Augmented Intensity Vectors for Direction of Arrival Estimation in the Spherical Harmonic Domain

Sina Hafezi, *Student Member, IEEE,* Alastair H. Moore, *Member, IEEE,* and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Pseudointensity vectors (PIVs) provide a means of Direction of Arrival (DOA) estimation for Spherical Microphone Arrays (SMAs) using only the zeroth and the first-order spherical harmonics. An Augmented Intensity Vector (AIV) is proposed which improves the accuracy of PIVs by exploiting higher order spherical harmonics. We compared DOA estimation using our proposed AIVs against PIVs, Steered Response Power (SRP) and subspace methods where the number of sources, their angular separation, the reverberation time of the room and the sensor noise level are varied. The results show that the proposed approach outperforms the baseline methods and performs at least as accurately as the state-of-the-art method with strong robustness to reverberation, sensor noise and number of sources. In the single and multiple source scenarios tested, which include realistic levels of reverberation and noise, the proposed method had average error of 1.5°and 2°, respectively.**

***Index Terms*—spherical microphone arrays, localization, direction of arrival estimation, spherical harmonic, intensity vector.**

I. Introduction

DIRECTION of Arrival (DOA) estimation is an important acoustic signal processing task and has been used in areas including spatial filtering, source separation, source tracking, environment mapping, dereverberation and speech enhancement. As such it can be useful in applications such as teleconferencing, meeting diarization, robot audition and hearing aids. In a real-world scenario, various factors such as coherent reflections, sensor noise and the presence of multiple simultaneously active sources degrade the performance of DOA estimation.

In this work we focus on Spherical Microphone Arrays (SMAs) which have become popular [1]-[17] in contexts where their ability to analyse a sound field with equal resolution in all directions is important. By representing the sound field as a Spherical Harmonic (SH) expansion, the problem formulation can be made independent of the specific geometry of the SMA making the methods described here widely applicable. The first works on DOA estimation using SMA can be found in [18] and the first works on signal processing for SMAs are presented in [19], [2], [20].

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, UK (e-mails: {s.hafezi14, alastair.h.moore, p.naylor}@imperial.ac.uk).

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

Manuscript received December 21, 2016; revised May 19, 2017.

DOA estimation methods are generally categorised into four main groups. (1) Steered Response Power (SRP) methods [21] in which a beamformer such as Plane-Wave Decomposition (PWD) [22], [14], [11] is used to steer a beam into multiple directions and the source direction is found as the direction with the highest power. Due to limitations on spatial resolution, SRP methods can fail to localize closely spaced sources [23]; (2) Sub-space methods [24], [25] such as MUSIC (MUltiple SIgnal Classification) [6], [26] employ the eigenvalue decomposition to decompose the noisy signals spatial covariance matrix into the signal and noise subspaces, which then are used to estimate the DOAs. Although they are somewhat robust to noise and reverberation, their performance normally depends on a user-defined threshold. (3) Maximum Likelihood (ML) methods [27], [28] employ optimization to minimize a defined cost function [29], [30]. They mainly require accurate statistical models (e.g. for noise) and can be computationally expensive. Statistical DOA estimation methods using directional sparsity of sound sources have also been proposed in [31], [32]. (4) Intensity-based methods [33] determine the direction and the magnitude of the flow of sound energy within a narrow frequency band [17], [34], [35], [16], [36], [37]. Pseudointensity vector (PIV) [17] methods use sound field information with low spatial resolution. PIVs are computationally efficient and have been shown in [17] to offer good localization accuracy for a single source in the absence of strong reflections. In common with most localization algorithms however, localization accuracy is reduced as the level of reverberation, the sensor noise or the number of sources increase [16], [38], [37]. An extension of PIV is Subspace PIV (SSPIV) [37] where the low order (≤ 1) spatial information of signal subspace is used to enhance the accuracy of PIV DOA estimates.

In [38], [23] we proposed Augmented Intensity Vectors (AIVs) which exploits eigenbeams of order ≥ 2 to form vectors with improved DOA accuracy compared to PIVs. These vectors are obtained using spatially constrained grid search to minimize a cost function with initialisation derived from PIVs. By including higher order information in the DOA estimation, the method is more accurate and robust to reverberation, noise and multiple speakers. However AIVs suffer from spatial limitation due to limited search window size. On the other hand, full grid search causes high computational cost. In this paper, we propose an alternative solution for AIV in which the gradient descent approach is used in order to overcome

the problem of spatial limitation of grid search approach while keeping the computational cost low. A theoretical analysis of the optimized DOA error as a function of noise and harmonic order is also presented.

This paper is structured as follows. Section II briefly reviews the background theory of spherical harmonics, the baseline methods, PIV, SSPIV and SRP, and the state-of-the-art, MUSIC with Direct Path Dominance (DPD) test. Section III presents our proposed method, a theoretical error analysis and two variations of solution to the problem based on grid search and gradient descent optimization. Section IV evaluates the accuracy and robustness of our method compared with the baseline and the state-of-the-art methods for single and multiple source scenarios. We also evaluate the effect of the maximum spherical harmonic order on the accuracy of localization. Section V presents the visual performance and accuracy of methods using real recording signals in a real room. Finally section VI provides a rough analysis of the computational complexity of each method.

II. Technical Background

In this section, we review the Spherical Harmonic Domain (SHD), Spherical Harmonic Transform (SHT), its discrete approximation as applied to SMAs, and briefly introduce the methods used in our comparative evaluation.

A. Spherical Harmonic Domain

Consider the sound pressure field $p(\tau, k, r, \Omega)$ which is a function of wavenumber k, time frame τ and the point location $(r, \Omega) = (r, \theta, \varphi)$ in spherical coordinates with range r, inclination θ, and azimuth φ. The SHT of this field is given by [39]

$$p_{lm}(\tau, k, r) = \int_{\Omega \in S^2} p(\tau, k, r, \Omega) Y_{lm}^{*}(\Omega)\, d\Omega, \tag{1}$$

where $\int_{\Omega \in S^2} d\Omega = \int_0^{2\pi} \int_0^{\pi} \sin(\theta)\, d\theta d\varphi$, and $(.)^*$ denotes the complex conjugate.

The spherical harmonic basis functions $Y_{lm}(\Omega)$ of order l and degree m (satisfying $|m| \leq l$) are given by [39]

$$Y_{lm}(\Omega) = \sqrt{\frac{(2l+1)}{4\pi} \frac{(l-m)!}{(l+m)!}} P_{lm}(\cos(\theta))\, e^{im\varphi}, \tag{2}$$

where P_{lm} is the associated Legendre function and $i^2 = -1$.

Using the inverse SHT, the sound pressure field can be reconstructed as [12]

$$p(\tau, k, r, \Omega) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} p_{lm}(\tau, k, r) Y_{lm}(\Omega), \tag{3}$$

where the coefficients p_{lm} are spherical harmonic coefficients.

Considering a SMA with radius r_a and Q microphones, each with angle Ω_q, the integral in (1) is approximated as

$$p_{lm}(\tau, k, r_a) \approx \sum_{q=1}^{Q} w_{q,lm} p(\tau, k, r_a, \Omega_q), \tag{4}$$

where the weights $w_{q,lm}$ are chosen to ensure that (4) is an accurate approximation of (1), which in the case of a uniform sensor distribution are simply $w_{q,lm} = \frac{4\pi}{Q} Y_{lm}^{*}(\Omega_q)$. To avoid spatial aliasing due to discrete sampling of the sound field the lower bound on Q is [7]

$$Q \geq (L+1)^2, \tag{5}$$

where L is the maximum SH order considered. Further information regarding the analysis of spatial aliasing errors and the selection of an appropriate spatial sampling scheme can be found in [40].

The range dependence of the SH coefficients is a function of wavenumber and array configuration, for example, open or rigid baffle. This dependence is captured by the mode strength, which for microphones mounted on a rigid sphere of radius $r = r_a$, as used in our study, is [39]

$$b_l(kr_a) = 4\pi i^l \left[j_l(kr_a) - \frac{j_l^{'}(kr_a)}{h_l^{(2)'}(kr_a)} h_l^{(2)}(kr_a) \right], \tag{6}$$

where j_l is the spherical Bessel function of order l, $h_l^{(2)}$ is the spherical Hankel function of the second kind and of order l, and $(.)^{'}$ denotes the first derivative.

Eigenbeams with mode strength compensation are obtained as

$$a_{lm}(\tau, k) = \frac{p_{lm}(\tau, k, r_a)}{b_l(kr_a)}. \tag{7}$$

In the case of a single plane wave S with DOA $\Omega_u = (\theta_u, \varphi_u)$, the compensated eigenbeams are simply [40], [12]

$$a_{lm}(\tau, k) = S(\tau, k) Y_{lm}^{*}(\Omega_u). \tag{8}$$

B. PWD-SRP in SHD

The baseline DOA estimator used for comparison is the SRP approach [21] implemented in the SHD using a PWD beamformer [22]. The output of the beamformer steered to look direction Ω is given as [11]

$$y(\tau, k, \Omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm}(\tau, k) Y_{lm}(\Omega). \tag{9}$$

The narrow-band power response of SRP at bin (τ, k) is

$$P_{SRP}(\tau, k, \Omega) = \mid y(\tau, k, \Omega) \mid^2 . \tag{10}$$

In the conventional wideband SRP, the spatial spectrum is obtained as $\sum_{\tau,k} P_{SRP}(\tau, k, \Omega)$ in which the position of the peaks represent the estimated DOAs. In [38] and [23], we have shown that the conventional wideband PWD-SRP fails in localization of multiple sources especially for adjacent sources as the summation of power spectra can result in merging of adjacent peaks associated to different sources resulting in an erroneous estimated DOA between the sources as seen in Fig. 1. In order to preserve the DOA information at each TF bin and obtain a DOA per TF bin, we use the narrow-band PWD-SRP [41] in which the global peak represents the estimated narrow-band DOA

$$\Omega_{SRP}(\tau, k) = \arg\max_{\Omega} P_{SRP}(\tau, k, \Omega). \tag{11}$$

C. PIVs

Sound intensity is a measure of the flow of sound energy through a surface per unit area, in a direction perpendicular to this surface. The intensity vector I, which defines the magnitude and the direction of the energy flow can be determined by calculating the flow of sound energy through the three unit surfaces perpendicular to the Cartesian axes as [42]

$$\mathbf{I} = \frac{1}{2}\Re\left\{q^*\mathbf{v}\right\}, \tag{12}$$

where q is the sound pressure, $\mathbf{v} = [v_x, v_y, v_z]^T$ is the particle velocity vector in Cartesian coordinates, and $\Re\{.\}$ denotes the real part of a complex number.

Due to the difficulty of measuring the particle velocity at the sensors, PIV approximates the intensity vector using low-order ($l \leq 1$) eigenbeams. PIV at time frame τ and wavenumber k is calculated as [17]

$$\mathbf{I_{piv}}(\tau,k) = \frac{1}{2}\Re\left\{a_{00}(\tau,k)^* \begin{bmatrix} D_{-x}(\tau,k,\mathbf{a_{lm}}) \\ D_{-y}(\tau,k,\mathbf{a_{lm}}) \\ D_{-z}(\tau,k,\mathbf{a_{lm}}) \end{bmatrix}\right\}, \tag{13}$$

where

$$D_\nu(\tau,k,\mathbf{a_{lm}}) = \sum_{m=-1}^{1} Y_{1m}(\phi_\nu)a_{1m}(\tau,k), \quad \nu \in \{-x,-y,-z\} \tag{14}$$

are dipoles steered in the negative direction of Cartesian axes, given by $\phi_{-x} = (\pi/2,\pi)$, $\phi_{-y} = (\pi/2,-\pi/2)$ and $\phi_{-z} = (\pi,0)$ and $\mathbf{a_{lm}} = \left[a_{00}, a_{1(-1)}, a_{1(0)}, a_{1(1)}, \ldots, a_{LL}\right]^T$ denotes the set of eigenbeams with maximum SH order L.

The estimated DOA unit vector $\mathbf{u}(\tau,k)$ pointing towards the source is given by

$$\mathbf{u}(\tau,k) = -\frac{\mathbf{I_{piv}}(\tau,k)}{\|\mathbf{I_{piv}}(\tau,k)\|}, \tag{15}$$

where $\|.\|$ indicates ℓ_2-norm.

D. SSPIVs

The PIVs use low order SHs ($l \leq 1$) and only work well in single source scenarios. In [37] authors extend the concept of PIVs by taking advantage of higher order SHs and frequency smoothing to enhance the accuracy of DOAs.

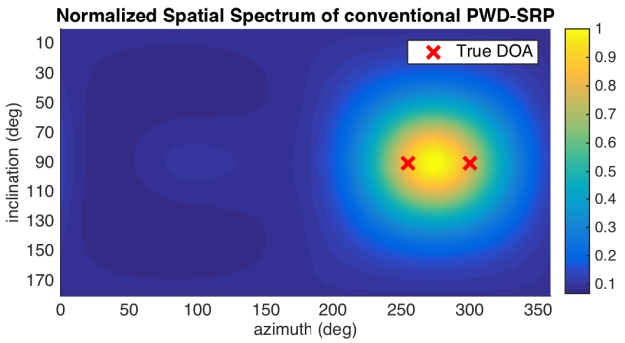

Fig. 1. Normalized spatial spectrum of conventional wideband PWD-SRP.

In the TF domain, the covariance matrix of the observed eigenbeams is given as

$$\mathbf{R}(\tau,k) = E\left[\mathbf{a_{lm}}(\tau,k)\mathbf{a_{lm}}^H(\tau,k)\right], \tag{16}$$

where $E[.]$ denotes the expectation and $(.)^H$ indicates Hermitian transpose. In single source scenario using Singular Value Decomposition (SVD), the covariance matrix of the observed noisy eigenbeams is decomposed as

$$\mathbf{R}(\tau,k) = \mathbf{U_s}\mathbf{\Sigma_s}\mathbf{U_s}^H + \mathbf{U_v}\mathbf{\Sigma_v}\mathbf{U_v}^H, \tag{17}$$

where $\mathbf{U_s} = \left[\hat{a}_{00}, \hat{a}_{1(-1)}, \hat{a}_{1(0)}, \hat{a}_{1(1)}, \ldots, \hat{a}_{LL}\right]^T$ is the one-dimensional signal subspace matrix, $\mathbf{U_v}$ is the noise subspace with $(L+1)^2-1$ dimensions and $\mathbf{\Sigma}$ is the rectangular diagonal singular value matrix. Note that (τ,k) are omitted here for notational simplicity.

Using the estimated de-noised low order eigenbeams with PIV formulation, the SSPIV is given by [37]

$$\mathbf{I_{sspiv}}(\tau,k) = \frac{4\pi\sqrt{4\pi}}{3}\Re\left\{\hat{a}_{00}(\tau,k)^* \begin{bmatrix} D_{-x}(\tau,k,\mathbf{U_s}) \\ D_{-y}(\tau,k,\mathbf{U_s}) \\ D_{-z}(\tau,k,\mathbf{U_s}) \end{bmatrix}\right\}. \tag{18}$$

Although only low order SHs ($l \leq 1$) components of $\mathbf{U_s}$ are used in (18), their value depends on high order eigenbeams in (16) and (17).

E. DPD-MUSIC in SHD

A state-of-the-art DOA estimator used for comparison is DPD-MUSIC algorithm, also implemented in the SHD [6]. DPD-MUSIC consists of two stages of DPD test and MUSIC.

DPD test [6] is proposed to identify the Time Frequency (TF) regions where the direct path of a single source is estimated to be significantly dominant and the significance of dominance is defined by a user-defined threshold.

The selected set of TF bins in DPD test is given as

$$\Upsilon_{\mathrm{DPD}} = \{(\tau,k) : \mathrm{erank}\left(\mathbf{R}(\tau,k)\right) = 1\}, \tag{19}$$

where

$$\mathrm{erank}\left(\mathbf{R}(\tau,k)\right) = 1 \text{ if } \eta_{DPD}(\tau,k) > \epsilon \tag{20}$$

is the effective rank, the Singular Value Ratio (SVR) η_{DPD} is the ratio of the largest and the second largest singular values of $\mathbf{R}$ in (17) and ϵ is a threshold.

Only in the TF bins passed by the DPD test, the well-known MUSIC method estimates a DOA using the noise subspace of the spatial covariance matrix. Using the estimated noise subspace in (17), the narrow-band MUSIC spectrum for a single source is given as [24]

$$P_{MUSIC}(\tau,k,\Omega) = \frac{1}{\|\mathbf{U_v}^H(\tau,k)\mathbf{Y_{lm}}^*(\Omega)\|^2}, \tag{21}$$

where $\mathbf{Y_{lm}} = \left[Y_{00}, Y_{1(-1)}, Y_{1(0)}, Y_{1(1)}, \ldots, Y_{LL}\right]^T$ are the column vector of SH basis functions given in (2). Note that (Ω) are omitted for notational simplicity.

The two alternative approaches [6] to obtain DOAs in DPD-MUSIC are discussed next.

1) Incoherent DPD-MUSIC: In the first approach the MUSIC spectra in (21) are simply summed over the selected TF bins $(\tau,k)\in\Upsilon_{DPDtest}$ so that

$$P_{incoh-MUSIC}(\Omega)=\sum_{(\tau,k)\in\Upsilon_{DPDtest}} P_{MUSIC}(\tau,k,\Omega), \quad (22)$$

where the set of N highest peaks in the final spectrum indicates the overall estimated DOAs.

2) Coherent DPD-MUSIC: The second approach performs coherent fusion of the directional information from the selected TF bins. The set of one dimensional signal spaces from the selected TF bins, $\{\mathbf{U_s}(\tau,k)\}_{(\tau,k)\in\Upsilon_{\mathrm{DPD}}}$, are clustered using one-run K-means clustering with random initialization into N clusters with centroids $\{\mathbf{U_s}^n\}_{n=1}^N$ where each centroid signal space is associated with one speaker. The DOA of each individual speaker is selected as the global peak in the coherent MUSIC spectrum of the speaker n which is given as

$$\begin{aligned}P^n_{coh-MUSIC}(\Omega)&=\frac{1}{\|(\mathbf{U_v}^n)^H\mathbf{Y_{lm}}^*(\Omega)\|^2} \quad (23)\\ &=\frac{1}{\mathbf{Y_{lm}}^T(\Omega)(I-\mathbf{U_s}^n(\mathbf{U_s}^n)^H)\mathbf{Y_{lm}}^*(\Omega)}.\end{aligned}$$

III. Augmented Intensity Vector method

The PIVs only use zeroth- and first-order SHs ignoring the higher order SHs. Since the spatial frequency of Y_{lm} increases with the SH order, employing higher order information increases the spatial resolution. Therefore in this section we propose to estimate a vector for each TF bin which uses information from higher order $(l>1)$ spherical harmonics to augment the PIV.

A. Signal Model

Consider a plane wave $S(\tau,k)$ with DOA $\Omega_u=(\theta_u,\varphi_u)$ arriving from a single source in the far-field in an anechoic environment. From (8) the eigenbeams of the sound field are

$$a_{lm}(\tau,k)=S(\tau,k)Y^*_{lm}(\Omega_u)+n_{lm}(\tau,k), \quad (24)$$

where $n_{lm}(\tau,k)$ represents the noise.

In case of a noise-free scenario, $n_{lm}=0$, considering (24) for $l\in[0,L]$ and $-l\le m\le l$, we would have $(L+1)^2$ complex equations with two unknowns S and Ω_u. In this case even for $L=1$ we would have an overdetermined system of equations in which the solution can be accurately obtained. By increasing L we still achieve the same solution that is the accurate true DOA.

Considering the noisy case with non-zero and unknown n_{lm}, we have an underdetermined system of equations. In such scenario, we aim to find the Ω which best satisfies all the $(L+1)^2$ equations of (24) for up to SH order L.

B. Cost function

The zeroth SH order has the noise-reducing characteristic since the noise signals at the individual sensors are averaged and reduced as in (4). For spatially-white noise this reduction is approximately 10 dB. Using this noise-reducing property of $l=0$, we assume $n_{00}(\tau,k)=0$ (no noise or reverberation only at $l=0$), which for moderate sensor noise level is a suitable approximation, to approximate $S(\tau,k)$ by substituting (2) into (24) for $l=0$ and $m=0$

$$S(\tau,k)=\sqrt{4\pi}a_{00}(\tau,k). \quad (25)$$

Substituting (25) into rearranged (24), for an arbitrary look direction Ω, we define a direction-dependant error $E_{lm}(\tau,k,\Omega)$

$$E_{lm}(\tau,k,\Omega)=a_{lm}(\tau,k)-\sqrt{4\pi}a_{00}(\tau,k)Y^*_{lm}(\Omega), \quad (26)$$

which leads to our proposed cost function

$$\Psi(\tau,k,\Omega)=\sum_{l=0}^{L}\sum_{m=-l}^{l}\mid E_{lm}(\tau,k,\Omega)\mid^2. \quad (27)$$

The optimized DOA $\Omega_{aiv}(\tau,k)$ is

$$\Omega_{aiv}(\tau,k)=\arg\min_{\Omega}\Psi(\tau,k,\Omega). \quad (28)$$

To form the Augmented Intensity Vector (AIV) we combine the optimized direction $\Omega_{aiv}(\tau,k)$ with the norm of the original PIV in (13)

$$\mathbf{I_{aiv}}(\tau,k)=-\mathbf{u_{aiv}}(\tau,k)\|\mathbf{I_{piv}}(\tau,k)\|, \quad (29)$$

where $\mathbf{u}_{aiv}(\tau,k)$ is the Cartesian unit vector of $\Omega_{aiv}(\tau,k)$.

Figure 4 shows an example which demonstrates that $\Psi(\tau,k,\Omega)$ is non-convex. However, calculation of (27) over all possible directions at each TF bin is computationally expensive. We first provide a theoretical error analysis in the presence of noise, then present our previously published grid search approach and finally propose our gradient descent approach, both of which use the PIV solution to form an initial estimate for optimization.

C. DOA Error Analysis

In this section we present a theoretical DOA error analysis for the noise-free and noisy scenarios. For the formulation in this section, (τ,k) are omitted for notational simplicity.

In a noise-free scenario as in (8), consider $\tilde{a}_{lm}=SY^*_{lm}(\Omega_u)$ as the clean eigenbeams of the direct path. For an arbitrary look direction Ω, we have the clean eigenbeam error function

$$\begin{aligned}\tilde{E}_{lm}(\Omega)&=\tilde{a}_{lm}-SY^*_{lm}(\Omega)\\ &=SY^*_{lm}(\Omega_u)-SY^*_{lm}(\Omega_u)g^*_{lm}(\Omega_u,\triangle\Omega)\\ &=SY^*_{lm}(\Omega_u)\left(1-g^*_{lm}(\Omega_u,\triangle\Omega)\right), \quad (30)\end{aligned}$$

where by using (2) we have

$$g_{lm}(\Omega_u,\triangle\Omega)=\frac{P_{lm}(\cos(\theta_u+\triangle\theta))}{P_{lm}(\cos(\theta_u))}e^{im\triangle\varphi}, \quad (31)$$

where $\Delta\Omega=\Omega-\Omega_u$.

Now assume the noisy scenario where the noisy cost function in (27) is decomposed into clean $\tilde{E}_{lm}$ and the noise eigenbeam n_{lm} resulting in

$$\begin{aligned}\Psi(\Omega)=\sum_{lm}\mid\tilde{E}_{lm}(\Omega)-n_{lm}\mid^2&=\tilde{\Psi}(\Omega)+C_n\\ &+2\sum_{lm}\mid\tilde{E}_{lm}(\Omega)\mid\mid n_{lm}\mid\cos(\Gamma_{lm}(\Omega)), \quad (32)\end{aligned}$$

where $\tilde{\Psi}(\Omega) = \sum_{lm} \mid \tilde{E}_{lm}(\Omega) \mid^2$ is the noise-free cost function, $C_n = \sum_{lm} \mid n_{lm} \mid^2$ is a noise-based constant, $\sum_{lm} = \sum_{l} \sum_{m=-l}^{l}$ and

$$\Gamma_{lm}(\Omega) = \angle \tilde{E}_{lm}(\Omega) - \angle n_{lm} = \angle S + \angle Y_{lm}^*(\Omega_u) + \angle\left(1 - g_{lm}^*(\Omega_u, \triangle\Omega)\right) - \angle n_{lm}, \quad (33)$$

where $\angle(.)$ denotes the phase of complex number.

The derivative of the noisy cost function in (32) is

$$\Psi^{'}(\Omega) = \tilde{\Psi}^{'}(\Omega) + 2\sum_{lm} \mid n_{lm} \mid \left(\mid \tilde{E}_{lm}(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega)\right)\right)^{'}, \quad (34)$$

where $(.)^{'} = \frac{d}{d\Omega}(.)$ is the derivative operator.

For $\triangle\theta \approx 0$ in (31) we have $g_{lm}^*(\Omega_u, \triangle\Omega) = e^{-im\triangle\varphi}$, which results in

$$\left(1 - g_{lm}^*(\Omega_u, \triangle\Omega)\right) = 2\sin\left(\frac{m\triangle\varphi}{2}\right) e^{i(\frac{\pi}{2} - \frac{m\triangle\varphi}{2})}. \quad (35)$$

Substituting (35) into (30) we have

$$\tilde{\Psi}^{'}(\Omega) = \sum_{lm} 2m \mid SY_{lm}^*(\Omega) \mid^2 \sin\left(m\triangle\varphi\right), \quad (36)$$

and

$$\left(\mid \tilde{E}_{lm}(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega)\right)\right)^{'} = m \mid SY_{lm}^*(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega) - \frac{m\triangle\varphi}{2}\right). \quad (37)$$

At the optimized look direction Ω_s, we have $\Psi^{'}(\Omega_s) = 0$, which by substituting (36) and (37) into (34) gives

$$\sum_{lm} \mid SY_{lm}^*(\Omega_u) \mid^2 m\sin\left(m\triangle\varphi_s\right) = \sum_{lm} \mid n_{lm} \mid\mid SY_{lm}^*(\Omega_u) \mid m\sin\left(\Lambda_{lm}(S, \Omega_u, n_{lm}) - m\triangle\varphi_s\right), \quad (38)$$

where $\Lambda_{lm}(S, \Omega_u, n_{lm}) = \angle S + \angle Y_{lm}^*(\Omega_u) - \angle n_{lm}$ is the combined phase from the direct path and the noise eigenbeams.

Simplifying (38) gives

$$\sum_{j=0}^{L} \sqrt{A_j^2 + B_j^2} \sin\left(j\triangle\varphi_s - \arctan\left(\frac{B_j}{A_j}\right)\right) = 0, \quad (39)$$

where

$$A_j = \frac{1}{\mu} \sum_{l=0}^{L} \sum_{|m|=j} \mid m \mid\mid \tilde{a}_{lm} \mid^2 \left(\mu + \cos\left(\Lambda_{lm}\right)\right), \quad (40)$$

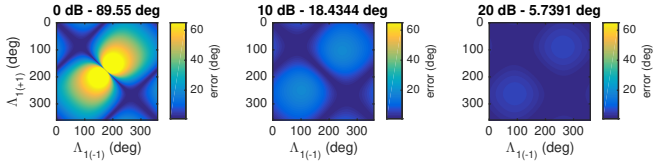

Fig. 2. Azimuth error for all possible combinations of Λ_{lm} for $L = 1$. The title contains the SNR=$\{0, 10, 20\}$ dB and the worst error.

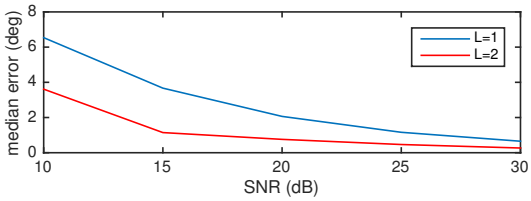

Fig. 3. Median error among all possible combinations of combined phases for varying SNR and maximum SH order L.

and

$$B_j = \frac{1}{\mu} \sum_{l=0}^{L} \sum_{|m|=j} m \mid \tilde{a}_{lm} \mid^2 \sin\left(\Lambda_{lm}\right), \quad (41)$$

where $\mu^2 = \frac{|SY_{lm}^*(\Omega_u)|^2}{|n_{lm}|^2} = \frac{|\tilde{a}_{lm}|^2}{|n_{lm}|^2}$ is the SNR for spatially white noise (equal noise level across all microphones) and is fixed for all (l, m). Note that (S, Ω_u, n_{lm}) are omitted from Λ_{lm} for notational simplicity.

For up to the first SH order ($L = 1$), the azimuth error is

$$\triangle\varphi_s = \arctan(\frac{B_1}{A_1}). \quad (42)$$

Figure 2 presents the azimuth error $\triangle\varphi_s$ in degrees for all possible combinations of Λ_{lm} for $L = 1$ with angle resolution of $\pi/100$ for varying SNR with a random true DOA $\Omega_u = (\phi_u, \theta_u) = (20, 45)^\circ$ and $\mid S \mid = 1$. We can clearly see the decrease in maximum error as the SNR increases.

For up to the second order ($L = 2$), (39) can be simplified into a quartic equation with one variable $\sin\left(\triangle\varphi_s - \arctan\left(B_1/A_1\right)\right)$ and parameters as a function of Aj and B_j. Among the real roots of the quartic equation, we consider the one which results into the minimum $\triangle\varphi_s$. Figure 3 presents the median of azimuth errors $\triangle\varphi_s$ in degrees across all possible combinations of Λ_{lm} with angle resolution of $2\pi/10$ for varying SNR and $L = \{1, 2\}$ with the same true DOA as in Fig. 2. It can be clearly seen that the increase in the maximum SH order of AIV cost function at least doubles the expected accuracy.

D. Grid search optimization

On the discrete spatial domain sampled with 1 degree resolution across azimuth and inclination, we define our search domain as set of look directions $\{\Omega_M\}$ covered within a spherical cap with a chosen radius centred at the initial DOA estimated by PIV. Note that larger search window size and higher grid resolution both increase the accuracy of estimation as well as the computational cost. The optimized DOA $\Omega_s(k)$ is obtained using (28) for $\Omega \in \{\Omega_M\}$.

E. Gradient descent optimization

Grid search solution suffers from spatial limitation or computational cost for small or large window size respectively. Gradient descent approach can be used to overcome the problem of spatial limitation with low computational cost. Starting from the initial angle Ω_0, using the objective cost

function in (27) we can formulate an iterative gradient descent as

$$\Omega_{n+1}(k) = \Omega_n(k) - \gamma_n(k)\nabla\Psi(k,\Omega_n), \quad n \geq 0 \tag{43}$$

where γ_n is the step at the n^{th} iteration and $\nabla(.) = \frac{\partial}{\partial\theta}(.)\hat{\theta} + \frac{\partial}{\partial\varphi}(.)\hat{\varphi}$ denotes the gradient operator.

For a convex cost function, convergence to a global minimum can be guaranteed. However when there are multiple active talkers or an active direct path plus one or more active reflection at the same TF bin, $\Psi(k,\Omega)$ is nonconvex as illustrated in Fig. 4. In this case, convergence to a global minimum is only achieved if the initial point is close enough to the global minimum.

The gradient at angle $\Omega = (\theta, \varphi)$ can be found by substituting (26) into (27) giving

$$\begin{aligned}
\nabla\Psi(k,\Omega) &= \nabla\{\sum_{lm}^{L}|E_{lm}(k,\Omega)|^2\} \\
&= \sum_{lm}^{L}\nabla\{|a_{lm}(k) - \sqrt{4\pi}a_{00}(k)Y_{lm}^*(\Omega)|^2\} \\
&= \sum_{lm}^{L}\nabla\{|a_{lm}(k)|^2 + |\sqrt{4\pi}a_{00}(k)|^2|Y_{lm}^*(\Omega)|^2 \\
&- 2|a_{lm}(k)||\sqrt{4\pi}a_{00}(k)||Y_{lm}^*(\Omega)|\cos\left(\lambda_{lm}(k) - \angle Y_{lm}^*(\Omega)\right)\} \\
&= 4\pi|a_{00}(k)|^2\sum_{lm}^{L}\left\{\nabla\{|Y_{lm}^*(\Omega)|^2\}\right\} - 2\sqrt{4\pi}|a_{00}(k)| \\
&\times\sum_{lm}^{L}\{|a_{lm}(k)|\nabla\{|Y_{lm}^*(\Omega)|\cos\left(\lambda_{lm}(k) - \angle Y_{lm}^*(\Omega)\right)\}\},
\end{aligned} \tag{44}$$

where $\sum_{lm}^{L} = \sum_{l=0}^{L}\sum_{m=-l}^{l}$ is the summation over all the harmonic orders and degrees up to the maximum order L, $\lambda_{lm} = \angle a_{lm} - \angle a_{00}$.

The gradient of the components in the final expression in (44) can be calculated individually for each harmonic order and degree using (2) as shown in Table I.

F. DOA extraction from Intensity Vectors

Considering either PIVs or AIVs, the intensity vectors are calculated for all TF bins. A 2D histogram (inclination vs azimuth) using the quantized directions of all intensity vectors is formed. Note that only the directions of the intensity vectors are used and not their vector length. As shown in [37] in case of multiple arriving plane-waves in a TF bin, it is possible to have an erroneous resulting intensity vector with direction in between of or away from the sources and a norm higher than the intensity vector norm in the presence of a single source depending on the relative amplitude and phase of the impinging plane waves. In order to avoid the accuracy-loss effect of the erroneous intensity vector with high norm, the norms are ignored and only the cardinality of the quantized directions are considered in the histogram. An advantage of histogram is to eliminate the weakening impact of the erroneous directions with low cardinality on the position of the peaks in the histogram although they are present unlike the averaging technique in [17] in which all intensity vectors are summed to estimate the final DOA where erroneous directions reduce the accuracy if they are not spatially diffused.

Due to noisy observations and the presence of multiple irregular peaks, we employ smoothing on the constructed DOA histogram using a Gaussian kernel. The Gaussian kernel, centred on the look direction Ω, for an angle $\Omega_{\theta_i,\varphi_i}$ with inclination θ_i and azimuth φ_i is expressed as

$$K(\Omega, \Omega_{\theta_i,\varphi_i}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\angle(\Omega, \Omega_{\theta_i,\varphi_i})^2}{2\sigma^2}\right), \tag{45}$$

where σ denotes the standard deviation, which is chosen empirically as described in section IV. The kernel is truncated by removing the entries with $K < 0.001$. For N_s sources, the positions of the largest N_s peaks in the smoothed histogram are taken as the estimated DOAs. Figure 5 shows an example of unsmoothed (raw) and smoothed histograms for two sources with $45°$ separation with simulation configuration as in section IV-B2. The choice for the σ, which represents the smoothness of the histogram, is studied in our previous paper [43], which concludes that a suitable σ requires the knowledge of the

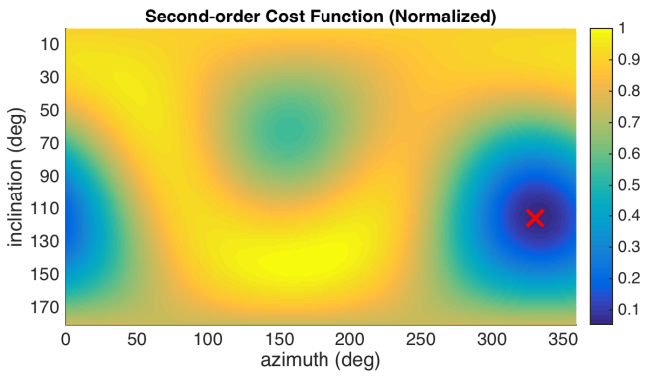

Fig. 4. Normalized second-order cost function for the entire space at a particular TF-bin for a single source with true DOA marked by a red cross, $T_{60} = 0.5$ s and sensor noise level with SNR=20 dB.

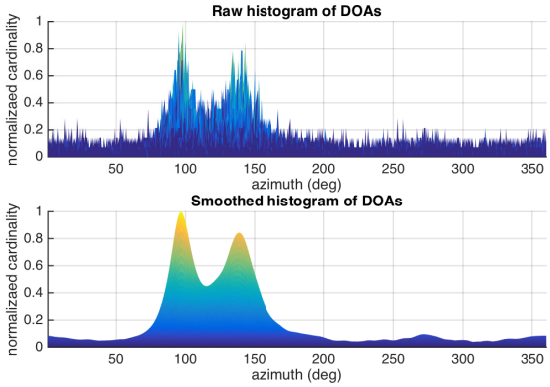

Fig. 5. An example side view of the raw and the smoothed 2D histograms of the estimated narrow-band DOAs with $\sigma = 4°$.

$l(m)$	$Y^*_{lm}(\theta,\varphi)$	$\nabla\{\mid Y^*_{lm}(\Omega)\mid^2\}$	$\nabla\left\{\mid Y^*_{lm}(\Omega)\mid\cos\left(\lambda_{lm}-\angle Y^*_{lm}(\Omega)\right)\right\}$
0(0)	$\sqrt{\frac{1}{4\pi}}$	0	0
1(-1)	$\sqrt{\frac{3}{8\pi}}\sin(\theta)e^{i\varphi}$	$(\frac{3}{8\pi})\sin(2\theta)\hat{\theta}$	$\sqrt{\frac{3}{8\pi}}\left\{\cos(\theta)\cos\left(\lambda_{1(-1)}-\varphi\right)\hat{\theta}+\sin(\theta)\sin\left(\lambda_{1(-1)}-\varphi\right)\hat{\varphi}\right\}$
1(0)	$\sqrt{\frac{3}{4\pi}}\cos(\theta)$	$(\frac{-3}{4\pi})\sin(2\theta)\hat{\theta}$	$-\frac{1}{2}\sqrt{\frac{3}{\pi}}\sin(\theta)\cos\left(\lambda_{10}\right)\hat{\theta}$
1(1)	$-\sqrt{\frac{3}{8\pi}}\sin(\theta)e^{-i\varphi}$	$(\frac{3}{8\pi})\sin(2\theta)\hat{\theta}$	$-\sqrt{\frac{3}{8\pi}}\left\{\cos(\theta)\cos\left(\lambda_{1(1)}+\varphi\right)\hat{\theta}-\sin(\theta)\sin\left(\lambda_{1(1)}+\varphi\right)\hat{\varphi}\right\}$
2(-2)	$\sqrt{\frac{15}{32\pi}}\sin^2(\theta)e^{2i\varphi}$	$(\frac{15}{32\pi})4\sin^3(\theta)\cos(\theta)\hat{\theta}$	$\sqrt{\frac{15}{32\pi}}\left\{\sin(2\theta)\cos\left(\lambda_{2(-2)}-2\varphi\right)\hat{\theta}+2\sin^2(\theta)\sin\left(\lambda_{2(-2)}-2\varphi\right)\hat{\varphi}\right\}$
2(-1)	$\sqrt{\frac{15}{8\pi}}\frac{1}{2}\sin(2\theta)e^{i\varphi}$	$(\frac{15}{8\pi})\sin(2\theta)\cos(2\theta)\hat{\theta}$	$\sqrt{\frac{15}{8\pi}}\left\{\cos(2\theta)\cos\left(\lambda_{2(-1)}-\varphi\right)\hat{\theta}+\frac{1}{2}\sin(2\theta)\sin\left(\lambda_{2(-1)}-\varphi\right)\hat{\varphi}\right\}$
2(0)	$\sqrt{\frac{5}{16\pi}}(3\cos^2(\theta)-1)$	$(\frac{-15}{8\pi})\sin(2\theta)\left(3\cos^2(\theta)-1\right)\hat{\theta}$	$-\sqrt{\frac{5}{16\pi}}3\sin(2\theta)\cos\left(\lambda_{20}\right)\hat{\theta}$
2(1)	$-\sqrt{\frac{15}{8\pi}}\frac{1}{2}\sin(2\theta)e^{-i\varphi}$	$(\frac{15}{8\pi})\sin(2\theta)\cos(2\theta)\hat{\theta}$	$-\sqrt{\frac{15}{8\pi}}\left\{\cos(2\theta)\cos\left(\lambda_{2(1)}+\varphi\right)\hat{\theta}-\frac{1}{2}\sin(2\theta)\sin\left(\lambda_{2(1)}+\varphi\right)\hat{\varphi}\right\}$
2(2)	$\sqrt{\frac{15}{32\pi}}\sin^2(\theta)e^{-2i\varphi}$	$(\frac{15}{32\pi})4\sin^3(\theta)\cos(\theta)\hat{\theta}$	$\sqrt{\frac{15}{32\pi}}\left\{\sin(2\theta)\cos\left(\lambda_{2(2)}+2\varphi\right)\hat{\theta}-2\sin^2(\theta)\sin\left(\lambda_{2(2)}+2\varphi\right)\hat{\varphi}\right\}$
3(-3)	$\sqrt{\frac{35}{64\pi}}\sin^3(\theta)e^{3i\varphi}$	$(\frac{35}{64\pi})3\sin(2\theta)\sin^4(\theta)\hat{\theta}$	$\sqrt{\frac{35}{64\pi}}\{3\sin^2(\theta)\cos(\theta)\cos\left(\lambda_{3(-3)}-3\varphi\right)\hat{\theta}$ $+3\sin^3(\theta)\sin\left(\lambda_{3(-3)}-3\varphi\right)\hat{\varphi}\}$
3(-2)	$\sqrt{\frac{105}{32\pi}}\sin^2(\theta)$ $\times\cos(\theta)e^{2i\varphi}$	$(\frac{105}{32\pi})\sin(2\theta)\sin^2(\theta)$ $\times(2-3\sin^2(\theta))\hat{\theta}$	$\sqrt{\frac{105}{32\pi}}\{\sin(\theta)(2-3\sin^2(\theta))\cos\left(\lambda_{3(-2)}-2\varphi\right)\hat{\theta}$ $+2\sin^2(\theta)\cos(\theta)\sin\left(\lambda_{3(-2)}-2\varphi\right)\hat{\varphi}\}$
3(-1)	$\sqrt{\frac{21}{64\pi}}\sin(\theta)$ $\times(5\cos^2(\theta)-1)e^{i\varphi}$	$(\frac{21}{64\pi})(4-5\sin^2(\theta))$ $\times(4-15\sin^2(\theta))\sin(2\theta)\hat{\theta}$	$\sqrt{\frac{21}{64\pi}}\{\cos(\theta)(4-15\sin^2(\theta))\cos\left(\lambda_{3(-1)}-\varphi\right)\hat{\theta}$ $+\sin(\theta)\left(5\cos^2(\theta)-1\right)\sin\left(\lambda_{3(-1)}-\varphi\right)\hat{\varphi}\}$
3(0)	$\sqrt{\frac{7}{16\pi}}(5\cos^3(\theta)$ $-3\cos(\theta))$	$(\frac{7}{16\pi})3(2-5\sin^2(\theta))$ $\times(-4+5\sin^2(\theta))\sin(2\theta)\hat{\theta}$	$\sqrt{\frac{7}{16\pi}}3\sin(\theta)(-4+5\sin^2(\theta))\cos\left(\lambda_{3(0)}\right)\hat{\theta}$
3(1)	$-\sqrt{\frac{21}{64\pi}}\sin(\theta)$ $\times(5\cos^2(\theta)-1)e^{-i\varphi}$	$(\frac{21}{64\pi})(4-5\sin^2(\theta))$ $\times(4-15\sin^2(\theta))\sin(2\theta)\hat{\theta}$	$-\sqrt{\frac{21}{64\pi}}\{\cos(\theta)(4-15\sin^2(\theta))\cos\left(\lambda_{3(1)}+\varphi\right)\hat{\theta}$ $-\sin(\theta)\left(5\cos^2(\theta)-1\right)\sin\left(\lambda_{3(1)}+\varphi\right)\hat{\varphi}\}$
3(2)	$\sqrt{\frac{105}{32\pi}}\sin^2(\theta)$ $\times\cos(\theta)e^{-2i\varphi}$	$(\frac{105}{32\pi})\sin(2\theta)\sin^2(\theta)$ $\times(2-3\sin^2(\theta))\hat{\theta}$	$\sqrt{\frac{105}{32\pi}}\{\sin(\theta)(2-3\sin^2(\theta))\cos\left(\lambda_{3(2)}+2\varphi\right)\hat{\theta}$ $-2\sin^2(\theta)\cos(\theta)\sin\left(\lambda_{3(2)}+2\varphi\right)\hat{\varphi}\}$
3(3)	$-\sqrt{\frac{35}{64\pi}}\sin^3(\theta)e^{-3i\varphi}$	$(\frac{35}{64\pi})3\sin(2\theta)\sin^4(\theta)\hat{\theta}$	$-\sqrt{\frac{35}{64\pi}}\{3\sin^2(\theta)\cos(\theta)\cos\left(\lambda_{3(3)}+3\varphi\right)\hat{\theta}$ $-3\sin^3(\theta)\sin\left(\lambda_{3(3)}+3\varphi\right)\hat{\varphi}\}$

TABLE I
GRADIENT OF THE COST FUNCTION FOR THE HARMONIC ORDERS AND DEGREES UP TO THE THIRD ORDER

minimum angular separation of the sources and the choice of $3 \leq \sigma \leq 6$ provides robust and well distinguished peaks for $\geq 30^\circ$ separation.

IV. EVALUATION

The proposed DOA estimation algorithms are evaluated in terms of their accuracy and robustness to Reverberation Time (RT) [44], sensor noise level, number of sources, and angular separation of sources using simulated data for one talker and for multiple simultaneous talkers. The Acoustic Impulse Responses (AIRs) of a 32-element rigid SMA with radius of $4.2\,\text{cm}$ (corresponding to the em32 Eigenmike®) in a $5 \times 6 \times 4\,\text{m}$ shoebox room were simulated using Spherical Microphone arrays Impulse Response Generator (SMIRgen) [45] based on Allen & Berkley's image method [46]. For all methods, a sampling frequency of $8\,\text{kHz}$ was used with a Short-Time Fourier Transform (STFT) window size of $8\,\text{ms}$ and 50% overlap of time frame. The processing band was set to $500\,\text{Hz}$ to $3850\,\text{Hz}$ to avoid spatial aliasing and ensuring $kr < L$ for $L = 3$ as in [47], [6] and to avoid excessive noise amplification due to mode strength compensation at lower frequencies.

The proposed methods, denoted AIV Grid Search (AIV-GS) and AIV Gradient Descent (AIV-GD), are compared to the previously presented PIV, SSPIV, PWD-SRP as the baseline methods and both variations of DPD-MUSIC as the state-of-the-art. For AIVs we evaluate the effect of the choice of maximum SH order $L = 2, 3$. Accordingly, the evaluated algorithms are denoted AIV-GS2, AIV-GS3, AIV-GD2 and AIV-GD3. For the rest of the evaluation $L = 3$ is considered for the methods using high order SHs.

In order to compare the narrow-band PWD-SRP with our AIV-GS under same spatial limitation, we employ Spatially Constrained (SC)-SRP which uses the same search window as AIV-GS and PIV as its centre of window. The employed SC-SRP is the generalized variation of the proposed method in [48] in which SC-SRP is applied only on the TF bins with an active single source unlike our SC-SRP which is applied on all TF bins. The radius of spherical cap search window in AIV-GS and SC-SRP was set to 10° as, in our experiments, more than 95% of PIVs were within 10° of the true DOA. For AIV-GD,

the optimization function 'fminunc' based on 'trust-region' algorithm from MATLAB Optimization Toolbox™ was used and was set to be terminated if the new angle is less than $0.5°$ away from the current angle or if the number of calls to the cost function exceeds 100. These termination conditions were determined empirically. The covariance matrix $\mathbf{R}$ in (16) used by SSPIV and DPD-MUSIC is approximated as the average covariance matrix over a local TF neighbourhood [6], [37]

$$\mathbf{R}(\tau,k) = \frac{1}{J_\tau J_k}\sum_{j_\tau=0}^{J_\tau-1}\sum_{j_k=0}^{J_k-1}\mathbf{a_{lm}}(\tau+j_\tau,k+j_k) \times \mathbf{a_{lm}}^H(\tau+j_\tau,k+j_k), \quad (46)$$

where $J_\tau = 6$ and $J_k = 4$ are the width (number of bins) of averaging window over time and frequency respectively giving 500Hz and 32ms of window size in the TF domain based on our frequency and time resolution. The threshold ϵ in (20) for DPD-MUSIC was empirically set to 6, which is also the choice in its original paper [6]. The DOA histogram and MUSIC spectrum were constructed with 1 degree resolution along inclination and azimuth (181×360 points respectively). In DOA histogram smoothing, the kernel had the standard deviation of $\sigma = 4°$ which was chosen empirically from a range of $2°$ to $6°$.

A. Single Source

The array is placed at $(2.52,\ 3.11,\ 1.97)$ m and the source signal consists of an anechoic speech signal, using the same utterance for all trials [49] with duration 5 s, convolved with the simulated AIRs to each microphone and white Gaussian sensor noise added. We consider 40 different source positions at the distance of 1 m from the centre of array with a DOA randomly selected from a uniform distribution around the sphere. For each source position, the test was repeated over a range of RT, $T_{60} = \{0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6\}$ s, and signal-to-noise ratio, SNR=$\{10,\ 20,\ 30\}$ dB.

For each method, the DOA estimation error ε between the true DOA unit vector $\mathbf{u}_o$ and the estimated DOA unit vector $\mathbf{u}_e$ was computed in degrees as

$$\varepsilon = \cos^{-1}\left(\mathbf{u}_o^T\mathbf{u}_e\right). \quad (47)$$

Results are presented in two parts. In the first we compare the second- and third-order of two variations of our methods, grid search and gradient descent AIVs with PIV method. In the second, we compare the most accurate AIV approach with the baseline and the state-of-the-art methods.

1) Quantification of the improvements due to higher order spherical harmonics: Figure 6 shows the mean DOA estimation error for each method as a function of T_{60} for all SNRs. As expected due to utilisation of higher spatial resolution from higher SHs, the AIV approaches significantly outperform PIV for all T_{60} and SNRs. AIV approaches also show noticeably more robustness to reverberation and noise. Comparing AIV-GS and AIV-GD, advantage of gradient descent becomes noticeable as the noise increases. This is due to spatial freedom of search for gradient descent since the AIV-GS's search window centred on PIVs, which are prone to noise, are more likely to not include the global minimum of cost function. Moreover, the results demonstrate that the increase in SH order (2 vs 3) has a larger impact on improvement of accuracy and robustness than the change in optimization method (GS vs GD) highlighting the higher importance of the cost function quality over the optimization method used for it.

2) Comparison with baselines and state-of-the-art: In this section our AIV-GD and AIV-GS are compared with SSPIV, SC-SRP and incoherent DPD-MUSIC.

In terms of accuracy, AIV-GD shows the second best accuracy of $\leq 2°$ after DPD-MUSIC. The worst performance is by PIV as it only uses the low order eigenbeams. Although SSPIV uses the same formulation as PIV, it performs significantly better than PIV as it employs high order SH in SVD to estimate de-noised low order eigenbeams. SC-SRP and AIV-GS outperform the previous two methods due to utilisation of high order eigenbeams but perform very similar to each other as they use the same eigenbeams, search window and TF bins although their formulations differ. AIV-GD outperforms all previously stated methods due to high order eigenbeams, compared to PIV and SSPIV, and lack of spatial limitation compared to SC-SRP and AIV-GS. DPD-MUSIC leads in the performance wih $0.5°$ accuracy due to utilisation of denoised high order eigenbeams using SVD, spatial freedom due to full

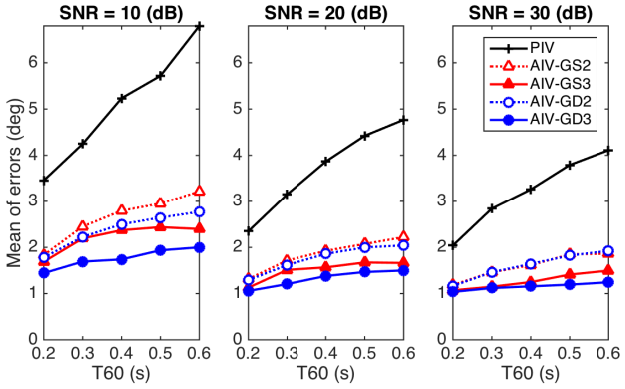

Fig. 6. Effect of T_{60} and SNR on mean DOA estimation error for PIV and AIV methods in the single source scenario.

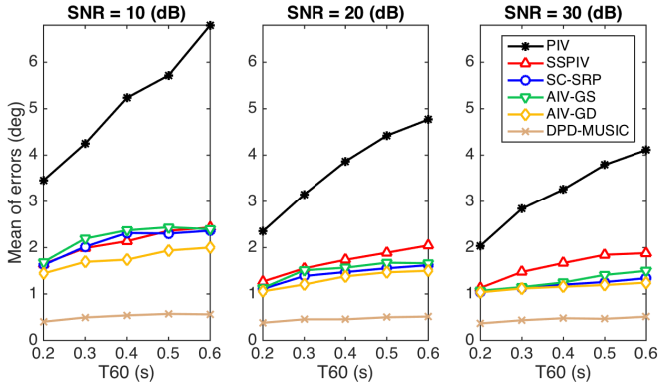

Fig. 7. Effect of T_{60} and SNR on mean DOA estimation error for the proposed, baseline and state-of-the-art methods in the single source scenario.

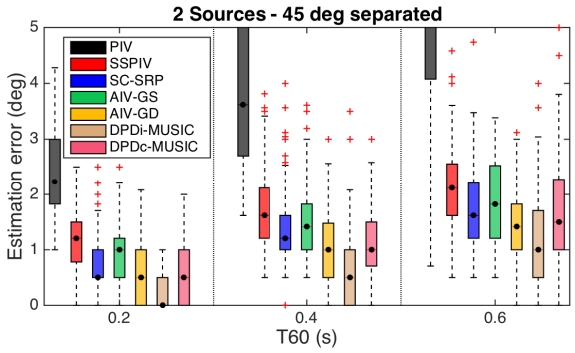

Fig. 8. Distribution of DOA estimation errors for two sources 45° apart with varying T_{60}.

grid search and DPD test which estimates reliable TF bins in which accuracy of DOA estimation would be high.

In terms of robustness to noise and reverberation, subspace techniques such as SSPIV and DPD-MUSIC lead as they take advantage of decomposition of noisy eigenbeams into signal and noise subspaces. Although AIV-GD uses noisy eigenbeams, it shows almost as strong robustness as subspace techniques due to its spatially-unconstrained optimization which minimizes the effect of noise on the optimized DOA estimate.

B. Multiple Sources

In this section we evaluate the effect of reverberation, number of sources and angular separation of the sources in the multiple source scenario.

In order to systematically evaluate the effect of source separation with varying number of sources we used multi-source distribution with similar angular separation between them. We chose the distribution of the sources on the same horizontal plane as the microphone array for simplicity of understanding of the result and maximizing the clarity of systematic evaluation of the effect of source separation. The experiments in section IV-A and V demonstrate the effectiveness of the method in varying azimuth-inclination condition. In total, 100 trials were used where in each trial the azimuth of the first source is chosen randomly from a uniform distribution around the circle and the subsequent sources are placed at regularly spaced azimuth intervals $\Delta\phi_s$. The number of sources N_s and the angular separation $\Delta\phi_s$ vary in each experiment, as described below.

The constraint of fixed inclination reduces the range of variations in distance-to-the-closest-wall and so the strongest reflection in compare to the single source scenario in which both the azimuth and inclination vary per trial. In order to compensate the reduction in range of variation of the strongest reflection, we displaced the microphone array slightly away from the centre of the room at $(2.52,\ 4.48,\ 1.45)$ m in multi-source scenario while the distance of the sources to the centre of SMA stays 1 m.

The source signals consist of different anechoic speech signals randomly selected for each trial from the APLAWD database [50]. The active level of each speech source according to ITU-T P.56 [51], as measured at p_{00}, is set to be equal across all trials. Spatio-temporally white Gaussian noise is added to the microphone signals to produce a signal to incoherent noise ratio (SNR) of 25 dB at p_{00} for each source. In order to analyse the effect of reverberation, number of sources and the angular separation of the sources on multiple source localization, the evaluation includes two experiments to investigate: 1) the effect of T_{60} and 2) the effect of N_s and $\Delta\phi_s$.

For multiple sources and an equal number of estimated DOAs, the average DOA estimation error depends on how we associate the true DOAs and the estimated DOAs in (47). In order to avoid any ambiguity due to data association uncertainty in our results, best case data association was used to obtain the mean estimation error using (47).

1) Experiment 1: The effect of T_{60} is evaluated here for the illustrative case with $N_s = 2$ and $\Delta\phi_s = 45°$. Figure 8 shows the distribution of DOA estimation errors of all methods for $T_{60} = \{0.2,\ 0.4\,,0.6\}$ s. The black dot in each box shows the median while the boxes show the upper and lower quartiles, and the whiskers which is extend to 1.5 times the interquartile range.

As studied in [37], the accuracy of PIV is severely degraded in strong reverberation when multiple sources are simultaneously active. This can be seen in Figure 8 where PIV median error increases from 2.2° to up to 8.7° (out of Y-axis limit) as T_{60} increases. AIV-GD, after incoherent DPD-MUSIC, shows the second best accuracy with median errors of $\{0.5,\ 0.9,\ 1.4\}$ ° and robustness to reverberation of 1° similar to coherent DPD-MUSIC for all T_{60}s. Although incoherent DPD-MUSIC leads in accuracy it shows the same robustness to reverberation as others as its median error varies for around 1° from lowest to highest T_{60}. The coherent DPD-MUSIC provides less accuracy than incoherent DPD-MUSIC since

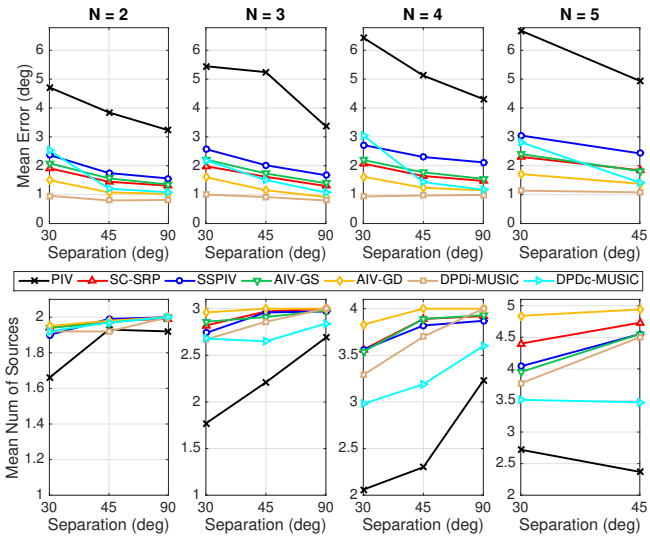

Fig. 9. Mean error and Mean NoDS for incremental N_s and varying $\Delta\phi_s$ with $T_{60} = 0.4\ s$.

DPDc-MUSIC is more prone to TF bins with erogenous signal space due to high sensitivity of clustering to outliers.

2) Experiment 2: The effect of the number of sources and the angular separation of the sources is evaluated for incremental N_s from 2 to 5 sources with $\Delta\phi_s = \{30,\ 45\,, 90\}^\circ$ for up to 4 sources and $\Delta\phi_s = \{30,\ 45\}^\circ$ for 5 sources with $T_{60} = 0.4\,\text{s}$. The performance of each method is evaluated using two metrics: Mean Number of Detected Sources (NoDS) and Mean error respectively representing the robustness and accuracy of DOA estimators.

Mean NoDS: Having obtained N DOA estimates ($N \leq N_s$), we use best case data association to find the best assignment of N estimated DOAs and N_s true DOAs. In the best case assignment, the error between each pair of estimated and true DOAs are calculated using (47). The number of pairs with error $\leq 15^\circ$ is considered as the number of detected sources. The mean number of detected sources is the average number of detected sources across all trials.

Mean error: The mean error is simply the average estimation error among the trials with successful localization where the number of detected sources is equal to the number of true sources. Figure 9 shows the mean errors and mean NoDS for all methods with incremental N_s and varying $\Delta\phi_s$. In terms of accuracy, AIV-GD shows the second best performance with the worst mean error of 1.8° after DPDi-MUSIC with the worst mean error of 1°. AIV-GS performs as accurate as SC-SRP with the worst mean error of 2.1° as they both utilise the same eigenbeams, initial DOA estimates and search window although differ in cost function. DPDc-MUSIC shows noticeable accuracy loss of 2° for adjacent sources with $\Delta\phi_s = 30^\circ$ for all the values of N_s. In contrast to the results in the original work on DPD-MUSIC by Nadiri and Rafaely where the sources are widely separated by 60° and 70°, in scenarios with lower separation, as in our evaluation, DPDc-MUSIC, compared to DPDi-MUSIC, does not show a better accuracy. This is caused as the clustering in DPDc-MUSIC becomes highly prone to adjacent sources and results in the merge of two adjacent clusters of signal subspaces giving erroneous centroid signal subspace especially in the presence of outliers signal subspaces. In terms of mean NoDS, AIV-GD has the highest robustness to angular separation and number of sources. Apart from PIV, which generally performs poorly in multi-source scenario in all cases, SSPIV, SC-SRP, AIV-GS and AIV-GD show more robustness to number of sources than DPD-based methods. Both DPD-MUSIC variations significantly lose mean NoDS as the number of sources increases due to static threshold for SVR in DPD test which causes the reduction of number of TF bins that passed the test with the increase of number of sources. Figure 10 presents the percentage of the passed TF bins in DPD test for incremental number of sources. As expected the percentage reduces with the increase of N_s as the likelihood of dominant single source in a bin drops. It can also be observed that the percentage increases as the angular separation of the sources decreases since the likelihood of strong unity rank in (20) increases as the two adjacent sources may be considered as a single erroneous intermediate dominant source. Although increase in source separation decreases the percentage of the passed bins, the accuracy and robustness increase as shown in Fig. 9 due to having less erroneous passed TF bins.

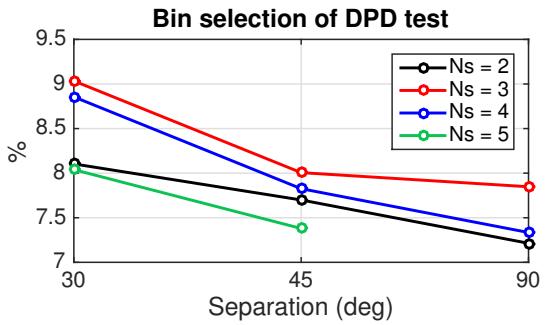

Fig. 10. Percentage of the bins passed in DPD test for varying N_s and $\Delta\phi_s$.

V. Experimental Verification

To demonstrate the performance of each method, real recording of 4 s speech in a real room with approximate dimensions of $10 \times 9 \times 2.5\,\text{m}$ and reverberation time of $0.4\,\text{s}$ was used using an Eigenmike 32-channel rigid spherical microphone array with radius of $4.2\,\text{cm}$ placed close to the centre of the room. Four talkers were simultaneously active and were located $1.5\,\text{m}$ away from the centre of the array at approximately 60° intervals while their inclinations alternated to be above or below the horizontal plane of the array. Figure 11 shows the normalized smoothed histogram for PIV, SSPIV, SC-SRP, AIV-GS and AIV-GD as well as normalized MUSIC spectrum for incoherent DPD-MUSIC. Due to approximate knowledge of the position of sources and array, we cannot obtain accurate numerical estimation error. The approximate mean estimation error for all methods is 3° except PIV which is 4.5°. Although all methods successfully estimate peaks corresponding to all four sources due to well separation, the distinctness, the sharpness and the height of the peaks clearly present the relative performance of the methods. In order to provide a numerical evaluation, for each peak, a measure of 'peak strength' is proposed which is the ratio of the peak height over the peak smoothness where the peak smoothness is defined as the average height in the normalized peak distribution within its range of $r_p = 25^\circ$ neighbourhood. Table II presents the peak strength of each peak for all methods.

AIV-GD and SSPIV lead as they both estimate the most prominent peaks. AIG-GS and SC-SRP performs similarly as explained in previous sections. SSPIV, due to noise suppression in eigenbeams by sub-space decomposition, manages to

Peak	PIV	SC-SRP	DPD-M	SSPIV	AIV-GS	AIV-GD
1	2.08	5.66	3.31	6.23	5.24	6.88
2	1.96	4.45	3.04	6.12	4.18	5.37
3	1.82	3.54	2.39	5.32	3.50	3.35
4	0.99	1.50	0.49	2.31	1.46	1.17
Mean	1.71	3.79	2.31	5.00	3.59	4.19

TABLE II
Peak Strength of each peak for all methods

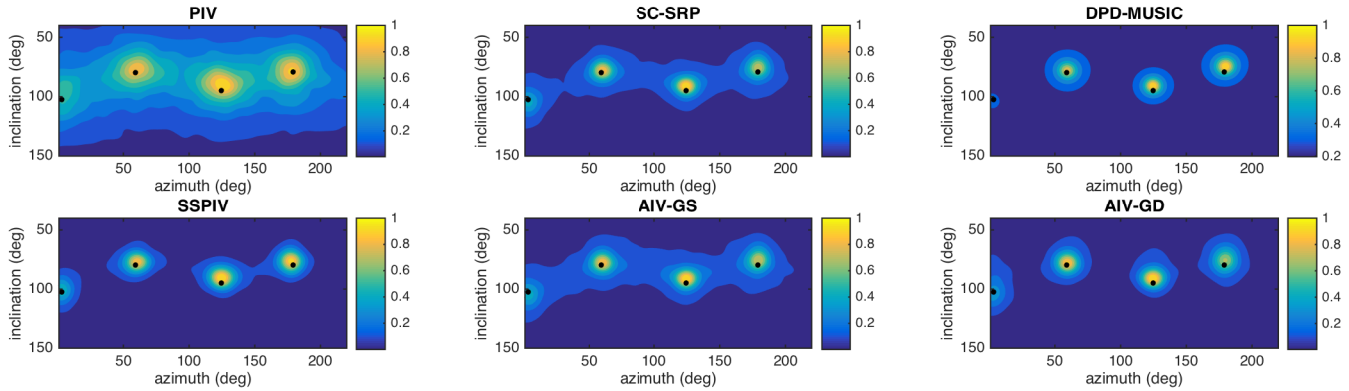

Fig. 11. Normalized smoothed histogram of PIV, SSPIV, SC-SRP, AIV-GS, AIV-GD and normalized incoherent DPD-MUISC spectrum using real recording. The black dot represents the approximate true DOA.

successfully estimate accurate DOAs in the majority of TF bins where PIV estimates an erroneous DOA. The performance improvement of AIV-GD compared to AIV-GS, due to utilisation of spatially unconstrained optimization, is clearly observable in Fig. 11 and in Table II as erroneous DOAs in AIV-GS, which are mainly distributed between and around the peaks, are more concentrated around the peaks in AIV-GD resulting into sharper peaks. Note that the number of DOA estimates in all methods except DPD-MUSIC are equal. Comparing the sharpness and the sparsity of the histograms of PIV and AIV-GD, we can also see the significant accuracy improvement for AIV-GD since majority of the erroneous DOAs in PIV histogram have had their accuracy improved in AIV-GD due to employment of high order eigenbeams. DPD-MUSIC shows a poor peak strength in Table II due to having a very low-height, although sharp, peak (peak 4) as seen in Fig. 11. This is caused since an increase in the number of sources results in reduction of the number of passed bins in DPD test, as previously shown in Fig. 10, which can result in having low-height peaks in the MUSIC spectrum and therefore potentially missing a source.

VI. Computational Complexity

In this section we discuss the number of computations required in each method for a single TF bin in terms of the number of real multiplications. Note that the multiplication of two complex numbers is counted as four real multiplication while $|\ .\ |^2$ is counted as two real multiplications. We do not include the number of multiplications in (2) as we pre-calculate and store all the required $Y_{lm}\left(\Omega\right)$.

For subspace methods, we have $4(L+1)^4 J_\tau J_k$ multiplications for covariance matrix in (46) as well as $3(L+1)^6$ for SVD in (17).

For PIV, we have 48 $(3\times 3\times 4+3\times 4)$ operations where the numbers in parentheses respectively represent the number of axes, harmonic modes and the real multiplications in (14) and the number of axes and the real multiplications in (13).

For AIV and SRP cost functions using (27) and (10), we have $48+(L+1)^2\times(2+4+2)$ operations for a single look direction where the numbers respectively correspond to the PIV, number of eigenbeams up to the order L, a real-complex followed by a complex-complex multiplications, and squared magnitude.

For DPD-MUSIC, as well as subspace computation we have $4((L+1)^2-1)(L+1)^2$ multiplications for MUSIC spectrum in (21) for a single look direction. Coherent DPD-MUSIC is excluded from our consideration due to unknown and highly dependant complexity in clustering.

We empirically achieved an average of 5 iterations for gradient descent in AIV-GD. Considering numerical gradient using the four neighbouring look directions, we call AIV cost function 5 times per iteration which results in average of 25 look directions for AIV-GD. With spherical cap window of radius 10° for AIV-GS and SC-SRP, we have an average of 100 look directions. MUSIC uses a full grid search of 181×360 look directions.

Using the setting in this paper, the overall approximate number of real multiplications of each method per TF bin is as follow: 48 for PIV, 37 thousands for SSPIV, 13 thousands for SC-SRP and AIV-GS, 3 thousands for AIV-GD, and 250 millions for DPD-MUSIC assuming an average of 10% of the bins pass the DPD test. Our proposed AIV-GD leads in computation after PIV while the state-of-the-art DPD-MUSIC shows an expensive computational cost due to covariance matrix calculation, SVD and full grid search although it is performed on few percent of the total TF bins.

VII. Conclusions

In this paper we proposed a novel DOA estimation method for spherical microphone arrays. This method exploits a new measure denoted the augmented intensity vectors. It uses high order spherical harmonics to enhances the accuracy and robustness of DOA estimates in PIV. Two alternative implementations of our method were evaluated, one based on grid search and the other on gradient descent optimization. It is shown that the gradient descent approach shows a better performance in accuracy and robustness compared to spatially limited grid search approach. Simulation and real recording results have been presented for single and multiple sources with different sensor noise levels, reverberation times, number

of sources, and angular separation of sources. The results also show that using up to the third order spherical harmonics has significant advantages over second order harmonics for AIV and the increase of order has more impact on accuracy than the choice of optimization technique. For the third-order gradient descent AIV in the presence of realistic reverberation and sensor noise level, we found the worst average error of $1.5°$ for single source and $2°$ for up to 5 sources with down to $30°$ angular separation. It also outperforms the baseline PIV, SSPIV and SC-SRP and is at least as accurate as the state-of-the-art method DPD-MUSIC. It is shown that AIV-GD leads in terms of robustness to number of sources and separation. In addition, a rough analysis of computational complexity indicates that our proposed AIV-GD technique outperforms the state-of-the-art method in terms of computational complexity with a few thousands real multiplications per bin.

REFERENCES

[1] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 727–732, 2007.

[2] G. W. Elko and J. Meyer, "Spherical microphone arrays for 3D sound recordings," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., 2004, ch. 3, pp. 67–89.

[3] E. Fisher and B. Rafaely, "The nearfield spherical microphone array," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2008, pp. 5272–5275.

[4] ——, "Near-field spherical microphone array processing with radial filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 256–265, 2011.

[5] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, 2007.

[6] 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.

[7] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[8] ——, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[9] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007.

[10] B. Rafaely, "The spherical-shell microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 740–747, May 2008.

[11] 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.

[12] B. Rafaely, *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer, 2015.

[13] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.

[14] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[15] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 2–16, 2010.

[16] 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.

[17] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[18] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Philadelphia, PA, USA, Mar. 2005, pp. iii/89–iii/92.

[19] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Proc. Audio Eng. Soc. (AES) Convention*, New York, NY, USA, Oct. 2003, pp. 1–9.

[20] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Orlando, FL, USA, May 2002, pp. 1949–1952.

[21] 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.

[22] B. Rafaely, "Plane-wave decomposition of the pressure on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004.

[23] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, September 2016.

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

[25] 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.

[26] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[27] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[28] C. Chen, R. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[29] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1539–1551, October 2015.

[30] J. L. Yuxiang Hu and X. Qiu, "A maximum likelihood direction of arrival estimation method for open-sphere microphone arrays in the spherical harmonic domain," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 791–794, 2015.

[31] N. Epain and C. Jin, "Independent component analysis using spherical microphone arrays," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 91–102, 2012.

[32] T. Noohi, N. Epain, and C. Jin, "Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 346–349.

[33] S. Tervo, "Direction estimation based on sound intensity vectors," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Scotland, August 2009, pp. 700–704.

[34] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[35] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[36] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, September 2015.

[37] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudo-intensity vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[38] S. Hafezi, A. H. Moore, and P. A. Naylor, "3D acoustic source localization in the spherical harmonic domain based on optimized grid

search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Shanghai, China, March 2016.
[39] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999.
[40] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer, 2016.
[41] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, September 2016.
[42] M. J. Crocker and F. Jacobsen, "Sound intensity," in *Handbook of Acoustics*, M. J. Crocker, Ed. Wiley-Interscience, 1998, ch. 106, pp. 1327–1340.
[43] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.
[44] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.
[45] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132.
[46] 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.
[47] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2009, pp. 221–224.
[48] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D DOA estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 96–100.
[49] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, Corpus LDC93S1, 1993.
[50] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Technical Report, Jun. 1987.
[51] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Dec. 2011. [Online]. Available: http://www.itu.int/rec/T-REC-P.56-201112-I/en

Sina Hafezi received the BEng degree in Electronic Engineering in 2012 and the MSc degree in Digital Signal Processing in 2013 both from Queen Mary, University of London, UK. He worked in the Centre for Digital Music as a researcher and software engineer on multiple projects related to autonomous equalization, which led to patent and commercial application. He then started the PhD in Acoustic Signal Processing at Imperial College London, UK in 2014. His research interests are in the field of audio signal processing especially spherical microphone arrays, source localization, modeling of room acoustics and spatial audio.

Alastair H. Moore (M'13) received the MEng degree in Electronic Engineering with Music Technology Systems in 2005 and the PhD degree in 2010, both from the University of York, UK. He spent 3 years as a hardware design engineer for Imagination Technologies plc designing digital radio and networked audio consumer electronics products. In 2012, he joined the Department of Electrical and Electronic Engineering at Imperial College London as a postdoctoral research associate. His research interests are in the field of speech and audio processing, especially microphone array signal processing, acoustic scene awareness, dereverberation and spatial audio perception with applications for robot audition and hearing aids.

Patrick A. Naylor (M'89, SM'07) is a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. He received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, UK, and the PhD. degree from Imperial College London, UK. His research interests are in the areas of speech, audio and acoustic signal processing. He has worked in particular on adaptive signal processing for speech dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single and multi-channel speech enhancement and speech production modelling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the UK, USA and in Europe. He is the past-Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and a director of the European Association for Signal Processing (EURASIP). He has served as an associate editor of IEEE Signal Processing Letters and is currently a senior area editor of IEEE Transactions on Audio Speech and Language Processing.