Direction of Arrival Estimation in the Spherical Harmonic Domain using Subspace Pseudo-Intensity Vectors

Alastair H. Moore, *Member, IEEE,* Christine Evers, *Senior Member, IEEE,* and Patrick A. Naylor, *Senior Member, IEEE*

***Abstract*—Direction of Arrival (DOA) estimation is a fundamental problem in acoustic signal processing. It is used in a diverse range of applications, including spatial filtering, speech dereverberation, source separation and diarization. Intensity vector-based DOA estimation is attractive, especially for spherical sensor arrays, because it is computationally efficient. Two such methods are presented which operate on a spherical harmonic decomposition of a sound field observed using a spherical microphone array. The first uses Pseudo-Intensity Vectors (PIVs) and works well in acoustic environments where only one sound source is active at any time. The second uses Subspace Pseudo-Intensity Vectors (SSPIVs) and is targeted at environments where multiple simultaneous sources and significant levels of reverberation make the problem more challenging. Analytical models are used to quantify the effects of an interfering source, diffuse noise and sensor noise on PIVs and SSPIVs. The accuracy of DOA estimation using PIVs and SSPIVs is compared against the state-of-the-art in simulations including realistic reverberation and noise for single and multiple, stationary and moving sources. Finally, robust performance of the proposed methods is demonstrated using speech recordings in real acoustic environments.**

***Index Terms*—Direction of arrival estimation, DOA localization, speaker tracking, robot audition, microphone array processing, spherical microphone array, spherical harmonics**

I. Introduction

MANY applications of acoustic signal processing rely on Direction of Arrival (DOA) estimation, including spatial filtering, speech dereverberation, source separation and diarization. Estimation of the DOA of a sound source is particularly important in the context of robot audition where tracking the directions of one or more moving sources enables an ‘awareness’ of the local environment, which is a requirement for effective human-robot interaction.

To estimate both the vertical and horizontal angles of arrival requires a three-dimensional microphone array. Array geometries which sample the sound field such that it can be represented in the Spherical Harmonic (SH) domain are attractive because this representation allows the sound field to be analyzed with equal resolution in all directions using algorithms which are independent of the specific array geometry [1]–[4].

A wide variety of DOA estimation algorithms have been proposed for use in the SH domain [5]–[14]. Most of these compute a metric over a dense azimuth-inclination grid before identifying its peak(s) as the DOA(s). Such methods include those that compute the Steered Response Power (SRP) due to a beamformer which is steered towards all potential source directions and those that compute the spatial spectrum using subspace methods based on Multiple Signal Classification (MUSIC) [15].

Many current DOA estimation methods make use of the spatial covariance matrix [7], [9], [10], [12]. For example, the SRP map produced by a Minimum Variance Distortionless Response (MVDR) beamformer optimally rejects background noise for each look direction by adjusting its beam pattern according to the spatial covariance matrix and MUSIC [15] directly decomposes the spatial covariance matrix into signal and noise subspaces. However, in reverberation, coherent reflections distort the spatial covariance matrix. For the MVDR beamformer this is manifested as incorrectly placed attenuation in the beam pattern. For MUSIC the fact that the reflections are linearly dependent on the direct path signals means the rank of the covariance matrix is reduced and division between signal and noise subspaces can be prone to errors.

Frequency Smoothing (FS) [16] has been shown to improve the accuracy of DOA estimation using MUSIC [7] and MVDR-SRP [10]. The procedure decorrelates coherent reflections by combining information across multiple frequency bands. In the spatial domain, where microphone signals are processed directly, special focussing matrices and an initial DOA estimate are required. In the SH domain, FS can be applied as a straightforward average by assuming frequency independence of the (mode strength compensated) array manifold [7] [10].

To estimate multiple source DOAs, a number of authors have proposed methods which exploit the sparsity of speech in the Time-Freqeuncy (TF) domain. By identifying TF-regions where a single source is dominant, single source DOA estimation methods can be employed locally to those regions [12], [17], [18]. This class of methods exploits the principle that, for a single dominant source, the rank of the spatial covariance matrix is unity. In [18] pairwise correlations between adjacent

A. H. Moore, C. Evers and P. A. Naylor are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: alastair.h.moore@imperial.ac.uk; c.evers@imperial.ac.uk; 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.

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M026698/1].

microphones of a circular array were estimated by averaging over frequency bins within a single time frame. In [17] the spatial covariance matrix between all microphones was estimated at each frequency bin by averaging over time frames. In [12] it was shown that estimating the spatial covariance matrix by averaging (smoothing) over time and frequency decorrelates the reflections and so the rank is only unity when a single direct path is dominant. Accuracy of the subsequent DOA estimation is substantially improved but the Direct-Path Dominance (DPD) test described in [12] is reported to be passed in only 3% of TF-regions. This may lead to time frames in which there are no DOA estimates, which is problematic in applications where the sources are moving.

Methods for DOA estimation in the SH domain which exploit the directional sparsity of sound sources have been proposed in a series of related works [19]–[22]. In [19] Independent Component Analysis (ICA) of the SH domain signals was performed and the DOAs estimated by comparing the columns of the unmixing matrix to the steering vectors for plane waves from all possible directions. In [20] the directional component of the SH domain signals was obtained by subtracting an estimate of the diffuse component, which were determined using a subspace approach. An iterative optimization was then performed to find a sparse set of weights for a dense dictionary of plane wave elements. The directions associated with the selected elements represent the estimated DOAs. In [21] and [22] various approaches to combining the methods of [19] and [20] were proposed, each with their own success in a particular application scenario. However, none of these included live recordings of real-world audio, where small source movements may be important.

Intensity-based DOA estimation [8], [11], [13], [14] differs from the previously discussed methods because, by directly computing the direction of energy flow, there is no need to compute a spatial cost function. This has the potential for significant computational savings. The component of intensity in a particular direction has been measured using two types of intensity probe [23]. One approximates particle velocity using the difference between two closely spaced omnidirectional pressure sensors while the other measures particle velocity directly [24]. The former approach is more common but is sensitive to phase mismatch and sensor noise. Using an array of intensity probes yields an intensity vector in 2 or 3 dimensions, from which the DOA can be found [25], [26].

In [8] DOA estimation using a spherical microphone array was proposed whereby a large number of microphones was used to transform the sound field into the SH domain from which the particle-velocity was approximated. The resulting vectors were termed Pseudo-Intensity Vectors (PIVs). Those initial results demonstrated the effectiveness of the method for single source DOA estimation in a noise-free environment. In [11] DOA estimation of multiple sources was achieved using k-means clustering of PIVs. In both [27] and [28] a DOA was obtained for each TF-bin by first finding the PIV and then refining the direction by evaluating a cost function using higher order SHs over a spatially constrained grid around the PIV direction. The final estimates of multiple sources' DOAs were then obtained by identifying the peaks of a histogram of all the individual direction estimates.

In this current paper we review the formulation and use of PIVs presented in [8]. We then provide a novel formulation of the PIV which follows directly from the SH domain representation of a sound field expressed in Cartesian form and develop an extended analysis of PIVs under non-ideal conditions. Further, we propose the Subspace Pseudo-Intensity Vector (SSPIV) which we show to be more robust to noise and reverberation than the PIV. Like DPD-MUSIC, it exploits FS and subspace decomposition and assumes TF-sparsity of the input signal. However, by directly computing a DOA for each TF-region, rather than evaluating the spatial spectrum over all possible directions, it is computationally more efficient. We investigate the criteria under which smoothed histograms of PIVs and SSPIVs give accurate estimates of the DOAs of multiple sources in a noisy reverberant environment, including when sources are moving. Some of the first steps of an earlier version of the SSPIV method were presented in [13] and [29]. The current paper extends both the theoretical analysis and the evaluation of the PIV method compared to [8], especially in the context of multiple and moving speakers and in real-world applications.

The remainder of this paper is organized as follows. Sec. II reviews the SH domain representation of a sound field. Sec. III presents the PIV and SSPIV methods. Sec. IV analyses PIV and SSPIV under non-ideal conditions, whether these be caused by an interfering (independent or correlated) sound source, diffuse noise or sensor noise. Sec. V presents simulated experiments comparing the intensity-based methods to classical and state-of-the-art DOA estimation methods. Sec. VI demonstrates the effectiveness of the methods in real-world tests. Finally the paper is concluded in Sec. VII.

II. Review of SH representation of a sound field

The SH representation of a sound field [4], [30] around a particular point in space is determined by the complex-valued plane-wave density $a(k,\theta,\phi)$, which is a function of wavenumber k, inclination θ and azimuth ϕ. A unit vector pointing towards the n-th plane wave, $\mathbf{x}_n = \begin{bmatrix} x_n & y_n & z_n \end{bmatrix}^T$, where $(\cdot)^T$ is the transpose operator, has DOA, $\Psi_n = (\theta_n, \phi_n)$, given by

$$\theta_n = \arccos(z_n), \qquad \phi_n = \operatorname{arctan2}(y_n/x_n) \tag{1}$$

where arctan2 is the arctangent function mapped to the correct quadrant according to the signs of x_n and y_n. A plane-wave density composed of N plane waves is given by

$$a(k,\theta,\phi) = \sum_{n=1}^{N} \delta\left(\cos\theta - \cos\theta_n\right)\delta\left(\phi - \phi_n\right)s_n\left(k\right) \tag{2}$$

where $s_n\left(k\right)$ is the amplitude of the n-th plane wave and $\delta\left(\cos\theta\right)\delta\left(\phi\right)$ is the Dirac delta function on the sphere, which is zero everywhere on the sphere except $(\theta,\phi) = (\pi/2, 0)$. The complex SHs of order l and degree $m \in \{-l,\ldots,l\}$ provide a set of orthogonal basis functions defined over the unit sphere [30]

$$Y_l^m(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m\left(\cos\theta\right)e^{im\phi} \tag{3}$$

where $P_l^m(\cdot)$ is the associated Legendre function such that

$$a(k,\theta,\phi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}a_{lm}(k)Y_l^m(\theta,\phi). \tag{4}$$

Substituting $\Omega=(\theta,\phi)$, the weights of each SH are the Spherical Fourier Transform (SFT) of $a(k,\theta,\phi)$

$$a_{lm}(k)=\int_{\Omega\in S^2}a(k,\Omega)\left[Y_l^m(\Omega)\right]^*d\Omega \tag{5}$$

where $\int_{\Omega\in S^2}d\Omega=\int_0^{2\pi}\int_0^{\pi}\sin\theta d\theta d\phi$ is the integral over the unit sphere and $(\cdot)^*$ denotes conjugation. Substituting (2) into (5) gives

$$a_{lm}(k)=\sum_{n=1}^{N}\left[Y_l^m(\Psi_n)\right]^*s_n(k). \tag{6}$$

Considering the $(L+1)^2$ SHs up to $l\leq L$, (6) is expressed in stacked vector notation as [12]

$$\mathbf{a}_{lm}(k)=\mathbf{Y}(\boldsymbol{\Psi})^H\mathbf{s}(k) \tag{7}$$

where subscript lm on a vector denotes that the elements are SH coefficients, $\mathbf{s}(k)=\left[s_1(k)\ldots s_N(k)\right]^T$, $\boldsymbol{\Psi}=\left[\Psi_1\ldots\Psi_N\right]^T$,

$$\mathbf{Y}(\boldsymbol{\Psi})=\begin{bmatrix}\mathbf{y}(\Psi_1)\\ \vdots\\ \mathbf{y}(\Psi_N)\end{bmatrix}, \tag{8}$$

$\mathbf{y}(\Psi_n)=\left[Y_0^0(\Psi_n)\,Y_1^{-1}(\Psi_n)\,Y_1^0(\Psi_n)\,Y_1^1(\Psi_n)\ldots Y_L^L(\Psi_n)\right]$ and $(\cdot)^H$ denotes the conjugate transpose.

The SH domain representation of the plane-wave density, as expressed in (7), is useful because the steering vectors, $\mathbf{y}(\Psi_n)$, are analytic functions which are independent of frequency. In order to obtain this representation, the sound field in the vicinity of the point of interest must be observed. The pressure at a particular point is related to the plane-wave density by the mode strength, which depends on the distance of the point from the origin and whether a rigid scatterer is present [2], [3], [30]. Although irregular sampling schemes are possible, for mathematical convenience we use the pressure on the surface of a sphere of radius r centered at the origin, $p(k,r,\Omega)$, for which the mode strength can be denoted $b_l(kr)$. The SFT of this function is

$$\mathbf{p}_{lm}(k,r)=\mathbf{B}(kr)\mathbf{a}_{lm}(k) \tag{9}$$

where $\mathbf{B}(kr)=\operatorname{diag}\{b_0\,b_1\,b_1\,b_1\ldots b_L\}$, $\mathbf{p}_{lm}(k,r)=\left[p_{00}\,p_{1(-1)}\,p_{10}\,p_{11}\ldots p_{LL}\right]^T$ is a vector of SH coefficients and the functional dependence of the stacked terms has been omitted for clarity. Sampling $p(k,r,\Omega)$ at Q points with directions $\{\Omega_q\}_1^Q$, the SFT is approximated using the discrete SFT [4]

$$\mathbf{p}_{lm}(k,r)\cong\mathbf{Y}(\boldsymbol{\Omega})^H\mathbf{W}\mathbf{p}(k,r) \tag{10}$$

where $\mathbf{p}(k,r)=\left[p_1\ldots p_Q\right]^T$ is the pressure at each of the sample points, $\mathbf{W}=\operatorname{diag}\{w_1\,w_2\ldots w_Q\}$, where $\{w_q\}_1^Q$ are the weights of the sampling scheme, and $\mathbf{Y}(\boldsymbol{\Omega})$ is a $Q\times(L+1)^2$ matrix defined as in (8) but with the SHs evaluated at $\{\Omega_q\}_1^Q$. For the approximation in (10) to hold up to the maximum spherical harmonic order, L, requires that there are sufficient microphones, $Q\geq(L+1)^2$, and that they are adequately distributed over the sphere [31]. Furthermore, for a given radius, the error in the approximation of (10) increases with frequency. In practice the upper threshold is commonly taken as $kr<L$ [7], [12], although to avoid spatial aliasing requires $kr\ll L$ [2], [31]. It has also been shown that to accurately reproduce the pressure at a point due to a plane wave using the inverse SFT requires a much more conservative threshold [32].

Equating (9) and (10) the plane-wave density can be obtained as

$$\mathbf{a}_{lm}(k)=\mathbf{B}(kr)^{-1}\mathbf{Y}(\boldsymbol{\Omega})^H\mathbf{W}\mathbf{p}(kr). \tag{11}$$

Let $\mathbf{x}(k,r)=\mathbf{p}(k,r)+\mathbf{v}(k)$ be the observation of $\mathbf{p}(k,r)$ in the presence of sensor noise which we assume to be zero-mean, normally distributed, uncorrelated between sensors and uncorrelated with $\mathbf{s}(k)$. Applying the SFT and compensating for the mode strength, the observed plane-wave density is

$$\tilde{\mathbf{x}}_{lm}(k)=\mathbf{Y}(\boldsymbol{\Psi})^H\mathbf{s}(k)+\tilde{\mathbf{v}}_{lm}(k) \tag{12}$$

where

$$\tilde{\mathbf{v}}_{lm}(k)=\mathbf{B}(kr)^{-1}\mathbf{Y}(\boldsymbol{\Omega})^H\mathbf{W}\mathbf{v}(k). \tag{13}$$

III. Pseudo-intensity vector formulation

The pseudo-intensity vector was proposed in [8] as an approximation to the active intensity vector. This approach is reviewed in Sec. III-A while in Sec. III-B an equivalent vector is derived directly from the SH representation of the sound field. Finally, in Sec. III-C, the SSPIV is formulated.

A. Review of sound intensity and pseudo-intensity

The active intensity vector is defined as the time-averaged magnitude and direction of the net flow of energy and is given by [30]

$$\mathbf{I}(k)=\frac{1}{2}\mathcal{R}\left\{p(k)^*\mathbf{u}(k)\right\} \tag{14}$$

where $p(k)$ is the omnidirectional pressure, $\mathbf{u}(k)=\left[u_x(k)\;u_y(k)\;u_z(k)\right]^T$ is a vector of the particle velocities in the Cartesian directions and $\mathcal{R}\{\cdot\}$ is the real operator. It is useful for DOA estimation because acoustic energy flows in the direction of wave propagation. For a planewave, the particle velocity vector is related to direction of arrival (θ,ϕ) as [8]

$$\mathbf{u}(k)=-\frac{p(k)}{\rho_0 c}\begin{bmatrix}\sin\theta\cos\phi\\ \sin\theta\sin\phi\\ \cos\theta\end{bmatrix} \tag{15}$$

where ρ_0 and c are the ambient density and speed of sound in the medium, respectively. It can be seen that the elements of $\mathbf{u}(k)$ have dipole directivity patterns aligned with the Cartesian axes and that the resulting vector points in the opposite direction from the DOA.

A beamformer with a dipole directivity pattern can be obtained directly from first order SH coefficients as

$$D(k,\varphi,\mathbf{a}_{lm}(k))=\sum_{m=-1}^{1}Y_1^m(\varphi)a_{1(m)}(k) \tag{16}$$

where φ is the steering direction. Therefore to approximate (14) using the SH coefficients of the plane-wave density function the PIV is formulated [8]

$$\mathbf{I}\left(k\right) = \frac{1}{2}\mathcal{R}\left\{ a_{00}(k)^* \begin{bmatrix} D(k, \varphi_{-x}, \mathbf{a}_{lm}(k)) \\ D(k, \varphi_{-y}, \mathbf{a}_{lm}(k)) \\ D(k, \varphi_{-z}, \mathbf{a}_{lm}(k)) \end{bmatrix} \right\} \tag{17}$$

where $\varphi_{-x} = (\pi/2, \pi)$, $\varphi_{-y} = (\pi/2, -\pi/2)$ and $\varphi_{-z} = (\pi, 0)$.

B. Alternative formulation of PIV

From (6) the plane-wave decomposition for the n-th plane wave is $a_{lm}^{(n)}(k) = \left[Y_l^m(\Psi_n)\right]^* s_n\left(k\right)$. Expressing the first order coefficients in Cartesian form gives

$$\begin{aligned} a_{1(-1)}^{(n)}(k) &= s_n(k)\sqrt{3/8\pi}\left(x_n + iy_n\right) && \text{(18a)} \\ a_{10}^{(n)}(k) &= s_n(k)\sqrt{3/4\pi}z_n && \text{(18b)} \\ a_{11}^{(n)}(k) &= s_n(k)\sqrt{3/8\pi}\left(-x_n + iy_n\right). && \text{(18c)} \end{aligned}$$

where the SHs are evaluated on the unit sphere. Rearranging (18) gives

$$\begin{aligned} s_n(k)x_n &= \sqrt{\frac{8\pi}{3}}\frac{1}{2}\left(a_{1(-1)}^{(n)}(k) - a_{11}^{(n)}(k)\right) && \text{(19a)} \\ s_n(k)y_n &= \sqrt{\frac{8\pi}{3}}\frac{1}{2i}\left(a_{1(-1)}^{(n)}(k) + a_{11}^{(n)}(k)\right) && \text{(19b)} \\ s_n(k)z_n &= \sqrt{\frac{8\pi}{3}}\frac{1}{\sqrt{2}}a_{10}^{(n)}(k). && \text{(19c)} \end{aligned}$$

which can be interpreted as a weighted sum of the 1-order plane-wave decomposition coefficients. Moreover, the weight corresponding to each $a_{1(m)}^{(n)}(k)$ is proportional to the order 1, degree m SH evaluated in the required axial direction as

$$\begin{aligned} s_n(k)\varpi_n &= \frac{4\pi}{3}\sum_{m=-1}^{1} Y_1^m(\varphi_\varpi)a_{1(m)}^{(n)}(k) && \text{(20)} \\ &= \frac{4\pi}{3}D(k, \varphi_\varpi, \mathbf{a}_{lm}^{(n)}(k)) && \text{(21)} \end{aligned}$$

where (16) has been used to obtain (21), $\varpi \in \{x, y, z\}$, $\varphi_x = (\pi/2, 0)$, $\varphi_y = (\pi/2, \pi/2)$ and $\varphi_z = (0, 0)$. To obtain a vector pointing towards the n-th DOA, we note that $a_{00}^{(n)}(k) = \sqrt{\frac{1}{4\pi}}s_n(k)$ and evaluate (23) for $\varpi \in \{x, y, z\}$ leading to

$$\begin{aligned} \tilde{\mathbf{I}}(k) &= \frac{4\pi\sqrt{4\pi}}{3}\mathcal{R}\left\{ a_{00}^{(n)}(k)^* \begin{bmatrix} D(k, \varphi_x, \mathbf{a}_{lm}^{(n)}(k)) \\ D(k, \varphi_y, \mathbf{a}_{lm}^{(n)}(k)) \\ D(k, \varphi_z, \mathbf{a}_{lm}^{(n)}(k)) \end{bmatrix} \right\} && \text{(22)} \\ &= \mathcal{R}\left\{ \left|s_n(k)\right|^2 \mathbf{x}_n \right\} && \text{(23)} \end{aligned}$$

where, for a single plane wave in noise free conditions, the argument to the real operator is intrinsically real but the real operator may be needed in practical implementations with finite precision. The direction of the PIV in spherical coordinates can be extracted using (1) from the unit vector given by $\tilde{\mathbf{I}}(k)/\left\|\tilde{\mathbf{I}}(k)\right\|$ where $\|\cdot\|$ denotes the ℓ_2-norm.

The formulation of (22) is structurally identical to (17), but $\mathbf{I}\left(k\right)$ and $\tilde{\mathbf{I}}(k)$ point in opposite directions due to the steering of the dipoles. Moreover, the inclusion of the $4\pi\sqrt{4\pi}/3$ normalizing constant in (22) leads to the simplified form of (23) which will make the notation of the subsequent analysis more straightforward. For historical reasons $\tilde{\mathbf{I}}(k)$ is hereafter referred to as the PIV but its orientation towards the DOA is preferred for simplicity of describing the methods.

C. Subspace PIV

The SSPIV extends the concept of PIVs to take advantage of higher order SHs and frequency smoothing and is aimed at providing more accurate and reliable DOA estimates in the presence of multiple and interfering sound sources and reverberation. It follows from (7) that the covariance of $\mathbf{a}_{lm}$ is [12]

$$\begin{aligned} \mathbf{R}_{\mathbf{a}_{lm}} &= E\left\{\mathbf{a}_{lm}\mathbf{a}_{lm}^H\right\} && \text{(24)} \\ &= \mathbf{Y}^H(\mathbf{\Psi})\mathbf{R}_\mathbf{s}\mathbf{Y}(\mathbf{\Psi}) && \text{(25)} \end{aligned}$$

where $\mathbf{R}_\mathbf{s} = E\left\{\mathbf{s}\mathbf{s}^H\right\}$. Singular Value Decomposition (SVD) leads to

$$\mathbf{R}_{\mathbf{a}_{lm}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H = \left[\mathbf{U}_s\mathbf{U}_n\right]\begin{bmatrix} \mathbf{\Sigma}_s & 0 \\ 0 & \mathbf{\Sigma}_n \end{bmatrix}\begin{bmatrix} \mathbf{U}_s^H \\ \mathbf{U}_n^H \end{bmatrix} \tag{26}$$

where $\mathbf{U}$ is a unitary matrix, $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values of $\mathbf{R}_{\mathbf{a}_{lm}}$ and $\mathbf{U}_s$ and $\mathbf{U}_n$ respectively, represent the conventional partitioning into signal and noise subspaces [15]. In the simplest case of a single plane wave, $\mathbf{U}_s = \left[\hat{a}_{00}\,\hat{a}_{1(-1)}\,\hat{a}_{10}\,\hat{a}_{11}\ldots\hat{a}_{LL}\right]^T$ is a column vector and is proportional to the steering vector for the plane wave DOA, $\mathbf{y}(\Psi_n)$. The SSPIV method applies the PIV method (c.f. (22) and (16)) to the one-dimensional signal subspace as

$$\tilde{\mathbf{I}}_{\text{ss}} = \frac{4\pi\sqrt{4\pi}}{3}\mathcal{R}\left\{ \hat{a}_{00}^* \begin{bmatrix} D(k, \varphi_x, \mathbf{U}_s) \\ D(k, \varphi_y, \mathbf{U}_s) \\ D(k, \varphi_z, \mathbf{U}_s) \end{bmatrix} \right\}. \tag{27}$$

to obtain a vector pointing towards the source. Whilst (27) depends only on the 0 and 1st order components of $\mathbf{U}_s$, through (25) and (26), their values do depend on the higher order SH terms of $\mathbf{a}_{lm}$. As with the PIV method, the benefit of this approach is that a direction is obtained for each TF-region directly, i.e. without evaluating all possible directions. The implications of violating the assumption that a single plane wave is present is addressed in Sec. IV.

IV. PIV AND SSPIV DISTRIBUTIONS FOR REPRESENTATIVE EXAMPLE SOUND FIELDS

As described in Sec. II, an arbitrary sound field can be decomposed into a sum of plane waves. In this section we consider how the PIVs and SSPIVs are affected by amplitude, phase and directional relationships between two plane waves. These simplified cases provide some insight into the behavior of pseudo-intensity vectors in real acoustic environments.

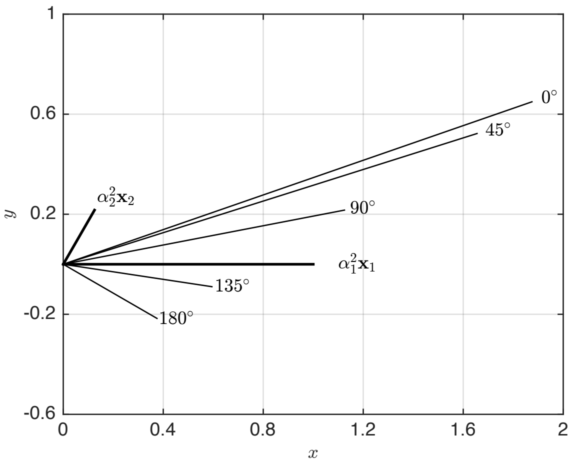


Fig. 1. $\tilde{\mathbf{I}}$ for selected values of $|\beta_1 - \beta_2|$ with fixed α_1, α_2, $\mathbf{x}_1$, and $\mathbf{x}_2$. $\alpha_1^2\mathbf{x}_1$ and $\alpha_2^2\mathbf{x}_2$ are shown for reference.

A. Two plane waves - general case

For two plane waves with DOAs given by the unit vectors, $\mathbf{x}_n$, $n = \{1, 2\}$, and source signals $s_n(k) = \alpha_n(k)e^{i\beta_n(k)}$, where $\alpha_n(k)$ and $\beta_n(k)$ are the magnitude and phase at the origin, respectively, the PIV is obtained from (2) and (22) as

$$\tilde{\mathbf{I}} = \mathcal{R}\left\{(s_1 + s_2)^*(s_1\mathbf{x}_1 + s_2\mathbf{x}_2)\right\}. \tag{28}$$

$$= \alpha_1^2\mathbf{x}_1 + \alpha_2^2\mathbf{x}_2 + (\mathbf{x}_1 + \mathbf{x}_2)\,\alpha_1\alpha_2\cos(\beta_1 - \beta_2) \tag{29}$$

where for brevity the dependence on k is assumed. This is interesting because it implies that the resulting vector lies on the plane containing the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ but that it does not necessarily lie between the two. To illustrate this point, Fig. 1 shows $\tilde{\mathbf{I}}$ for various values of $|\beta_1 - \beta_2|$. The resulting vector is nominally distributed about the direction of the stronger source (i.e. $\mathbf{x}_1$) but is either drawn towards the direction of the weaker source (i.e. $\mathbf{x}_2$) or repelled from it, depending on the relative amplitudes and phases of the signals.

The SSPIV depends on the SVD of $\mathbf{R}_{\mathbf{a}_{lm}}$, which is determined primarily by the source covariance, $\mathbf{R}_\mathbf{s} = \begin{bmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, where σ_1^2 and σ_2^2 are the variances of the two plane waves and σ_{12}, σ_{21} is their covariance. The dimensionality of $\mathbf{R}_{\mathbf{a}_{lm}}$ depends on the maximum SH order, L, but is independent of the number of plane waves.

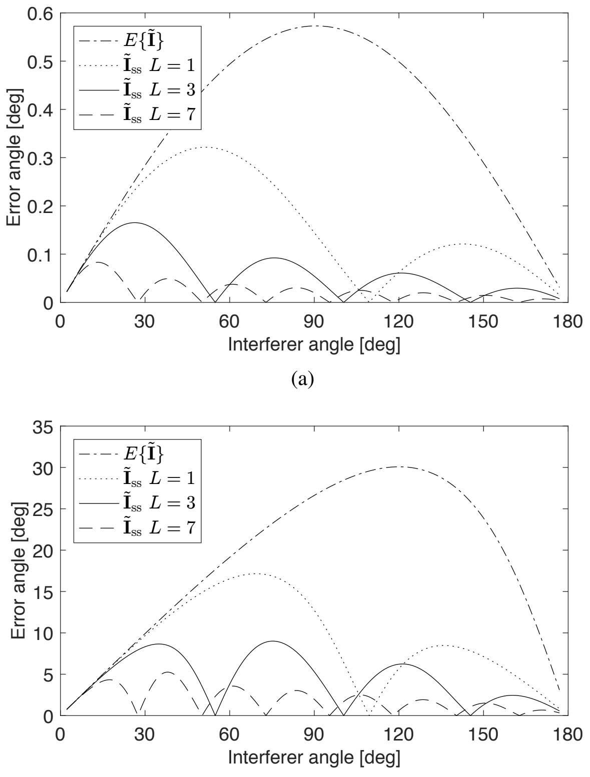
B. Uncorrelated sources

Consider two uncorrelated sources in a free-field with fixed DOAs and amplitude ratio. We assume that β_1 and β_2 are independent with identical uniform distribution $\mathcal{U}(0, 2\pi)$ such that $\Delta\beta = \beta_1 - \beta_2$ is a triangular distribution over the interval $\Delta\beta \in [-2\pi, 2\pi]$ which, due to periodicity of the phase, reduces to $\Delta\beta \in [-\pi, \pi]$ with probability $p(\Delta\beta) = 1/(2\pi)$. The expected value of $\tilde{\mathbf{I}}$ is obtained by integrating (29) with respect to $\Delta\beta$,

$$E\left\{\tilde{\mathbf{I}}\right\} = \int_{-\pi}^{\pi} \tilde{\mathbf{I}} p(\Delta\beta)\, d\Delta\beta \tag{30}$$

$$= \alpha_1^2\mathbf{x}_1 + \alpha_2^2\mathbf{x}_2 + \frac{(\mathbf{x}_1 + \mathbf{x}_2)\,\alpha_1\alpha_2}{2\pi}\int_{-\pi}^{\pi}\cos(\beta_1 - \beta_2)\, d\Delta\beta \tag{31}$$

$$= \alpha_1^2\mathbf{x}_1 + \alpha_2^2\mathbf{x}_2. \tag{32}$$

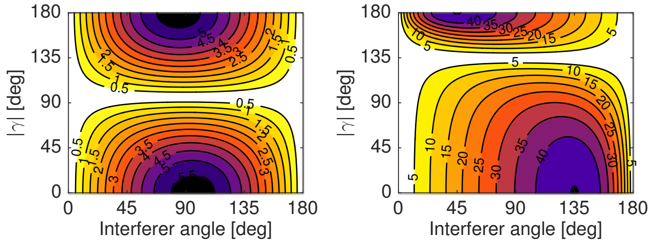


(a)

(b)

Fig. 2. Error in $E\left\{\tilde{\mathbf{I}}\right\}$ and $\tilde{\mathbf{I}}_{\text{ss}}$ for for $L = \{1, 3, 7\}$ as function of $\angle(\mathbf{x}_1, \mathbf{x}_2)$ with (a) SIR 20 dB and (b) SIR 3 dB.

The SSPIV is determined by the source covariance, $\mathbf{R}_\mathbf{s} \propto \begin{bmatrix} \alpha_1^2 & 0 \\ 0 & \alpha_2^2 \end{bmatrix}$, the DOAs and the maximum order of SHs considered. Without loss of generality, let $\mathbf{x}_1$ point in the direction of the desired source such that the Signal-to-Interference Ratio (SIR) in dB is $10\log_{10}(\alpha_1^2/\alpha_2^2) \geq 0$. Figure 2 shows the error angle $\angle(\mathbf{x}_1, \tilde{\mathbf{I}}_{\text{ss}})$ as a function of the interferer angle $\angle(\mathbf{x}_1, \mathbf{x}_2)$ for SIRs of 20 dB and 3 dB and for different values of L. These plots were produced by collating, without averaging, SSPIVs calculated according to (25), (26) and (27) for interferers at 794 approximately equally distributed directions and 5 random target directions. The variation in error as a function of interferer angle has multiple peaks and nulls corresponding to the number of lobes in the real (or imaginary) part of highest order SH considered but is independent of the target direction. Increasing L reduces the worst case error, which confirms that higher order SHs are being utilized by the SSPIV. Also shown is $\angle(\mathbf{x}_1, E\left\{\tilde{\mathbf{I}}\right\})$ where $E\left\{\tilde{\mathbf{I}}\right\}$ is calculated according to (32). Figure 2(a) shows that PIVs and SSPIVs are both accurate to



(a) (b)

Fig. 3. Error in $\tilde{\mathbf{I}}$ due to coherent source as a function of incidence angle and phase difference with SIR of (a) 20 dB and (b) 3 dB. Higher errors shown as darker colors.

within 1° when a single source is dominant but Fig. 2(b) shows that when the SIR is low, SSPIVs are substantially more accurate than expected by averaging PIVs.

C. Coherent sources

Multipath propagation, as encountered in enclosed spaces, leads to coherent plane waves arriving at the microphone array. We consider here the simplified case of two incident plane waves where the second is a delayed and attenuated version of the first. The relative gain, $0 \leq g \leq 1$, and phase, $-\pi < \gamma \leq \pi$, of the second plane wave with the respect to the first give $\alpha_2 = g\alpha_1$ and $\beta_2 = \beta_1 + \gamma$. Therefore, from (29),

$$\tilde{\mathbf{I}} = \alpha_1^2\left[(1+g\cos\gamma)\mathbf{x}_1 + g(g+\cos\gamma)\mathbf{x}_2\right] \qquad (33)$$

which by inspection has a number of special cases. If $\cos\gamma = -g$ or $\mathbf{x}_2 = -\mathbf{x}_1$ the terms in $\mathbf{x}_2$ cancel leaving $\tilde{\mathbf{I}} = \left(\alpha_1^2(1-g^2)\mathbf{x}_1\right)$ which is the desired direction. From the law of sines, the error angle is

$$\angle(\mathbf{x}_1, \tilde{\mathbf{I}}) = \cos^{-1}\left(\frac{c + a\cos\theta}{\sqrt{(a^2 + c^2 + 2ac\cos\theta}}\right) \qquad (34)$$

where $c = 1 + g\cos\gamma$, $a = g(g + \cos\gamma)$ and $\theta = \angle(\mathbf{x}_1, \mathbf{x}_2)$. Figure 3 shows $\angle(\mathbf{x}_1, \tilde{\mathbf{I}})$ as a function of $|\gamma|$ and θ for two different values of g. The error is highly dependent on all the factors but for any interferer angle the error is zero when $\cos|\gamma| = -g$ and increases as $|\gamma| \to 0°$ and $|\gamma| \to 180°$.

If both wavefronts originate from an omnidirectional point source such that first wavefront is the direct path and the second has undergone one or more reflections, g is related to the individual wavefront amplitudes according to

$$\alpha_1 \propto \frac{1}{r}, \quad \alpha_2 \propto \frac{\rho(k)}{r+d}, \quad g = \frac{r\rho(k)}{r+d} \qquad (35)$$

where r is the distance of the source from the origin, d is the path difference and $\rho(k)$ is the accumulation of the reflection coefficients associated with all the reflections encountered by the interfering wavefront. Similarly, the phase difference is given by $\gamma = kd/(2\pi)$. Thus, a reflected wavefront will produce a systematic bias in $\tilde{\mathbf{I}}(k)$ which is periodic with frequency and independent of the direct path signal strength.

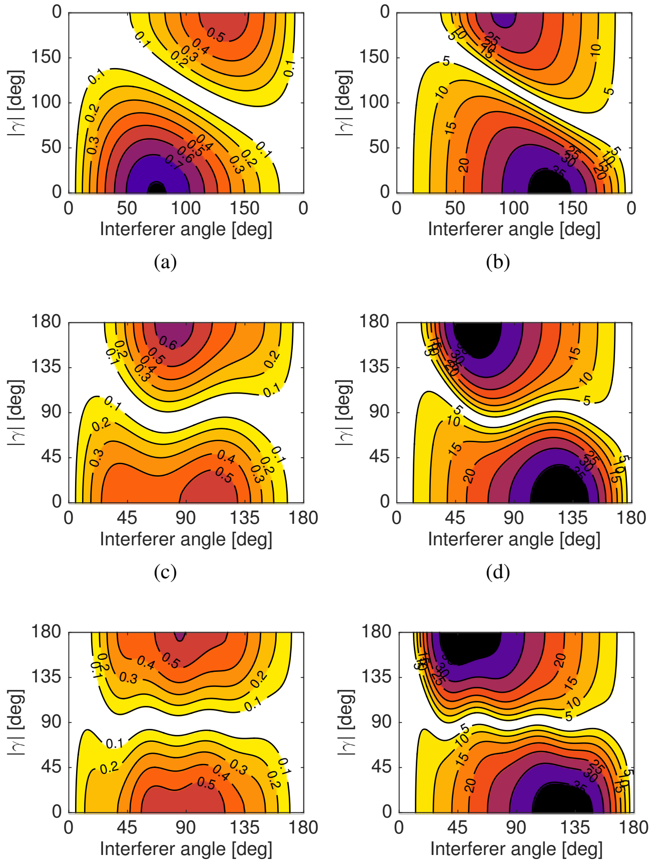


Fig. 4. Error in $\tilde{\mathbf{I}}_{\text{ss}}$ due to coherent source as a function of incidence angle and phase difference with SIR of (a,c,e) 20 dB and (b,d,f) 3 dB and (a,b) L=1, (c,d) L=3 and (e,f) L=7. Higher errors shown as darker colors.

Assuming g is independent of frequency over a bandwidth of Δ_k and $p(k) = 1/\Delta_k$, the expected PIV is

$$E\left\{\tilde{\mathbf{I}}(k)\right\} = \int_{k-\Delta_k/2}^{k+\Delta_k/2} \tilde{\mathbf{I}}(k)p(k)dk \qquad (36)$$

$$= \alpha_1^2\left(\mathbf{x}_1 + g^2\mathbf{x}_2 + f(k, \Delta_k)\right) \qquad (37)$$

where $f(k, \Delta_k) = \frac{4\pi g}{\Delta_k d}(\mathbf{x}_1 + \mathbf{x}_2)\sin\left(\frac{\Delta_k d}{4\pi}\right)\cos\left(\frac{kd}{2\pi}\right)$. The term $\frac{4\pi}{\Delta_k d}\sin\left(\frac{\Delta_k d}{4\pi}\right)$ is a sinc function which takes its maximum at $\Delta_k d = 0$. As Δ_k becomes large $E\left\{\tilde{\mathbf{I}}(k)\right\} \to \alpha_1^2\left(\mathbf{x}_1 + g^2\mathbf{x}_2\right)$ which is independent of k but over a limited bandwidth the $f(k, \Delta_k)$ term leads to significant variations.

The SSPIV for two correlated sources depends on the source covariance

$$\mathbf{R_s} = \begin{bmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \sigma_1^2 \begin{bmatrix} 1 & g^2\cos\gamma \\ g^2\cos\gamma & g^2 \end{bmatrix}.$$

The effect of off-diagonal elements in $\mathbf{R_s}$ on $\angle(\tilde{\mathbf{I}}_{\text{ss}}, \mathbf{x}_1)$ is shown in Fig. 4. In contrast to the uncorrelated case, increasing L does not substantially reduce the error. To reduce the correlation between the direct path and the reflection we

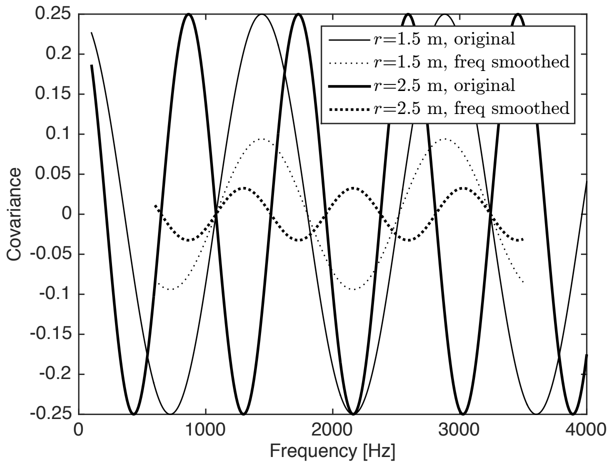


Fig. 5. Effect of frequency smoothing: covariance as a function of frequency for coherent reflection with g=0.5 for r=d=1.5 m and r=d=2.5 m before (solid) and after (dashed) frequency smoothing over bandwidth of 1 kHz.

consider Frequency Smoothing [7], [10], [12]. For a monopole source with single reflection as in (35) the covariance is

$$\sigma_{12}/\sigma_1^2 = g^2 \cos\left(\frac{kd}{2\pi}\right) \tag{38}$$

which varies periodically with frequency. Integrating (38) over frequency, similar to (36), gives

$$\begin{aligned}\int_{k-\Delta_k/2}^{k+\Delta_k/2} \frac{\sigma_{12}}{\sigma_1^2} p(k)dk =& \frac{4\pi g^2}{\Delta_k d}\sin\left(\frac{\Delta_k d}{4\pi}\right)\cos\left(\frac{kd}{2\pi}\right)\\ =& \frac{4\pi}{\Delta_k d}\sin\left(\frac{\Delta_k d}{4\pi}\right)\frac{\sigma_{12}}{\sigma_1^2} \qquad (39)\end{aligned}$$

where again the multiplicative factor introduced by the integration can be recognized as a sinc function whose absolute value is guaranteed to be less than one for all $\Delta_k d > 0$. This shows that FS decorrelates the coherent reflection. Figure 5 illustrates the extent of the decorrelation for two different values of d. The value of r is adjusted to maintain the same value of g in both cases. Larger path differences cause the covariance to change more rapidly with frequency such that a particular integration bandwidth achieves more decorrelation.

D. Single desired source in spherically isotropic noise

Diffuse (ambient) noise and late reverberation are well modeled by spherically isotropic noise. In this case $\mathbf{x}_2$ points in all directions with equal probability, $p(\mathbf{x}_2) = 1/(4\pi)$, so (32) integrated over all possible directions of $\mathbf{x}_2$ (i.e. over the surface of a unit sphere) becomes

$$\begin{aligned}E\left\{\tilde{\mathbf{I}}\right\} &= \int_{\phi_2}\int_{\theta_2} \left(\alpha_1^2\mathbf{x}_1 + \alpha_2^2\mathbf{x}_2\right) p(\mathbf{x}_2)\sin\theta_2 d\theta_2 d\phi_2\\ &= \alpha_1^2\mathbf{x}_1 \qquad (40)\end{aligned}$$

which is simply a scaled version of $\mathbf{x}_1$, the desired source direction. This suggests that regardless of the noise amplitude, the expected value of the PIV provides an unbiased estimate of the source direction.

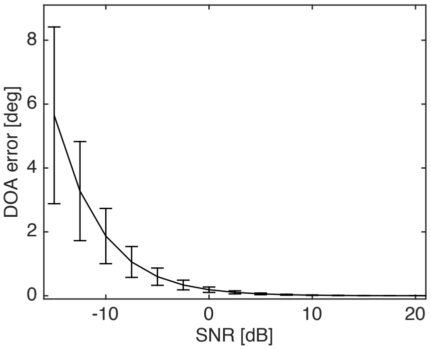


Fig. 6. Mean error in SSPIV DOA as a function of SNR of spherically isotropic noise averaged over 794 approximately equally distributed target directions, $L = 3$. Error bars show ±1 standard deviation.

The plane-wave density of spherically isotropic noise is given by substituting $a^{(2)}(k,\theta,\phi) = \alpha_2/(4\pi)$ into (5). The covariance between the SH coefficients is thus

$$\begin{aligned}a_{lm}^{(2)}\left[a_{l'm'}^{(2)}\right]^* &= \left(\frac{|\alpha_2|}{4\pi}\right)^2 \int_{\Omega\in S^2} \left[Y_l^m(\Omega)\right]^* Y_{l'}^{m'}(\Omega)d\Omega\\ &= \left(\frac{|\alpha_2|}{4\pi}\right)^2 \delta_{l-l',m-m'}\end{aligned}$$

where the simplification in the second line arises from the orthogonality of SHs [4, p. 11]. The complete noise covariance matrix is therefore

$$\mathbf{R}_{\mathbf{a}_{lm}}^{(2)} = \left(\frac{|\alpha_2|}{4\pi}\right)^2 \mathbf{I}_{[(L+1)^2\times(L+1)^2]} \tag{41}$$

and the total covariance is

$$\mathbf{R}_{\mathbf{a}_{lm}} = |\alpha_1|^2\,\mathbf{y}(\Psi_1)\mathbf{y}(\Psi_1)^H + \mathbf{R}_{\mathbf{a}_{lm}}^{(2)}. \tag{42}$$

The effect of diagonal loading of $\mathbf{R}_{\mathbf{a}_{lm}}$ due to spherically isotropic noise on the calculated SSPIV is shown in Fig. 6. The error is negligible for positive SNRs and the average error is less than 2° with SNR of -10 dB.

E. Single desired source plus spatially white noise

Sensor noise is typically modeled as spatially white noise such that $E\left\{\mathbf{v}\mathbf{v}^H\right\} = \sigma_v^2\mathbf{I}_{[Q\times Q]}$ where σ_v^2 is the variance of the noise, which is assumed to be the same at all sensors.

The spatial covariance matrix in the presence of spatially white noise is

$$\begin{aligned}\mathbf{R}_{\tilde{\mathbf{x}}_{lm}} &= E\left\{\tilde{\mathbf{x}}_{lm}\tilde{\mathbf{x}}_{lm}^H\right\} \qquad (43)\\ &= \mathbf{R}_{\mathbf{a}_{lm}} + \mathbf{R}_{\tilde{\mathbf{v}}_{lm}} \qquad (44)\end{aligned}$$

where $\mathbf{R}_{\mathbf{a}_{lm}}$ is defined in (25) and $\mathbf{R}_{\tilde{\mathbf{v}}_{lm}} = E\left\{\tilde{\mathbf{v}}_{lm}\tilde{\mathbf{v}}_{lm}^H\right\}$. Substituting (13) gives

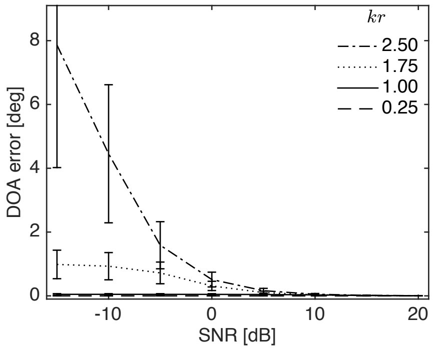


Fig. 7. Mean error in SSPIV DOA as a function of SNR of spatially white noise averaged over 794 approximately equally distributed target directions for different values of kr assuming a rigid spherical microphone array with $Q = 32$ and $L = 3$. Error bars show ±1 standard deviation.

$$\mathbf{R}_{\tilde{\mathbf{v}}_{lm}}(kr) = E\left\{\tilde{\mathbf{v}}_{lm}\tilde{\mathbf{v}}_{lm}^{H}\right\} \tag{45}$$

$$= E\Big\{\mathbf{B}(kr)^{-1}\mathbf{Y}\left(\boldsymbol{\Omega}\right)^{H}\mathbf{W}\mathbf{v}(k) \tag{46}$$

$$\cdot\mathbf{v}(k)^{H}\mathbf{W}^{H}\mathbf{Y}\left(\boldsymbol{\Omega}\right)\left[\mathbf{B}(kr)^{-1}\right]^{H}\Big\} \tag{47}$$

$$\approx \sigma_v^2\frac{4\pi}{Q}\mathbf{B}(kr)^{-1}\left[\mathbf{B}(kr)^{-1}\right]^{H} \tag{48}$$

where $\mathbf{Y}\left(\boldsymbol{\Omega}\right)^{H}\mathbf{W}\mathbf{W}^{H}\mathbf{Y}\left(\boldsymbol{\Omega}\right) \approx (4\pi/Q)\,\mathbf{I}_{[Q\times Q]}$. Thus $\mathbf{R}_{\tilde{\mathbf{v}}_{lm}}(kr)$ is a diagonal matrix whose elements vary with diag$\left\{\frac{1}{|b_0(kr)|^2}\;\frac{1}{|b_1(kr)|^2}\;\frac{1}{|b_1(kr)|^2}\;\frac{1}{|b_1(kr)|^2}\cdots\frac{1}{|b_L(kr)|^2}\right\}$ which is frequency dependent. At low frequencies, where $L \gg kr$, $b_l(kr)$ decreases with l and so the squared reciprocal of the higher order terms dominate $\mathbf{R}_{\tilde{\mathbf{v}}_{lm}}(kr)$. This turns out to have relatively little effect on the SSPIV accuracy. On the other hand, at higher frequencies, $b_l(kr)$ is more similar for different values of l making $\mathbf{R}_{\tilde{\mathbf{v}}_{lm}}(kr)$ closer to a scaled identity matrix. For a particular SNR this has a more detrimental effect on the SSPIV accuracy. Substituting the expressions in (25) and (48) into (44), the average error in the SSPIV DOA is shown in Fig. 7 as a function of SNR for different values of kr for 32 sensors on a rigid spherical baffle and $L = 3$. The results suggest that spatially white noise, such as sensor noise, at positive SNRs will cause $< 1°$ error in the SSPIVs. The effect of estimation errors in the spatial covariance matrices is addressed through numerical simulations and real experiments in Sec. V and VI, respectively.

V. Numerical Experiments

The PIV and SSPIV-based DOA estimation methods are compared to two baseline methods from the literature for single and multiple speech sources in a simulated reverberant environment. Performance in a real acoustic environment is later described in Sec. VI.

A. Simulation setup

Acoustic Impulse Responses (AIRs) were simulated up to 6th order for a 32-element rigid spherical microphone array with radius 4.2 cm centered at Cartesian co-ordinates (2.0 m, 2.5 m, 1.5 m) in a 5×6×3 m rectangular room using the image-source method [33] modified to account for the scattering of a rigid sphere [34].

Anechoic speech for each of 5 male and 5 female speakers arbitrarily selected from the TIMIT database [35] was concatenated (without inserting any pauses) and randomly segmented into ten 6-second sections per speaker. In each trial for each source a randomly selected segment was convolved with the simulated AIR to every microphone, the leading 2 seconds were removed to ensure that the amount of reverberation was consistent across the whole segment and the level adjusted such that the direct path was normalised according to [36]. Independent realizations of white Gaussian noise were added to each of the 32 microphone signals to give the desired direct path SNR with respect to each source.

Condition 1: In each trial, a single source was placed 1.5 m from the array in one of 24 directions given by all combinations of $\phi \in \{0°, 30°, \ldots, 330°\}$ and $\theta \in \{80°, 100°\}$. Within each test condition, 4 different speech signals were used, giving 96 test samples per condition.

Condition 2: Four sources were placed 1.5 m from the array at 60° intervals in azimuth alternating between 80° and 100° inclination. To ensure the specific locations of the sources did not bias the results, 12 possible source orientations were separately tested by rotating the azimuth angles of all 4 sources in 30° increments. For each orientation of the four sources, 8 combinations of speech signals were generated giving a total of 96 test samples per condition.

Condition 3: Three moving source trajectories were simulated lasting 10 seconds. For the first second the sources were positioned at (80°, 330°), (100°, 210°) and (80°, 90°). For the following 8 seconds the first source was stationary, while the azimuths of the second and third sources followed sinusoidal trajectories each with an amplitude of 30° and periods of 8 s and 16 s, respectively, ending at (100°, 150°) and (80°, 90°). For the last second all three sources were again stationary. The trajectories were quantized to the nearest 5° azimuth and the relevant segments of clean speech convolved with the appropriate AIR using an overlap-add scheme. For each of the 10 speakers a single 10 second clean speech segment was used in this case. In each of 30 trials a different random assignment of speakers to source trajectories was used.

B. Calculation of PIVs and SSPIVs

For each trial, the 32 discrete-time microphone signals were transformed into the Short Time Fourier Transform (STFT) domain using a Hamming window with 75% overlap. The stacked signal vector $\mathbf{x}(k, r)$ defined after (11) is then reformulated as $\mathbf{x}(\nu, \ell)$ where ν and ℓ are the STFT frequency index and frame index, respectively. An initial pilot study was performed to investigate the choice of frame size. The results indicated that frames of 4 to 8 ms gave the highest concentration of PIVs around the true DOAs compared to frames $\geq$16 ms. The short frames increase the probability that the Window-Disjoint Orthogonality (WDO) assumption [37] is true and by having more TF-bins the distribution of the vectors' directions is more regular.

The maximum SH order used was $L = 3$ giving a maximum frequency of 3850 Hz to ensure $kr < L$. The lowest frequency bin was centered at 500 Hz, which avoids excessive noise amplification due to mode strength compensation at lower frequencies. The SFT was applied as in (10) with $\mathbf{x}(\nu,\ell)$ replacing $\mathbf{p}(kr)$ and $w_q - 4\pi/Q \in [-0.0187, 0.0112]$.

The PIVs, $\tilde{\mathbf{I}}(\nu,\ell)$, were calculated according to (22) where k has been replaced with the time-frequency index. The SSPIVs, $\tilde{\mathbf{I}}_{\text{ss}}(\nu,\ell)$, were calculated as detailed in Sec. III-C but using $\hat{\mathbf{R}}_{\tilde{\mathbf{x}}_{lm}}(\nu,\ell)$ in place of $\mathbf{R}_{\mathbf{a}_{lm}}$, which is approximated in the vicinity of frame index ℓ and frequency index ν by

$$\hat{\mathbf{R}}_{\tilde{\mathbf{x}}_{lm}}(\nu,\ell) = \frac{1}{J_\nu J_\ell}\sum_{j_\nu=0}^{J_\nu-1}\sum_{j_\ell=0}^{J_\ell-1}\tilde{\mathbf{x}}_{lm}(\nu+j_\nu,\ell+j_\ell) \times \tilde{\mathbf{x}}_{lm}^{H}(\nu+j_\nu,\ell+j_\ell) \quad (49)$$

where J_ν and J_ℓ are the number of frequency bins and time frames, respectively, included in the average. The averaging across frequency is possible because in the SH domain the steering vectors are independent of frequency [7]. A comprehensive experimental study was performed to investigate the relationship between J_ν, J_ℓ and the STFT frame size. Considering J_ℓ giving time ranges of 4-256 ms, J_ν giving frequency ranges of 125-1000 Hz and STFT frame lengths of 4-64 ms, the results showed that calculating (49) over time and frequency range of 32 ms and 250 Hz, respectively, with STFT frame length of 8 ms gave the best results, although frame lengths in the range 4-16 ms gave very similar results.

C. *Baseline methods*

The proposed intensity-based methods were compared to the classical Plane-Wave Decomposition (PWD)-SRP approach [5] and state-of-the-art FS-MUSIC with DPD test method (DPD-MUSIC) [12].

1) PWD-SRP: The plane-wave decomposition (or regular) beamformer [38] output is formulated as

$$Z(\varphi,\nu,\ell) = \mathbf{w}_{lm}(\nu,\varphi)^H \mathbf{x}_{lm}(\nu,\ell) \quad (50)$$

where φ is the look direction, $\mathbf{w}_{lm}(\nu,\varphi) = \left[W_{00}\, W_{1(-1)}\, W_{10}\, W_{11}\, \ldots W_{LL}\right]^T$ and $W_{lm}(\nu,\varphi)^* = b_l(kr)Y_l^m(\varphi)$. The PWD beamformer maximizes the directivity index and is equivalent to the MVDR under the assumption of an uncorrelated diffuse noise field. The SRP follows directly as $S_{\text{PWD}}(\varphi) = \sum_{\nu,\ell} Z(\varphi,\nu,\ell)Z(\varphi,\nu,\ell)^*$ up to a constant factor. For consistency with the proposed methods, the frame length was 8 ms.

2) DPD-MUSIC: The MUSIC spectrum is calculated from the noise subspace of the frequency-smoothed spatial covariance matrix, as defined in (26), as [12], [15]

$$S_{\text{MUSIC}}(\varphi) = \sum_{(\nu,\ell)\in A} \frac{1}{\|\mathbf{U}_n^H(\nu,\ell)\mathbf{y}(\varphi)^*\|^2} \quad (51)$$

where A defines the subset of TF-region indices which pass the DPD test $A = \{(\nu,\ell) : \eta(\nu,\ell) > \varepsilon\}$ where $\eta(\nu,\ell)$ is the ratio of first to second singular values of $\hat{\mathbf{R}}_{\tilde{\mathbf{x}}_{lm}}(\nu,\ell)$ and $\varepsilon = 6$ is an algorithm parameter, which was set as in [12]. It is assumed that the effective rank of $\hat{\mathbf{R}}_{\tilde{\mathbf{x}}_{lm}}(\nu,\ell)$ in those TF-regions which pass the DPD test is unity and so the noise subspace has dimension $(L+1)^2 - 1$. The frame length and values of J_ν and J_ℓ were set as for SSPIV as this led to improved results in Condition 2 compared to the parameter choice in [12] and allows a direct comparison of computational complexity.

D. *DOA estimation from PIVs and SSPIVs*

A variety of approaches to DOA estimation from a set of vectors have been proposed to deal with single [8], [26] and multiple [11], [14] source situations. The analysis in Sec. IV has shown that the directions of the calculated PIVs or SSPIVs are expected to be concentrated around the DOA(s) of the dominant sound source(s). The approach taken here finds the peaks of a spatial cost function which approximates the probability distribution of vectors as a smoothed histogram over a regular 2D grid of directions.

For practicality of implementation we precompute a dictionary containing a Gaussian kernel centered at each direction lying on a regularly spaced $N_{K\theta} \times N_{K\phi}$ 2D grid. The j_θ, j_ϕ-th element is thus defined as

$$\mathbf{K}_{j_\theta,j_\phi}(\varphi) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\angle\left(\varphi,\psi_{j_\theta,j_\phi}\right)^2}{2\sigma^2}\right) \quad (52)$$

where φ is the direction of interest (or look direction), σ is the standard deviation of the Gaussian kernel, $j_\theta \in \{0\ldots N_{K\theta}-1\}$ and $j_\phi \in \{0\ldots N_{K\phi}-1\}$) denote the inclination and azimuth indices, respectively, of the DOA, $\psi_{j_\theta,j_\phi} = (j_\theta\pi/N_{K\theta}, j_\phi 2\pi/N_{K\phi})$. A sparse dictionary is enforced by setting entries smaller than λ to zero, i.e.

$$\hat{\mathbf{K}}_{j_\theta,j_\phi}(\varphi) = \begin{cases} 0 & \mathbf{K}_{j_\theta,j_\phi}(\varphi) < \lambda \\ \mathbf{K}_{j_\theta,j_\phi}(\varphi) & \text{otherwise} \end{cases}. \quad (53)$$

Defining $\chi(\nu,\ell) \triangleq (\theta_\chi, \phi_\chi)$ to be the direction associated with each TF-bin (region) calculated using PIV (SSPIV), the corresponding dictionary indices are determined according to $\mathcal{J}_\theta(\chi(\nu,\ell)) = \lfloor \theta_\chi N_{K\phi}/\pi + 0.5\rfloor$ and $\mathcal{J}_\phi(\chi(\nu,\ell)) = \lfloor \phi_\chi N_{K\phi}/(2\pi) + 0.5\rfloor$ where $\lfloor\cdot\rfloor$ is the floor operator. Thus the smoothed histogram is calculated without any multiplications as

$$\mathbf{H}(\varphi) = \sum_{(\nu,\ell)\in\mathcal{T}} \hat{\mathbf{K}}_{\mathcal{J}_\theta(\chi(\nu,\ell)),\mathcal{J}_\phi(\chi(\nu,\ell))}(\varphi) \quad (54)$$

where $\mathcal{T}$ is the set of TF-bins in the observation interval.

Finally, the DOAs are estimated as the directions corresponding to all the peaks in $\mathbf{H}$ or the N_d largest peaks, whichever is smaller. This allows an analysis of the trade-off between missed detections and clutter measurements, which is important for tracking moving sources.

The employed method of estimating DOAs from a set of vectors is similar to [18] in the formation of a smoothed histogram but it does noes not use matching pursuits to find the peak positions. This makes our approach more generic (because it does not require Reverberation Time (RT)-dependent dictionary elements) and computationally more efficient, which is significant when considering a reasonably dense 2D search grid ([18] only considered 1D DOA estimation).

The parameters specific to the proposed method were: $\sigma = 4°$; $N_{K\theta}$=91 and $N_{K\phi}$=180, which corresponds to 2° resolution in azimuth and inclination; and $\lambda = 0.001/(\sigma\sqrt{2\pi})$, which removes entries >15° from the look direction.

For all methods (PIV, SSPIV, PWD-SRP and DPD-MUSIC) the corresponding spatial cost function were computed over a 2D grid with 2° resolution in azimuth and inclination. In Conditions 1 (and 2) the number of sources was assumed known *a priori*. Thus, using $N_d = 1$ (and 4) a single (set of) estimated DOA(s) was obtained for each trial by setting the observation interval to the full length of the signal (4 seconds). In Condition 3 the number of sources was assumed unknown. In this case $N_d = 12$ DOA estimates were obtained every 100 ms using a causal 250 ms observation interval. For moving sources the optimal length of observation interval is a trade-off between robustness to noise and the ability to follow the true source direction.

E. Performance metrics

For Condition 1 the Root Mean Square (RMS) error in the DOA of the single source was computed directly. For Condition 2 the four DOA estimates were first associated with the true DOAs. Following the procedure of [12], the error was calculated for each possible permutation and a source was considered found if the error was less than 20°. This limit ensured that any estimated DOAs lying approximately midway between the true DOAs were excluded. The assignment which led to the maximum number of found sources was chosen. The RMS error in the estimated DOAs was calculated across those trials in which all 4 sources were found.

For Condition 3, with 3 moving sources and up to 12 estimated DOAs, there were potentially more estimated DOAs than sources. For each time step an estimated DOAs was assigned to a source if it was within 30° of the true source direction at that time where the wider limit compared to Condition 2 reflects the observed variance in the estimates around the true source DOAs. With this approach, each source could have more than one estimate assigned to it. The RMS error was calculated for each source for all assigned estimates. In many time steps a particular source had no estimates assigned to it. The miss rate for each source was the proportion of time steps in which this occurred. The clutter rate was calculated as the average number of estimates which were not assigned to a source on each time step.

F. Results

Figure 8 shows the angular error in estimated DOA averaged across 96 trials in Condition 1 for each combination of SNR, RT and algorithm. In general, the error increases with reverberation time and noise level. PWD-SRP is the only method not to achieve perfect performance under anechoic conditions while DPD-MUSIC achieves the best performance in all cases. Comparing PIV and SSPIV, any benefit of SSPIV is only apparent under the worst case conditions (RT: 0.7 s, SNR: 10 dB) where SSPIV is 0.4° (18%) more accurate than PIV. This suggests that for single source DOA estimation the PIV method is adequate in our tests.

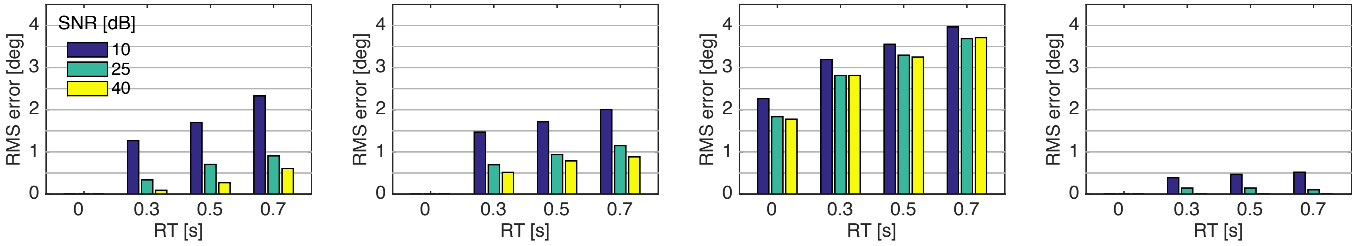
Figure 9 shows the number of found sources in Condition 2. For both PIV and PWD-SRP increasing reverberation has a strong effect on the ability to localize all the sources. SSPIV provides a clear improvement over both these methods while DPD-MUSIC is hardly affected by reverberation and only slightly by noise.

Figure 10 shows the RMS error in the estimated DOAs for those trials in which all four sources were found. Since for PIV with RT 0.7 s there were relatively few, if any, trials in which this condition was satisfied those results should be disregarded. In general, the error is substantially more than in Condition 1 (apart from when there is no reverberation, where again only PWD-SRP is less than perfect) and the general trend for the error to increase with RT is as expected. The SSPIV method achieves less than half the error of PWD-SRP and a substantial (1.8-3.8°) improvement over PIV. Even under the most challenging conditions tested (RT 0.7 s, SNR 10 dB) the RMS error is only 4.8°. Compared to DPD-MUSIC the increase in error for SSPIV under reverberant conditions is between 0.8° (RT 0.3 s, SNR 25 dB) and 1.9° (RT 0.7 s, SNR 40 dB).

Figure 11 shows the number of found sources as the duration of the observed signals increases for RT: 0.5 s and SNR: 25 dB. With 2 s of data, DPD-MUSIC is able to identify all the sources. Whilst longer observation intervals continue to improve the performance of both PIV and SSPIV, the performance of PWD-SRP plateaus after 2 s. With only 250 ms of data, none of the methods consistently finds all the sources. Condition 3 addresses the trade-off between the accuracy of DOA estimation, the probability of identifying each source and the number of erroneous estimates.

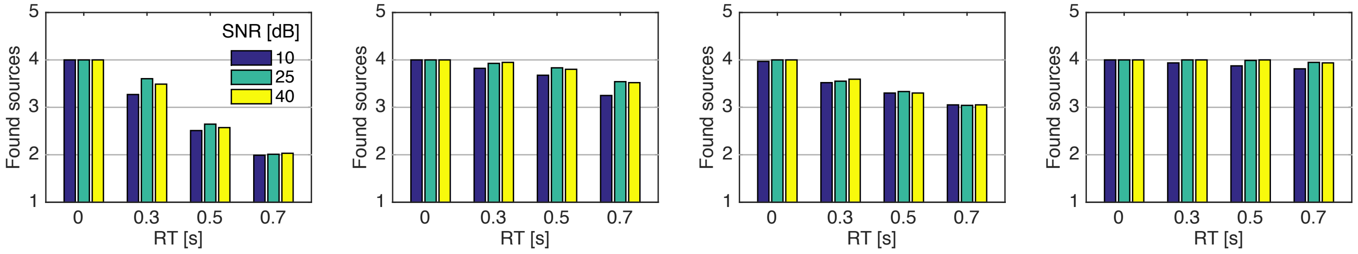
Figure 12 shows the estimated azimuth angles obtained in a representative trial in Condition 3. The symbols convey the assignment of each estimate to a particular source or classification as clutter based on knowledge of the ground truth. Qualitatively, it can be observed that SSPIV has a lower error and fewer missed detections than PIV. DPD-MUSIC has low error but a high amount of clutter. This is caused by DPD-MUSIC only computing the spatial spectrum over those TF-regions where one direct path is dominant (51). If a particular source was not dominant in any TF-regions during the observation interval no peak will exist at the corresponding position in the spatial spectrum. Since, in this example, the 3 largest peaks are always selected, regardless of their amplitude, one or more of those peaks can be erroneous.

For all methods the amount of clutter can be reduced by discarding peaks which fall below a threshold. However, this risks discarding weak observations which are in fact accurate. Figure 13(a) shows the relationship between RMS error and miss rate due to varying the threshold. As expected, for all methods, when only the largest peaks are retained the error is reduced but at the expense of more misses. Consistent with the results in Condition 2, SSPIV is more accurate than PIV, achieving approximately 4° less error for a given miss rate and DPD-MUSIC achieves the lowest RMS error. However, the relationship between clutter rate and miss rate, shown in Fig. 13(b), suggests that, for a particular miss rate, SSPIV achieves the lowest clutter rate. This is especially apparent for miss rates between 0.25 and 0.5 where DPD-MUSIC averages 0.7-2.3 clutter measurements per time step whereas SSPIV



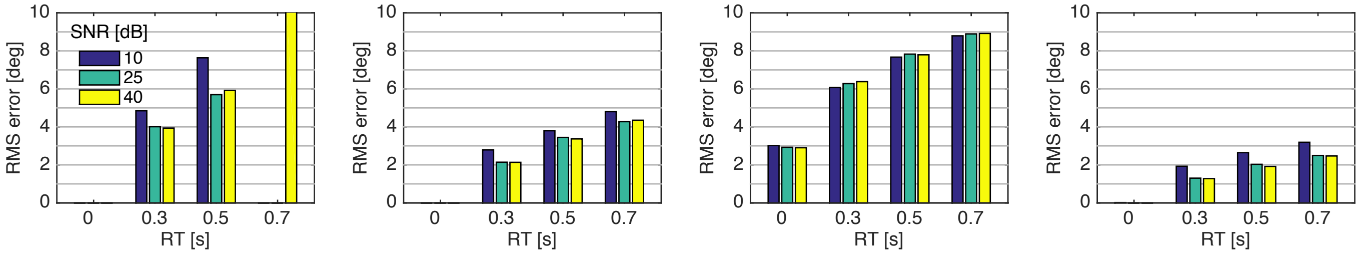
(a) (b) (c) (d)

Fig. 8. Effect of RT and SNR on RMS DOA estimation error for (a) PIV, (b) SSPIV, (c) PWD-SRP and (d) DPD-MUSIC for Condition 1 (one source).



(a) (b) (c) (d)

Fig. 9. Effect of RT and SNR on number of found sources for (a) PIV, (b) SSPIV, (c) PWD-SRP and (d) DPD-MUSIC for Condition 2 (maximum of 4).



(a) (b) (c) (d)

Fig. 10. Effect of RT and SNR on RMS DOA estimation error for (a) PIV, (b) SSPIV, (c) PWD-SRP and (d) DPD-MUSIC for Condition 2 (four sources).

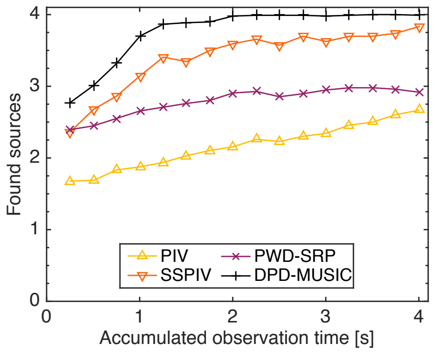
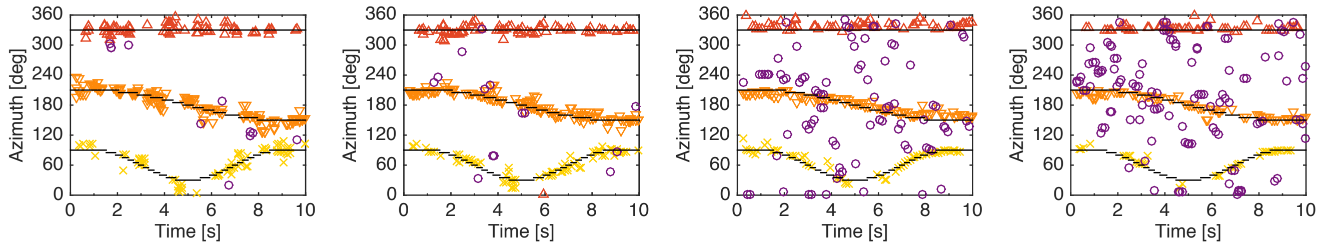


Fig. 11. Average number of found sources as a function of length of observation for Condition 2 with RT: 0.5 s and SNR: 25 dB.

averages less than 0.3. These results suggest that SSPIV is better suited to applications where clutter measurements and missed detections are more damaging than a small increase in absolute error.

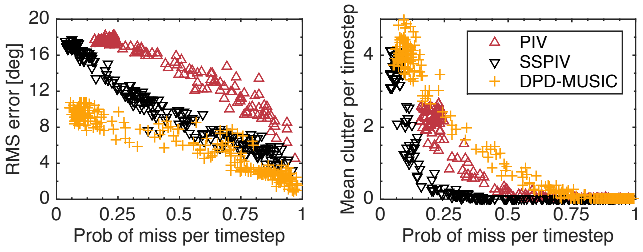
G. Computational cost

The computational costs of the proposed methods were compared to the baseline algorithms for a scenario with two simultaneous sources for different grid resolutions. The computational cost of the algorithms is evaluated in terms of their real-time factor (i.e. the ratio of elapsed time for the computation to the duration of the signal) as implemented in Matlab running on a general purpose computer (dual core Intel Core i5 processor, 2.6 GHz clock speed, 8 GB RAM). This metric does not include the precomputation time for signal independent variables. In our implementation the PWD-SRP and DPD-MUSIC algorithms use precomputed steering vectors while the PIV and SSPIV methods use precomputed dictionary elements. These took $\{0.0073,\ 0.0122,\ 0.0726,\ 0.2954\}$ s and



(a) (b) (c) (d)

Fig. 12. First 3 estimated azimuth angles as a function of time for one representative trial in Condition 3 using (a) PIV, (b) SSPIV, (c) PWD-SRP and (d) DPD-MUSIC. Lines show ground truth trajectories. Symbols represent source assignment used to calculate metrics (△: source 1, ▽: source 2, ×: source 3, ○: clutter).



(a) (b)

Fig. 13. The effect of varying the estimate acceptance threshold on the relationship between (a) RMS error and miss rate, and (b) clutter rate and miss rate for each algorithm (△: PIV, ▽: SSPIV, +: DPD-MUSIC).

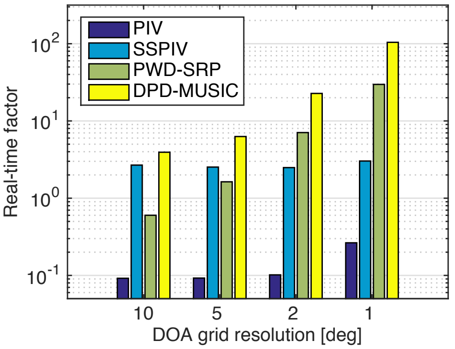


Fig. 14. Real-time factors of the compared algorithms for different angular resolutions.

$\{0.0040, 0.0181, 0.3051, 2.9103\}$ s, respectively, to compute for grid resolutions $\{10^\circ, 5^\circ, 2^\circ, 1^\circ\}$.

The results are shown in Fig. 14. From (22) and (27), the computational cost of calculating PIVs and SSPIVs does not depend on the resolution of the DOA analysis. However, the grid density determines the number of directions for which the summation in (54) must be performed. This weak dependence on grid resolution is noticeable in Fig. 14 for PIV, because the calculation of the PIVs themselves is very fast (10× faster than real-time), but not for SSPIV, where the computation time is dominated by (49) and (26). In contrast, both PWD-SRP and DPD-MUSIC have rapidly increasing computational cost as the grid density is increased because all directions in the grid are evaluated for every TF-bin. With a 2° resolution, as used in the reported performance evaluation, SSPIVs is an order of magnitude faster than DPD-MUSIC.

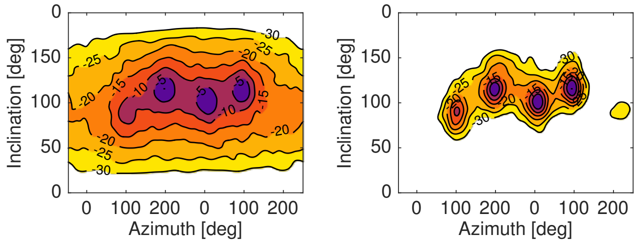
H. Discussion

The results for Condition 1 and the clear computational advantages suggest that for single source DOA estimation the PIV method is preferable to the other methods considered. However, Condition 2 demonstrates that when multiple talkers are simultaneously active subspace methods offer substantial improvements. By using the noise subspace and only considering TF-regions with a single dominant source, DPD-MUSIC achieves slightly better accuracy than SSPIV. However, Condition 3 shows that when short observation intervals are used, as is required for moving sources, this selectivity comes at the price of a higher proportion of clutter estimates for a given miss rate. Since SSPIV also requires significantly less computation than DPD-MUSIC at dense grid resolutions, it is particularly well suited to DOA estimation in situations involving multiple, moving speakers.

VI. Experimental verification

To demonstrate the efficacy of the proposed methods, speech was recorded in a real room with dimensions of approximately 10.3×9.2×2.6 m and a reverberation time of 0.4 s. Speech signals were recorded using an Eigenmike 32 channel rigid spherical microphone array with radius 4.2 cm located close to the centre of the room. In the first scenario four talkers were simultaneously active. These were arranged at approximately 60° intervals and their inclinations alternated to be above or below the horizontal plane of the array, according to whether they were seated or standing. In each case the projection of the source distance in the horizontal plane was approximately 1.5 m. Figure 15 shows smoothed histograms for a single 4 s observation interval of PIVs and SSPIVs. Peaks corresponding to the four sources are present in both cases but the definition of the peaks is much more distinct for SSPIV.

DOA estimates were calculated every 0.1 s using a sliding 4-second observation window of PIVs and SSPIVs. For each observation a smoothed histogram was computed and the largest four peaks selected as the DOA estimates. The azimuth angle estimates for a representative extract of the signal are



(a) (b)

Fig. 15. Smoothed histogram of (a) PIVs and (b) SSPIVs for observation interval at 14 s in experiment 1. Contours indicate histogram values in dB with respect to maximum value. Colormap is the same in both plots.

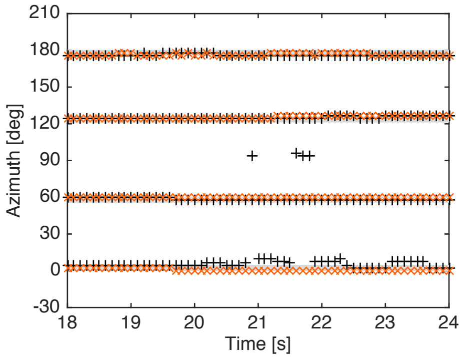
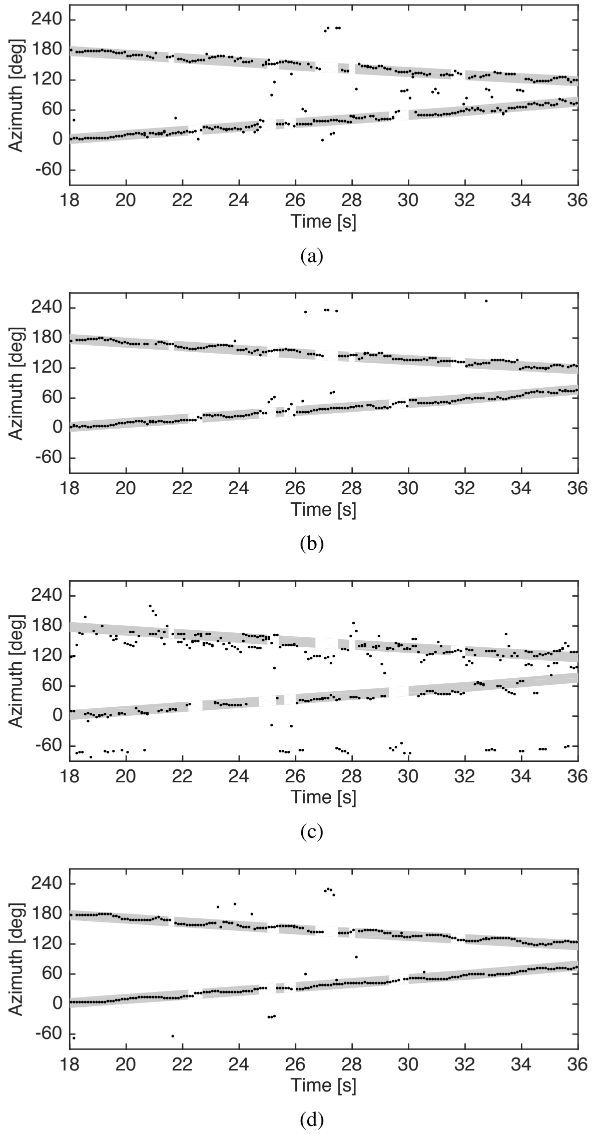


Fig. 16. Estimated azimuth of arrival as a function of time for (+) PIV and (×) SSPIV using a 4 second observation interval. Grey lines indicate ground truth azimuth.

shown in Fig. 16. The ground truth DOAs were measured with an estimated uncertainty of ±2.5° and are shown as thick grey lines. The DOA estimates produced by both the proposed methods coincide with the ground truth, but it can be seen that the SSPIV method has less variability and is free from erroneous estimates.

So as to be relevant to practical scenarios with moving sound sources, in the second scenario, two sources were recorded whilst moving around a radius of 1.5 m. In this case, in order to resolve the position, a sliding snapshot of 250 ms was used. Therefore each histogram was constructed from $^{1}/_{16}$ of the data points compared to the first scenario. The resulting azimuth estimates as a function of time are shown in Fig. 17. The estimated error in the ground truth is ±10°. Compared to the static case, the estimates are clearly more noisy but generally follow the ground truth trajectories. SSPIV and DPD-MUSIC clearly have fewer outlying estimates than PIV or PWD-SRP and in most cases these can be attributed to short pauses in the speech. Since it is assumed that both sources are active at all times, a peak in the histogram or spatial spectrum due to noise will yield an erroneous DOA. For all methods apart from PWD-SRP there is a clearly visible oscillatory component to the estimated trajectories. The ground truth indicates the overall trajectory but video analysis of the recordings reveals that the talkers' heads followed an oscillatory motion due to an inverted pendulum effect as they side-stepped radially around the microphone array. This suggests that in practical situations both the proposed methods would be suitable for tracking detailed source movements with the SSPIV method producing fewer outlier estimates.



(a)

(b)

(c)

(d)

Fig. 17. Estimated DOAs as a function of time for (a) PIV, (b) SSPIV (c) PWD-SRP and (d) DPD-MUSIC using a 0.25 s observation interval for two moving sources. Grey lines indicate ground truth trajectory and voice activity.

VII. Conclusions

Two intensity-based methods of DOA estimation operating in the SH domain have been presented and compared. The PIV method was shown to be computationally efficient and, in simulated experiments with a single source, was accurate to within about 1° across a range of SNRs (25-40 dB) and RTs (≤0.7 s) and to within 2.5° in the most challenging case tested (SNR 10 dB, RT 0.7 s).

The SSPIV method exploits frequency smoothing followed by subspace decomposition of the spatial covariance matrix. As a result it demonstrated better robustness to interfering sound sources, coherent reflections, diffuse noise and sensor noise than PIV. The SSPIV method is an order of magnitude faster than DPD-MUSIC, yet is only slightly less accurate. In simulated experiments with four simultaneously active sources, the mean angular error in the most challenging condition (SNR 10 dB, RT 0.7 s) was 4.8° for SSPIV, compared to 3.2° for DPD-MUSIC and 8.8° for PWD-SRP. Furthermore, it was shown that for moving sources SSPIV offered a better trade-off between the number of clutter measurements and the number of missed detections. This suggests the SSPIV method is particularly suitable for tracking applications such as in robot audition.

Finally a real world experiment demonstrated both PIVs and SSPIVs to be effective in practice for both stationary and moving sources.

References

[1] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 2002, pp. 1781–1784.

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

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

[4] ——, *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer, 2015.

[5] ——, "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.

[6] 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, Mar. 2005, pp. iii/89–iii/92.

[7] 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.

[8] 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.

[9] 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.

[10] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Am.*, vol. 131, pp. 2828–2840, 2012.

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

[12] 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, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[13] 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.

[14] 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)*, 2015.

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

[16] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, Aug. 1985.

[17] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Am.*, vol. 123, pp. 2136–2147, 2008.

[18] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[19] 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.

[20] ——, "Super-resolution sound field imaging with sub-space preprocessing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 350–354.

[21] 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.

[22] ——, "Super-resolution acoustic imaging using sparse recovery with spatial priming," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2414–2418.

[23] F. Jacobsen and H.-E. de Bree, "A comparison of two different sound intensity measurement principles," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1510–1517, 2005.

[24] H.-E. de Bree, P. Leussink, T. Korthorst, H. Jansen, T. S. Lammerink, and M. Elwenspoek, "The μ-flown: a novel device for measuring acoustic flows," vol. 54, no. 1-3. Elsevier, 1996, pp. 552–557.

[25] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, Sep. 1994.

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

[27] 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.

[28] 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 Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[29] A. H. Moore, C. Evers, and P. A. Naylor, "2D direction of arrival estimation of multiple moving sources using a spherical microphone array," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2016, (submitted).

[30] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999.

[31] 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.

[32] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 193–204, 2014.

[33] 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.

[34] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology (NIST), NIST Interagency/Internal Report (NISTIR) 4930, Feb. 1993.

[36] *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[37] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 2002, pp. 529–532.

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

Alastair 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, modelling and characterisation of room acoustics, dereverberation and spatial audio perception with applications for robot audition and hearing aids.

Christine Evers (M'14, SM'16) received her PhD from the University of Edinburgh, UK, in 2010, after having completed her MSc degree in Signal Processing and Communications at the University of Edinburgh in 2006, and BSc degree in Electrical Engineering and Computer Science at Jacobs University Bremen, Germany in 2005. After a position as a research fellow at the University of Edinburgh between 2009 and 2010, she worked as a senior systems engineer at Selex ES, Edinburgh, UK, between 2010 and 2014. Since 2014, she is a research associate in the Department of Electrical and Electronic Engineering at Imperial College London. Her research focuses on statistical signal processing for speech and audio applications, including sound source localization and tracking, acoustic simultaneous localization and mapping for robot audition, blind speech dereverberation, and sensor fusion. She is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.

Patrick Naylor (M'89, SM'07) received his BEng degree in Electronic and Electrical Engineering from the University of Sheffield, U.K., in 1986 and the PhD. degree from Imperial College, London, U.K., in 1990. Since 1990 he has been a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. His research interests are in the areas of speech, audio and acoustic signal processing. He has worked in particular on adaptive signal processing for 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 mainland Europe. He is the Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, a director of the European Association for Signal Processing (EURASIP) and formerly an associate editor of IEEE Signal Processing Letters and IEEE Transactions on Audio Speech and Language Processing.