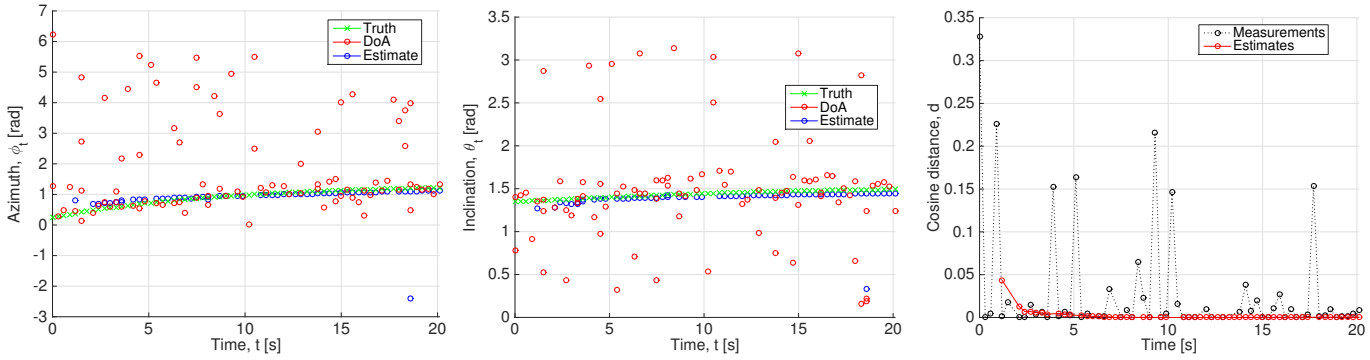
where N_b is the prior expected number of speaker initiations.

B. Range propagation with time

The update in (14) infers knowledge from the measurements through the innovation, $\left(\mathbf{z}_{t,m} - g(\mathbf{m}_{t|t-1}^{(j)})\right)$. Recalling that g transforms the three-dimensional Cartesian position to the two-dimensional DoA (see Section II-A), the innovation updates only the angular components of $\mathbf{m}_{t|t,m}^{(j)}$. Therefore, the range of surviving speakers sampled in (22) is extrapolated in time via (7), but is not corrected using measurements. Naturally, the range component is therefore expected to be subject to divergence.

V. Experimental results

Two experiments are presented to investigate the tracking performance of the proposed approach. In both experiments a speaker trajectory is initialized at absolute Cartesian position $(7, 3, 1.5)$ m in a $15 \times 15 \times 3$ m room with a stationary sensor at $(3, 2, 0.58)$ m corresponding to the height of a NAO robot standing on the ground. The speaker trajectory follows a straight line with $v_y = 0.5$ m/s towards the North facing wall of the room. To model the effects



(a) Azimuth, ϕ_t (b) Inclination, θ_t (c) Cosine distance, $d(\mathbf{z}_{t,m}, g(\hat{\mathbf{x}}_t))$

Fig. 1: Experiment 2 – Tracking performance in azimuth and inclination.

of reverberation on localization, clutter measurements as a Poisson process with clutter rate λ, uniformly distributed over the surveillance region. A spherical microphone array is assumed in order to faciliate 360×180 deg surveillance of the surrounding in azimuth and inclination. In both experiments, the GM-PHD filter is initialized with no components. The probability of survival is set to 0.98. The birth process samples 50 birth components with prior birth weight, $N_b = 10^{-3}$. To reduce the exponential growth of the filter, the pruning algorithm in [13] is used for GM reduction with a truncation threshold of $T = 10^{-9}$, a merging threshold of $M = 0.1$, and a maximum allowance number of 200 components.

1) Experiment 1: In the first experiment, the effects of the range prior are investigated. For this purpose, the clutter rate is set to $\lambda = 0$ with $p_d = 1$. The sensor noise is 0.5 rad in azimuth and 10^{-6} rad in inclination. For each realization of the experiment, the range prior mean is increased between $\hat{r} = 1.0, \ldots, 10$ m in steps of 1.5 m. The range estimates over time of the 20 realizations are plotted in Fig. 2.

As expected, the range cannot diverge to the ground truth as the range estimates are propagated without information update. The results in Fig. 2 indicate jumps of the range estimates. This behaviour is due to the fact that a small angle corresponds to small positional displacement at near-range but large displacement at far-range. Hence, the innovation term, $(\mathbf{z}_{t,m} - g(\mathbf{m}^{(j)}_{t|t-1})$, implicitly accounts for the range estimate, even though $\mathbf{z}_{t,m}$ and $g(\mathbf{m}^{(j)}_{t|t-1})$ contain angles only. Consequently, the weight decreases for GM components whose range gradually diverges. At the same time, the weight increases for components resembling the ground truth more closely. This eventually leads to a "jump" as the point estimates correspond to the highest weighted GM components as discussed in Section III-E.

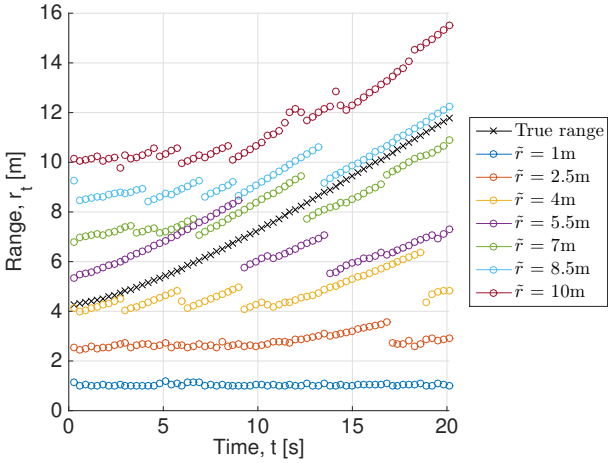


Fig. 2: Experiment 1: Range estiamtes for $\hat{r} = 0.5, 1, \ldots, 10$ m.

2) Experiment 2: The experiment is repeated for $\hat{r} = 6$ m with a standard deviation of $\sigma_{\hat{r}} = 10^{-3}$ m with a clutter rate of $\lambda_c = 0.5$ (i.e., between 0 and 3 clutter measurements per sample), detection probability of $p_d = 0.8$ and measurement noise of 3 deg in azimuth and 1 deg in inclination. The velocity is initialized to the ground truth with standard deviation $\sigma_v = 10^{-3}$ m/s in each direction. The cosine distance, $d(g(\mathbf{x}_t), g(\hat{\mathbf{x}}_t))$, between the track estimates and ground truth is evaluated as:

$$d(g(\mathbf{x}_t), g(\hat{\mathbf{x}}_t)) = 1 - \left\{ \frac{g(\mathbf{x}_t) \cdot g(\hat{\mathbf{x}}_t)}{\|g(\mathbf{x}_t)\| \|g(\hat{\mathbf{x}}_t)\|} \right\}, \tag{25}$$

where $\|\|$ denotes the vector norm and $\cdot$ is the dot product. Figures 1a and 1b compare the ground truth, estimates and measurements over time in azimuth and inclination. The cosine distance of the estimates to the ground truth is compared to the cosine distance of the measurements in Fig. 1c. These results illustrate that tracking performance improves significantly on the measurements with a maximum cosine distance of 0.043 for the estimates compared to 0.328 for the measurements.

VI. Conclusion

This paper proposed a novel approach to bearing-only acoustic tracking of speakers in the presence of clutter and missing detections. Experiments evaluated the performance for tracking of a single moving speaker. The results demonstrate that track estimates are robust against the spurious clutter measurements and probability of detection of 80%. The distance between tracks and the ground truth is significantly improved compared to the bearing-only measurements.

Acknowledgment

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

The research was partially supported by the Helmsley Charitable Trust through the Agricultural, Biological and Cognitive Robotics Center of Ben-Gurion University of the Negev.

REFERENCES

[1] H. G. Okuno, K. Nakadai, and H.-D. Kim, *Robotics Research*. Springer, 2011, ch. Robot Audition: Missing Feature Theory Approach and Active Audition, pp. 227–244.

[2] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[3] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[4] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, 1989.

[5] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, Jan. 2010, ch. 11.

[6] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Nice, France, Jul. 2014.

[7] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House, 1998.

[8] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *IEEE Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Munich, Germany, Apr. 1997, pp. 371 – 374.

[9] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Adv. Signal Process.*, Jun. 2006.

[10] M. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1228–1242, Aug. 2010.

[11] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[12] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[13] 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.

[14] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[15] J. A. Fawcett, "Effect of course maneuvers on bearings-only range estimation," *IEEE T. Acoust. Speech.*, vol. 36, no. 8, pp. 1193–1199, Aug. 1988.

[16] M. Beard and S. Arulampalam, "Performance of PHD and CPHD filtering versus JIPDA for bearings-only multi-target tracking," in *Proc. Conf. Information Fusion (FUSION)*, 2012, pp. 542–549.

[17] M. Beard, B. T. Vo, B.-N. Vo, and S. Arulampalam, "A partially uniform target birth model for Gaussian Mixture PHD/CPHD filtering," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 4, pp. 2835–2844, Oct. 2013.

[18] B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 2, pp. 1656–1668, Apr. 2012.

[19] X.-R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. part I: Dynamic models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[20] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes*. Springer, 2003, vol. I: Elementary Theory and Methods.

[21] R. L. Streit, *Poisson point processes*. Springer, 2010.

[22] D. J. Salmond, "Mixture reduction algorithms for point and extended object tracking in clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 2, pp. 667–686, Apr. 2009.

[23] ——, "Tracking in uncertain environments," Technical Memorandum AW 121, University of Sussex and Royal Aerospace Establishment, 1989.