DoA Reliability for Distributed Acoustic Tracking

Christine Evers, *Senior Member, IEEE,* Emanuël A. P. Habets, *Senior Member, IEEE,*
Sharon Gannot, *Senior Member, IEEE,* and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Distributed acoustic tracking estimates the trajectories of source positions using an acoustic sensor network. As it is often difficult to estimate the source-sensor range from individual nodes, the source positions have to be inferred from Direction of Arrival (DoA) estimates. Due to reverberation and noise, the sound field becomes increasingly diffuse with increasing source-sensor distance, leading to decreased DoA estimation accuracy. To distinguish between accurate and uncertain DoA estimates, this paper proposes to incorporate the Coherent-to-Diffuse Ratio as a measure of DoA reliability for single-source tracking. It is shown that the source positions therefore can be probabilistically triangulated by exploiting the spatial diversity of all nodes.**

***Index Terms*—Acoustic sensors, Smart homes, Bayes methods.**

I. Introduction

Autonomous systems and smart devices rely on accurate knowledge of the positions of surrounding objects for human-machine interaction [1]. Acoustic scene mapping [2], [3], [4] provides 3D representations of the sound sources and microphone arrays in the surrounding environment. However, in realistic acoustic conditions, it is often difficult to accurately infer the positions of sound sources distant to a single microphone array [3]. In large enclosures, e.g., in smart homes, networks of spatially distributed acoustic sensors can be exploited constructively for robust scene mapping [5], [6], [7]. Examples of network nodes include mobile phones [8], digital personal assistants [9], or robots [10].

For each node in the network, sound source localization algorithms [11] estimate the instantaneous positional information of sources for a sequence of time frames. Devices exploited for ad-hoc networks are typically equipped with compact microphone arrays, such that the range between the node's sensors and sources in the acoustic far-field is difficult to determine. Sound source localization therefore may only resort to estimates of the source Direction of Arrivals (DoAs) [12], [13], [14]. Furthermore, reverberation and noise often lead to missing and false DoA estimates as well as estimation errors [15], [16]. Distributed acoustic tracking [17], [18] can be used to obtain smoothed source trajectories from the instantaneous DoA estimates by incorporating spatio-temporal models of the source motion. Within the Bayesian framework [19], the uncertainty in the source motion model is traded off against the DoA reliability. However, most DoA estimators only provide point estimates of the source directions, but cannot quantify the reliability of the estimates [16], [20]. Therefore, source tracking algorithms typically express the DoA reliability as a constant covariance matrix that is obtained by prior empirical experimentation [21], [22], supervised learning [23], or using the confidence intervals of DoA histograms [24], [25].

However, in practice, DoA estimation accuracy is closely coupled to specific characteristics of each scenario, including the time-varying speech activity and source-sensor geometry. Reverberation [26] and ambient noise lead to diffuse noise fields [27], i.e., the acoustic energy approaches equal probability in every direction [28]. Furthermore, the energy in the direction of a source decreases with increasing source-sensor range. Therefore, the acoustic sound field at a microphone is decreasingly directional with increasing source-sensor range. The coupling between the DoA reliability and the sound field diffuseness is crucial for distributed sensor networks, where knowledge inferred from reliable nodes nearby a source must be distinguished from information at distant nodes.

The sound field diffuseness can be quantified by the Coherent-to-Diffuse Ratio (CDR) [29], [30]. The CDR was previously used to evaluate the speech presence probability for blind speech separation in [31] and source extraction in [32]. In [33], the DoA estimates from a single microphone array were modelled by a von Mises (vM) distribution [34] whose concentration parameter is a function of the CDR.

This paper proposes a novel approach to distributed acoustic tracking that incorporates the CDR as a DoA reliability. Building on the model in [33] and using directional statistics [34], [35] we derive a Bayesian filter that distinguishes reliable DoA estimates at nearby nodes from DoA estimates at distant nodes. In contrast to [31], performing triangulation using instantaneous DoA estimates, the proposed tracking algorithm probabilistically triangulates the Cartesian source positions using the DoA estimates from all nodes within a centralized communication scheme. Node-specific information about the source directions is inferred from each node's DoA estimates. Assuming a synchronized network, estimates of the source position are obtained by fusing statistics of the node-specific information from the spatially diverse nodes.

Section II formulates the problem, Section III derives the proposed methodology, Section IV presents the experimental evaluation using realistic room simulations of a human talker. Conclusions are drawn in Section V.

II. Problem Formulation

A. Likelihood of the DoA Estimates

Denote the 2D Cartesian position of a sound source relative to node $n = 1, \ldots, N$ at position $\mathbf{q}_n$ and time frame t as

C. Evers and P. A. Naylor are with the Dept. Electrical and Electronic Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK e-mail: c.evers@imperial.ac.uk and p.naylor@imperial.ac.uk.

E. A. P. Habets is with Intl. Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), 91058 Erlangen, Germany (e-mail: emanuel.habets@audiolabs-erlangen.de).

S. Gannot is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel (e-mail: Sharon.Gannot@biu.ac.il).

The research leading to these results has received funding from the UK EPSRC Fellowship grant no. EP/P001017/1.

$\mathbf{x}_{t,n} \triangleq \left[x_{t,n}, y_{t,n}\right]^T$. The source position relative to node n can be obtained from $\mathbf{x}_{t,\ell}$ relative to any node, $\ell \neq n$, as

$$\mathbf{x}_{t,n} = f_n(\mathbf{x}_{t,\ell}) \triangleq \boldsymbol{R}_{n|\ell}\,\mathbf{x}_{t,\ell} - \mathbf{q}_n + \mathbf{q}_\ell, \tag{1}$$

where $\boldsymbol{R}_{n|\ell}$ is the rotation matrix. In the following we assume, without loss of generality, that the axes are orientated equally for all nodes, such that $\boldsymbol{R}_{n|\ell}$ is a 2×2 identity matrix.

The direction and range between the source and node, n, can be obtained from the source position using the Cartesian-to-spherical transformation, such that $\mathbf{s}_{t,n} \triangleq \left[\phi_{t,n}, r_{t,n}\right]^T$ where $\phi_{t,n} = \arctan {}^{y_{t,n}}\!/_{x_{t,n}}$ is the source azimuth and $r_{t,n} = \sqrt{x_{t,n}^2 + y_{t,n}^2}$ is the source-to-node range.

For distant sound sources, estimation of the range is difficult when using compact sensor arrays. Hence, at each array, estimates, $\omega_{t,n} \in [0, 2\pi)$, of the true source DoA, $\phi_{t,n} \in [0, 2\pi)$, are available, whereas the range, $r_{t,n}$, is unmeasured. The likelihood function of each DoA estimate, $p\left(\omega_{t,n} \mid \mathbf{s}_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}\right)$, can be modelled by a vM distribution [33], [34] with mean, $\phi_{t,n}$, and concentration parameter, $\kappa_{t,n}^{(\mathrm{DoA})} > 0$:

$$\begin{aligned} p\left(\omega_{t,n} \mid \mathbf{s}_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}\right) &= \mathcal{M}\left(\omega_{t,n} \,|\, \phi_{t,n},\, \kappa_{t,n}^{(\mathrm{DoA})}\right) \\ &\triangleq \frac{(2\pi)^{-1}}{I_0(\kappa_{t,n}^{(\mathrm{DoA})})} \exp\{\kappa_{t,n}^{(\mathrm{DoA})}\, \cos\left(\omega_{t,n} - \phi_{t,n}\right)\}, \end{aligned} \tag{2}$$

where $I_p(\kappa)$ is the modified Bessel function of the first kind and order p.

B. Reliability Measure for DoA Estimates

Previous work showed that the accuracy of narrowband DoA estimates is highly dependent on the CDR. The CDR is also known to be related to the source-sensor range and the reverberation time of the enclosure. In contrast to the range and reverberation time, the CDR can be estimated efficiently and accurately with a small microphone array using, e.g., [36].

In our model, the concentration of the vM distribution can be used to indicate the reliability of the DoA estimates. Therefore, as in [33], we use the CDR, $\Gamma_{t,n}$, to determine the concentration parameter, $\kappa_{t,n}^{(\mathrm{DoA})}$, using

$$\kappa_{t,n}^{(\mathrm{DoA})} = \ell_{\min} + (\ell_{\max} + \ell_{\min}) \frac{10^{c\varrho/10}}{10^{c\varrho/10} + [\Gamma_{t,n}]^{\varrho}}. \tag{3}$$

The terms $\ell_{\min}, \ell_{\max}, c, \varrho$ respectively are the minimum and maximum concentration values, offset, and steepness of the transition region of the mapping from $\Gamma_{t,n}$ to $\kappa_{t,n}^{(\mathrm{DoA})}$.

The reliability of the DoA estimates, that is now reflected by the concentration of the vM distribution, can be used to distinguish between reliable DoA estimates from nearby nodes and less reliable DoA estimates from distant nodes.

C. Early Fusion of $\omega_{t,n}$ and $\kappa_{t,n}^{(DoA)}$ across Frequencies

Performing estimation in the Short Time Fourier Transform (STFT) domain results in one DoA and one CDR estimate per frequency bin, $k \in 1, \ldots, K$, time frame, and node. We assume that source tracking is provided with estimates per time frame and per node. The frequency-dependent estimates are therefore spectrally fused by evaluating their median values across all frequency bins, resulting in $\omega_{t,n}$ and $\kappa_{t,n}^{(\mathrm{DoA})}$.

D. Source Position Inference from DoA Estimates

Bayesian acoustic source tracking estimates the posterior Probability Density Function (pdf), $p\left(\mathbf{x}_{t,n_0} \mid \boldsymbol{\Omega}_{1:t}\right)$, of the current source position, $\mathbf{x}_{t,n_0}$, relative to a reference point, $\mathbf{q}_{n_0}$, with index, n_0, from $\boldsymbol{\Omega}_{1:t} \triangleq \{\boldsymbol{\omega}_{1:t}, \boldsymbol{\kappa}_{1:t}^{(\mathrm{DoA})}\}$, where $\boldsymbol{\omega}_{1:t} \triangleq \left[\omega_{1,1}, \ldots, \omega_{1,N}, \ldots, \omega_{t,N}\right]^T$, and $\boldsymbol{\kappa}_{1:t}^{(\mathrm{DoA})}$ is structured similar to $\boldsymbol{\omega}_{1:t}$. A point estimate of the source position at t, is given by the Maximum *a posteriori* (MAP) estimate:

$$\hat{\mathbf{x}}_{t,n_0}^{\mathrm{MAP}} = \underset{\mathbf{x}_{t,n_0}}{\arg\max}\, p\left(\mathbf{x}_{t,n_0} \mid \boldsymbol{\Omega}_{1:t}\right). \tag{4}$$

However, i) $\kappa_{t,n}^{(\mathrm{DoA})}$ is unknown in practice; ii) each node, n, only has access to its own DoA estimates, $\omega_{1,n}, \ldots, \omega_{t,n}$, such that (4) cannot be evaluated directly; and iii) the 2D source position, $\mathbf{x}_{t,n_0}$, must be inferred from the 1D DoA estimates.

III. Proposed Methodology

This section derives a novel approach for distributed source tracking that exploits (3) to distinguish reliable DoA estimates at nearby nodes from less reliable estimates at distant nodes.

A. Network Fusion for Distributed Acoustic Tracking

In practice, each node only has access to its own DoA estimates. Hence, the source position is tracked by separately propagating a node-specific pdf in time using that node's DoA estimates. The statistics of the node-specific pdfs are fused within the network. For approaches that do not account for the reliability of DoA estimates, network fusion can lead to track divergence as knowledge inferred from nearby nodes cannot be distinguished from uncertain information from distant nodes.

In contrast to approaches agnostic to DoA estimation reliability, exploiting the CDR (3) ensures that the node-specific pdfs of nearby nodes correspond to peaked distributions, whereas the pdfs of distant nodes do not contain significant modes. Therefore, the node-specific source pdfs are fused within the network, relative to a reference point, $\mathbf{x}_{t,n_0}$, as:

$$p\left(\mathbf{x}_{t,n_0} \mid \boldsymbol{\Omega}_{1:t}\right) \approx \prod_{n=1}^{N} p\left(\mathbf{x}_{t,n_0} \mid \omega_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}, \boldsymbol{\Omega}_{1:t-1}\right), \tag{5}$$

where the nodes are assumed to be independent due to their spatial diversity. By probability transformation of (1):

$$\begin{aligned} &p\left(\mathbf{x}_{t,n_0} \mid \omega_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}, \boldsymbol{\Omega}_{1:t-1}\right) \\ &\quad = p\left(f_{n_0}(\mathbf{x}_{t,n}) \mid \omega_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}, \boldsymbol{\Omega}_{1:t-1}\right). \end{aligned}$$

B. Probabilistic Triangulation for Range Inference

In order to estimate the node-specific pdfs required in (5), the density of the Cartesian source position must be updated from new information inferred from the DoA estimates. Contrary to bearing-only tracking, e.g., [37], we propose to infer the missing range estimates by exploiting the spatial diversity of the network nodes. As information is inferred from the DoA estimates, the node-specific pdfs must be propagated in polar coordinates as $p\left(\mathbf{s}_{t,n} \mid \boldsymbol{\Omega}_{1:t}\right)$. The posterior pdf of the polar coordinates can be estimated by 1) predicting the pdf using a model of the source dynamics and 2) updating the pdf with information inferred from the DoA estimates.

1) Prediction: The predicted pdf relative to node n is:

$$p\left(\mathbf{s}_{t,n} \mid \mathbf{\Omega}_{1:t-1}\right) = \int_{\mathbb{R}^3} p\left(\mathbf{s}_{t-1,n} \mid \mathbf{\Omega}_{1:t-1}\right) p\left(\mathbf{s}_{t,n} \mid \mathbf{s}_{t-1,n}\right) d\mathbf{s}_{t-1,n}, \tag{6}$$

where the prior pdf, $p\left(\mathbf{s}_{t,n} \mid \mathbf{s}_{t-1,n}\right)$, models the source dynamics. Assuming the DoA and range are independent:

$$\begin{aligned} p\left(\phi_{t,n} \mid \phi_{t-1,n}\right) &= \mathcal{M}\left(\phi_{t,n} \,|\, \phi_{t-1,n},\, \kappa_{t|t-1}\right) \\ p\left(r_{t,n} \mid r_{t-1,n}\right) &= \mathcal{U}\left(r_{t-1,n} - \rho_{t|t-1}, r_{t-1,n} + \rho_{t|t-1}\right) \end{aligned} \tag{7}$$

where $\mathcal{U}(\cdot)$ denotes the uniform distribution, and the transition concentration parameter and range, $\kappa_{t|t-1}$ and $\rho_{t|t-1}$ respectively, capture the source dynamics between time frames.

The range pdf is approximated by importance sampling J hypotheses, $\{\rho_{t,n}^{(j)}\}_{j=1}^{J}$, of $r_{t,n}$ from (7). Hence, the predicted pdf is given by the Probability Mass Function (PMF):

$$p\left(\mathbf{s}_{t,n} \mid \mathbf{\Omega}_{1:t-1}\right) \approx \sum_{j=1}^{J} \delta_{\rho_{t,n}^{(j)}}\left(r_{t,n}\right)\, \mathcal{M}\left(\phi_{t,n} \,|\, \tilde{\mu}_{t,n}^{(j)},\, \tilde{\kappa}_{t,n}^{(j)}\right), \tag{8}$$

where δ_a is the Dirac measure at a given state a, and the predicted mean, $\tilde{\mu}_{t,n}^{(j)}$, and covariance, $\tilde{\kappa}_{t,n}$, are given by (see Section I.A of the supplementary material to this paper):

$$\tilde{\mu}_{t,n}^{(j)} = \mu_{t-1,n}^{(j)}, \quad \tilde{\kappa}_{t,n} = A^{-1}\left(A(\kappa_{t-1,n}) A(\kappa_{t|t-1})\right), \tag{9}$$

with $A(\kappa) \triangleq {}^{I_1(\kappa)}/_{I_0(\kappa)}$ [34], and where, for any $\bar{r}$ [35]:

$$A^{-1}(\bar{r}) \approx \frac{2\bar{r} - \bar{r}^3}{1 - \bar{r}^2}. \tag{10}$$

2) Update: Using Bayes's theorem:

$$p\left(\mathbf{s}_{t,n} \mid \omega_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}, \mathbf{\Omega}_{1:t-1}\right) = \frac{p\left(\omega_{t,n} \mid \mathbf{s}_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}\right) p\left(\mathbf{s}_{t,n} \mid \mathbf{\Omega}_{1:t-1}\right)}{\int p\left(\omega_{t,n} \mid \mathbf{s}_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}\right) p\left(\mathbf{s}_{t,n} \mid \mathbf{\Omega}_{1:t-1}\right) d\mathbf{s}_{t,n}}. \tag{11}$$

Substituting (8) and (2) into (11):

$$p\left(\mathbf{s}_{t,n} \mid \omega_{t,n}, \kappa_{t,n}^{(\mathrm{DoA})}, \mathbf{\Omega}_{1:t}\right) = \sum_{j=1}^{J} w_{t,n}^{(j)}\, \delta_{\rho_{t,n}^{(j)}}\left(r_{t,n}\right) \mathcal{M}\left(\phi_{t,n} \,|\, \mu_{t,n}^{(j)},\, \kappa_{t,n}^{(j)}\right), \tag{12}$$

where the updated mean, $\mu_{t,n}^{(j)}$, concentration, $\kappa_{t,n}^{(j)}$, and weights, $w_{t,n}^{(j)}$, are (see supplementary material, Section I.B)[1]

$$\mu_{t,n}^{(j)} = -\tan^{-1}\left(-\frac{\tilde{\kappa}_{t,n}^{(j)} \cos\tilde{\mu}_{t,n}^{(j)} + \kappa_{t,n}^{\mathrm{DoA}} \cos\omega_{t,n}}{\tilde{\kappa}_{t,n}^{(j)} \sin\tilde{\mu}_{t,n}^{(j)} + \kappa_{t,n}^{\mathrm{DoA}} \sin\omega_{t,n}}\right), \tag{13a}$$

$$\kappa_{t,n}^{(j)} = \sqrt{[\tilde{\kappa}_{t,n}^{(j)}]^2 + [\kappa_{t,n}^{\mathrm{DoA}}]^2 + 2\tilde{\kappa}_{t,n}^{(j)}\, \kappa_{t,n}^{\mathrm{DoA}} (\cos\left(\tilde{\mu}_{t,n}^{(j)} - \omega_{t,n}\right))}, \tag{13b}$$

$$w_{t,n}^{(j)} = \frac{\tilde{w}_{t,n}^{(j)}}{\sum_{\ell=1}^{J} \tilde{w}_{t,n}^{(\ell)}} \quad \text{and} \quad \tilde{w}_{t,n}^{(j)} = \frac{I_0(\kappa_{t,n}^{(j)})}{I_0(\tilde{\kappa}_{t,n}^{(j)})}. \tag{13c}$$

[1]A von Mises filter for bearing-only tracking was previously proposed in [37]. It is important to note that the additive term in [37, Eq. (8)] may result in $|\mu_{t,n}| > \pi$. In contrast, by definition of the inverse tangent, (13a) ensures valid directions within the region of support, $\mu_{t,n} \in [-\pi, \pi]$.

Algorithm 1 PROST incorporating DoA reliability

```
for t = 1, ..., ∞ do
    for n = 1, ..., N do
        for k = 1, ..., K do
            Estimate DoA, ω_{t,n,k} [38];
            Estimate CDR, Γ_{t,n,k}, [36] and κ^(DoA)_{t,n,k}, (3);
        end for
        Spectral fusion of ω_{t,n} and κ^(DoA)_{t,n}, Section II-B;
        for j = 1, ..., J do
            Transform x^(j)_{t-1,n_0} to s^(j)_{t-1,n}, (1);
            Predict μ̃^(j)_{t,n}, κ̃^(j)_{t,n}, (9);
            Update w^(j)_{t,n}, μ^(j)_{t,n}, κ^(j)_{t,n}, (13);
            Transform s^(j)_{t,n} to x^(j)_{t,n_0}, (1);
        end for
    end for
    Network fusion, (5);
    Resample node-specific pdfs, Section III-B3;
end for
```

3) Network Fusion: Due to the product in (5), network fusion requires a continuous representation of the discrete PMF in (12). A continuous approximation, $\hat{p}\left(\mathbf{x}_{t,n_0} \mid \mathbf{\Omega}_{1:t}\right)$, of each node-specific pdf is obtained from the weighted Kernel Density Estimate (KDE) [39]. The resulting KDEs are used to evaluate (5). For propagation to $t+1$, the node-specific PMFs resulting from the network fusion are extracted by randomly selecting, with equal probability, J components of the non-fused PMFs in (12) from within the intersection region in (5).

Although the unmeasured range is only predicted in (11), resampling from the network posterior ensures that only stochastically relevant range hypotheses are propagated. Hence, the source positions are probabilistically triangulated by exploiting the spatial diversity of the nodes. PRObabilistic Source Triangulation (PROST) is summarized in Algorithm 1.

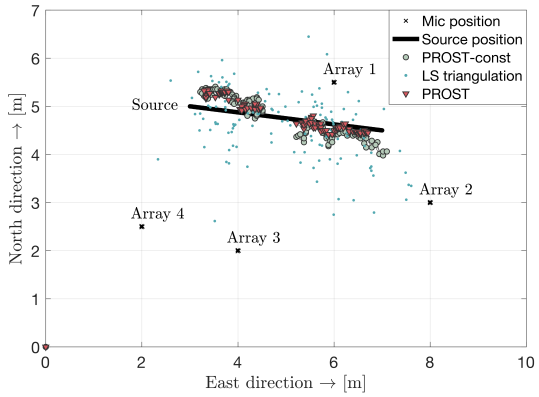
IV. Performance Evaluation

A. Experimental Setup

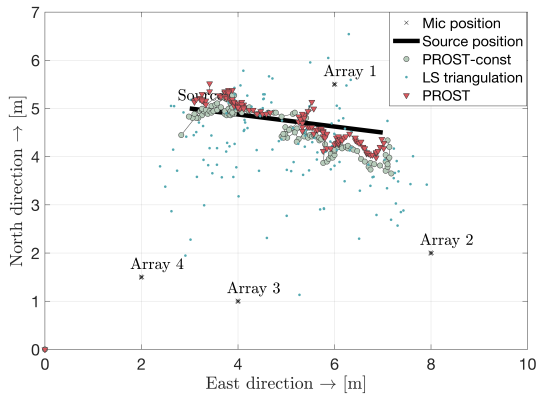
Simulations of a $10 \times 7 \times 2.5$ m^3 room at 16 kHz sampling frequency are used for evaluation. The positions of a source moving at 0.5 m/s are generated at 32 sample intervals along a straight line trajectory. The distributed sensor network contains four circular microphone arrays, each with 3 microphones spaced by 0.025 m. The array positions are simulated for a mid-range scenario (S1) and a far-range scenario (S2) as detailed in Table I. The Room Impulse Responses (RIRs) for each source-sensor configuration are simulated using the image-source method [40] for a reverberation time [26] of 0.5 s. The RIRs are convolved with anechoic speech of 8 s duration, consisting of two utterances by a female talker. The convolved signals are distorted by sensor and isotropic noise [41] with signal-to-noise ratios of 20 dB and 40 dB respectively. The STFT is evaluated for 64 ms frame lengths, 1024 discrete Fourier transform points, and using a Hamming window with 25% overlap. The Voice Activity Detector (VAD) in [42] is evaluated for a window length of 10 ms. Due to the compact array aperture, the propagation delay between

TABLE I
AVERAGE TRACKING ACCURACY FOR BOTH SCENARIOS, S1 AND S2, AND VOICE ACTIVITY PERIODS, P1 AND P2.

	$\mathbf{q}_1$ [m]	$\mathbf{q}_2$ [m]	$\mathbf{q}_3$ [m]	$\mathbf{q}_4$ [m]	LS				PROST-const				PROST			
					Mean [m]		Std [m]		Mean [m]		Std [m]		Mean [m]		Std [m]	
					P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
S1	6, 5.5, 1.8	8, 3, 1.8	4, 2, 1.8	2, 2.5, 1.8	0.63	0.68	0.44	0.5	0.48	0.26	**0.1**	0.13	**0.38**	**0.18**	0.12	**0.06**
S2	6, 5.5, 1.8	8, 2, 1.8	4, 1, 1.8	2, 2.5, 1.8	0.97	1.12	0.66	0.65	**0.46**	0.52	0.17	0.18	0.5	**0.32**	**0.1**	**0.12**



(a) Mid-range scenario (S1)



(b) Far-range scenario (S2)

Fig. 1. Birsdeye view of the acoustic scene map for the (a) mid-range and (b) far-range scenarios.

microphones for each array is negligible in the context of VAD. The VAD is hence evaluated only for channel 3 at each node. The VAD results are combined within the network by evaluating the maximum start time and minimum duration of each Voice Activity Period (VAP) across all nodes. During voice active periods, one CDR and one DoA per time frame and frequency bin are estimated. The DoAs are estimated by minimizing the discrepancy between the observed and expected inter-microphone phase differences as proposed in [38]. The CDR is estimated as a function of the spatial coherence between microphone 1 and 2 for each node using the approach in [36]. The corresponding estimates per frame are evaluated by the median across all frequency bins. $\kappa_{t,n}^{(\mathrm{DoA})}$ is obtained using (3) with $l_{min} = 0$, $l_{max} = 25$, $c = 6$, and $\rho = -2$. Algorithm 1 is initialized for each voice active period by drawing $J = 350$ source position hypotheses for each node. The initial hypotheses are sampled along each $\omega_{t,n}$ between $\mathbf{q}_n$ and the intersection of the DoA vector with the room boundary. For all subsequent frames, Algorithm 1 is evaluated using the transition parameters $\rho_{t|t-1} = 1.5$ m and $\kappa_{t|t-1} = 500$, chosen to enforce confidence in the dynamical model for robustness against false DoA estimates.

The performance of PROST is compared against 1) PROST-const, which uses a constant $\kappa_{t,n} = 5$ corresponding to the mean of $\kappa_{t,n}$ estimates during the first utterance; and 2) Least Squares Triangulation (LST) from the DoA estimates.

B. Results

1) Mid-range Scenario: The VAD detects two VAPs, P1 and P2. The DoA estimates correspond to average errors of $10°$ for P1 and $14°$ for P2. Fig. 1a compares the ground truth source positions against the estimates of PROST and the two benchmarks. LST produces a large-volume cloud of points. By tracking the source position across time, PROST-const results in tracks with large variation between positions. PROST results in smooth tracks near the source by incorporating the DoA reliability, achieving improvements in position accuracy of up to 0.5 m over LST and 0.1 m over PROST-const (see Table I).

2) Far-Range Scenario: Fig. 1b shows the estimated scene map for the far-range scenario. For the first VAP, all four arrays initially correspond to large source-sensor distances of over 4.5 m, such that PROST-const and PROST lead to comparable performance results. In the second VAP, Array 1 is within 1.04 m, leading to reliable DoA estimates and hence increased concentration parameters compared to the first VAP. PROST leads to an improvement of 0.2 m and 0.8 m compared to PROST-const and LST respectively. Hence, the benefits of incorporating the DoA reliability increase for far-range scenarios where at least one node is nearby the source.

V. CONCLUSIONS

We proposed to exploit the CDR as the DoA reliability for distributed acoustic tracking. The CDR is incorporated as the concentration parameter of the DoA likelihood function, modelled by a vM distribution. The source DoAs are then tracked in time at each individual node using a vM filter. To infer the unmeasured source-sensor range, the vM filter is evaluated for a cloud of uninformative range hypotheses. By network fusion, spatial diversity of the nodes is exploited in order to probabilistically triangulate stochastically relevant source positions, and hence range hypotheses. Realistic simulation results demonstrate improvements of up to 39% compared to the classical approach of a constant concentration parameter, and up to 74% compared to least-squares source triangulation.

REFERENCES

[1] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[2] C. Evers and P. A. Naylor, "Optimized self-localization for SLAM in dynamic scenes using probability hypothesis density filters," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 863–878, Feb. 2018.

[3] ——, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process. (in review)*, 2017.

[4] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding speakers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[5] C. Meesookho, U. Mitra, and S. Narayanan, "On energy-based acoustic source localization for sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 365–377, Jan. 2008.

[6] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 439–443, Feb. 2013.

[7] Y. Dorfan, A. Plinge, G. Hazan, and S. Gannot, "Distributed expectation-maximization algorithm for speaker localization in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 682–695, Mar. 2018.

[8] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Joint Workshop Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2011, pp. 127–132.

[9] H. T. Sullivan and S. Sahasrabudhe, "Envisioning inclusive futures: Technology-based assistive sensory and action substitution," *Futures*, vol. 87, pp. 140 – 148, 2017.

[10] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *IEEE Intl. Conf. Digital Sig. Proc. (DSP)*, July 2015, pp. 1206–1210.

[11] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87 – 112, 2015.

[12] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic source localization and beamforming: Theory and practice," *EURASIP J. Adv. Sig. Proc.*, vol. 2003, no. 4, p. 926837, Mar 2003.

[13] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 54 – 67, 2015.

[14] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Distributed 3d source localization from 2d doa measurements using multiple linear arrays," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[15] D. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer, 2016.

[16] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 178–192, Jan. 2017.

[17] O. Hlinka, F. Hlawatsch, and P. M. Djurić, "Consensus-based distributed particle filtering with distributed proposal adaptation," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3029–3041, Jun. 2014.

[18] A. Mohammadi and A. Asif, "Distributed consensus + innovation particle filtering for bearing/range tracking with communication constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 620–635, Feb. 2015.

[19] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Boston: Artech House, 2004.

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

[21] X. Zhong and J. R. Hopgood, "A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2356–2370, Dec. 2015.

[22] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1409–1415, May 2012.

[23] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 4, pp. 725–735, Apr. 2018.

[24] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 281–295, Feb. 2018.

[25] C. Evers, B. Rafaely, and P. A. Naylor, "Speaker tracking in reverberant environments using multiple directions of arrival," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Mar. 2017, pp. 91–95.

[26] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[27] H. Kuttruff, *Room Acoustics, Fifth Edition*. CRC Press, Jun. 2009.

[28] R. K. Cook, R. V. Waterhouse, R. D. Berendt, E. Seymour, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *J. Acoustic. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[29] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug 2011, pp. 1347–1351.

[30] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2018.

[31] M. Taseska and E. A. P. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, Mar. 2018.

[32] ——, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, p. 60, Aug 2017.

[33] ——, "Minimum Bayes risk signal detection for speech enhancement based on a narrowband DOA model," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 539–543.

[34] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Wiley-Blackwell, 1999.

[35] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Dec. 2005.

[36] O. Thiergart, G. D. Galdo, and E. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *The J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2337–2346, Dec 2012.

[37] I. Marković, J. Ćesić, and I. Petrović, "Von Mises mixture PHD filter," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2229–2233, Dec 2015.

[38] O. Thiergart, W. Huang, and E. A. P. Habets, "A low complexity weighted least squares narrowband DOA estimator for arbitrary array geometries," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 340–344.

[39] F. J. G. Gisbert, "Weighted samples, kernel density estimators and convergence," *Empirical Economics*, vol. 28, no. 2, pp. 335–351, Apr 2003.

[40] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[41] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 112, no. 6, pp. 3464–3470, Dec. 2007.

[42] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan 1999.