# Focusing and Frequency Smoothing for Arbitrary Arrays With Application to Speaker Localization

Hanan Beit-On and Boaz Rafaely, *Senior Member, IEEE*

***Abstract*—The coherent signal subspace method (CSSM) enables the direction-of-arrival (DoA) estimation of coherent sources with subspace localization methods. The focusing process that aligns the signal subspaces within a frequency band to its central frequency is central to the CSSM. Within current focusing approaches, a direction-independent focusing approach may be more suitable for reverberant environments since no initial estimation of the sources' DoAs is required. However, these methods use integrals over the steering function, and cannot be directly applied to arrays around complex scattering structures, such as robot heads. In this article, current direction-independent focusing methods are extended to arrays for which the steering function is available only for selected directions, typically in a numerical form. Spherical harmonics decomposition of the steering function is then employed to formulate several aspects of the focusing error. A case of two coherent sources is studied and guidelines for the selection of the frequency smoothing bandwidth are suggested. The performance of the proposed methods is then investigated for an array that is mounted on a robot head. The focusing process is integrated within the direct-path dominance (DPD) test method for speaker localization, originally designed for spherical arrays, extending its application to arrays with arbitrary configurations. Finally, experiments with real data verify the feasibility of the proposed method to successfully estimate the DoAs of multiple speakers under real-world conditions.**

***Index Terms*—Direct-path, direct-path dominance test, direction-of-arrival estimation, focusing, frequency smoothing, room reverberation, speaker localization, spherical harmonics.**

## I. Introduction

DIRECTION-OF-ARRIVAL (DoA) estimation is an important and timely challenge in audio signal processing with applications in acoustic scene analysis, signal enhancement, and speech processing [1], [2]. DoA estimation is often required to operate in reverberant and noisy environments; thus, there is a great need for robustness to reverberation and to noise. Furthermore, computational efficiency may be required for adaptive source tracking. A popular approach for DoA estimation is based on the MUSIC (MUltiple SIgnal Classification) algorithm [3], which is a high resolution estimator with relatively low computational burden. However, it fails in the presence of coherent sources, such as reflections in a reverberant environment, due to the deficient rank of the cross-spectrum matrix.

The coherent signal subspace method (CSSM) [4] enables the application of subspace localization methods such as MUSIC to coherent sources. This CSSM involves a focusing process followed by frequency smoothing. Different focusing approaches have been proposed, as summarized in [5]. Most focusing approaches rely on an initial estimate of the DoAs, leading to an iterative DoA estimation process. However, in addition to the computational cost associated with this iterative process, it may not be possible to obtain an initial estimate in reverberant environments with multiple coherent sources due to room reflections.

In contrast to the common focusing approach, the focusing methods proposed in [5]–[7] are direction-independent, and thus do not require initial DoA estimates nor an iterative process. This is achieved by formulating focusing matrices that minimize the mean square focusing error over all directions. These methods can be applied to arbitrary arrays and use integrals over the steering functions, assuming that an analytical expression of the steering function is available, e.g. for sensors in free-field. Therefore, these methods cannot be directly applied to arrays for which the steering function is available only for selected directions, and in a numerical form, such as arrays around complex scattering structures, e.g. robot head.

In this article, the current methods for direction-independent focusing are extended to arrays for which the steering function is available only for selected directions and in a numerical form. This is done by reformulating the current focusing transformations using a spherical harmonics coefficients matrix, which can be computed numerically without integral approximation. The factors affecting focusing error are studied, leading to conditions on the required number of directions, which may depend on frequency. Since the performance of the direction-independent focusing approach may strongly depend on the smoothing bandwidth, this is studied for the simple case of two coherent sources. The results indicate that a bandwidth of about 500 Hz could be appropriate under typical reverberant conditions. Focusing performance for this bandwidth is then investigated for an array that is mounted on a Nao robot head [8].

The proposed direction-independent focusing approach was then integrated with the direct-path dominance (DPD) test method [9] for speaker localization, leading to its generalization from spherical arrays to arbitrary arrays. This extension of the DPD test was previously described in [10], [11], where it was applied to a binaural array. The novelty of this article beyond [10], [11] is the analysis of smoothing bandwidth selection,

Manuscript received November 20, 2019; revised April 26, 2020; accepted July 7, 2020. Date of publication July 17, 2020; date of current version August 3, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Timo Gerkmann. *(Corresponding author: Hanan Beit-On.)*

The authors are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: hananbo26@gmail.com; br@bgu.ac.il).

Digital Object Identifier 10.1109/TASLP.2020.3010098

as well as the theoretical derivation of the focusing matrices and investigation of the factors affecting the focusing error. These additional contributions are crucial when applying the method for arbitrary arrays. Currently, alternative approaches for DoA estimation with arbitrary arrays in reverberant environments do exist [12]–[14]. However, in contrast to the speaker localization method proposed in [12], the proposed method based on the DPD test does not require speech-free time segments for estimating the noise statistics and can operate on short time segments. Also, unlike the binaural localization method proposed in [13], the proposed approach does not require training. Real recordings with an array mounted on a Nao robot head as part of the LOCalization And TrAcking (LOCATA) challenge [15], [16] were employed to evaluate and compare the performance of the proposed approach and the DPD test proposed in [14]. This method was chosen for comparison because it is also based on the DPD approach and can be applied to arbitrary arrays. However, [14] is strictly used for DoA estimation, while the proposed method can be used to compute smoothed cross-spectrum matrices for other applications, for example, speech enhancement [17], blind source separation [18], and beamforming [19]. The proposed method demonstrated performance comparable with the DPD test proposed in [14]. The results verify the feasibility of the proposed extension of the DPD test under real-world conditions. Moreover, the good DoA estimation performance implies that the proposed focusing and smoothing process works well and can be used for applications other than DoA.

## II. Array Model and Frequency Smoothing

Consider an array with $I$ omnidirectional microphones arranged in an arbitrary configuration. The position of the $i$-th microphone in a Cartesian coordinate system is $\boldsymbol{r}_i = (r_i \cos\phi_i \sin\theta_i, r_i \sin\phi_i \sin\theta_i, r_i \cos\theta_i)^T$. The elevation angle $\theta$ is measured downward from the z-axis, the azimuth angle $\phi$ is measured counterclockwise from the positive part of the x-axis and $r_i$ is the distance from the origin to the $i$-th microphone. Additionally, the sound field is assumed to be composed of $L$ plane waves originating from $L$ far-field sound sources, where a source can represent a direct-sound or, for example, a reflection due to room boundaries. Let $\psi_l = (\theta_l, \phi_l)$ and $\mathbf{k}_l = -(k \cos\phi_l \sin\theta_l, k \sin\phi_l \sin\theta_l, k \cos\theta_l)^T$ denote the DoA and the wave vector of the $l$-th plane wave, respectively. For a homogeneous medium the linear relation $\|\mathbf{k}\| = k = \frac{2\pi f}{c}$ holds, where $k$ is the wave number, $f$ is the frequency and $c$ is the speed of sound. We employ the multiplicative transfer function (MTF) approximation [20] that assumes that the time windows are sufficiently large with respect to the length of the microphone array's steering function (in time), such that the sound pressure measured by the microphone array can be modeled in the STFT (short-time Fourier transform) domain as

$$\mathbf{p}(\tau,\nu) = \mathbf{H}(\nu,\boldsymbol{\psi})\,\mathbf{s}(\tau,\nu) + \mathbf{n}(\tau,\nu), \tag{1}$$

where $\tau$ and $\nu$ are the time frame and the frequency indices, respectively. $\mathbf{p}(\tau,\nu) = [p_1(\tau,\nu),\ldots,p_I(\tau,\nu)]^T$, with $p_i(\tau,\nu)$ denoting the STFT of the received signal at the $i$-th microphone. $\mathbf{H}(\nu,\boldsymbol{\psi}) = [\mathbf{h}(\nu,\psi_1),\ldots,\mathbf{h}(\nu,\psi_L)]^T$ is an $I \times L$ steering matrix, where $\mathbf{h}(\nu,\psi_l) = [h_1(\nu,\psi_l),\ldots,h_I(\nu,\psi_l)]^T$ is the array's steering vector that corresponds to the $l$-th source with DoA $\psi_l$, such that $\boldsymbol{\psi} = [\psi_1,\ldots,\psi_L]^T$ is the DoA vector. $\mathbf{s}(\tau,\nu) = [s_1(\tau,\nu),\ldots,s_L(\tau,\nu)]^T$ is the source signals vector where $s_l(\tau,\nu)$ is the STFT of the $l$-th source signal denoting its complex amplitude at the origin. $\mathbf{n}(\tau,\nu) = [n_1(\tau,\nu),\ldots,n_I(\tau,\nu)]^T$ is the additive noise vector.

For coherent sources, such as a sound source and its reflections, the cross-correlation matrix of the source signals $\mathbf{S_s}(\tau,\nu)$ is singular, which may lead to the failure of the MUSIC algorithm [4]. The CSSM, which involves a focusing process followed by frequency smoothing, was proposed to enable the operation of subspace localization methods, such as MUSIC, for coherent sources. The focusing process is performed by multiplying the array signal $\boldsymbol{p}(\tau,\nu)$ at each frequency within a frequency band by a corresponding $I \times I$ focusing matrix $\mathbf{T}(\nu,\nu_0)$ that aligns the steering matrix $\mathbf{H}(\nu,\boldsymbol{\psi})$ to the central steering matrix $\mathbf{H}(\nu_0,\boldsymbol{\psi})$, where $\nu_0$ denotes the central frequency within the processed frequency band. The focusing matrix $\mathbf{T}(\nu,\nu_0)$ satisfies

$$\mathbf{T}(\nu,\nu_0)\,\mathbf{H}(\nu,\boldsymbol{\psi}) = \mathbf{H}(\nu_0,\boldsymbol{\psi}). \tag{2}$$

Defining the transformed array signal $\tilde{\mathbf{p}}(\tau,\nu)$ as

$$\tilde{\mathbf{p}}(\tau,\nu) = \mathbf{T}(\nu,\nu_0)\,\tilde{\mathbf{p}}(\tau,\nu), \tag{3}$$

and assuming $\mathbf{s}(\tau,\nu)$ and $\mathbf{n}(\tau,\nu)$ are uncorrelated, the cross-spectrum matrix of $\tilde{\mathbf{p}}(\tau,\nu)$ can be written as

$$\begin{aligned}\mathbf{S}_{\tilde{\mathbf{p}}}(\tau,\nu) &= E\left[\tilde{\mathbf{p}}(\tau,\nu)\,\tilde{\mathbf{p}}^H(\tau,\nu)\right] \\ &= \mathbf{H}(\nu_0,\boldsymbol{\psi})\,\mathbf{S_s}(\tau,\nu)\mathbf{H}^H(\nu_0,\boldsymbol{\psi}) \\ &\quad + \mathbf{T}(\nu,\nu_0)\,\mathbf{S_n}(\tau,\nu)\mathbf{T}^H(\nu,\nu_0),\end{aligned} \tag{4}$$

where $E[\cdot]$ is the expectation operation, $(\cdot)^H$ denotes the conjugate transpose, and $\mathbf{S_s}(\tau,\nu)$ and $\mathbf{S_n}(\tau,\nu)$ are the cross-correlation matrices of the source and noise signals, respectively. This equation shows the effect of the focusing operation. It removes the frequency dependence such that each transformed array signal shares the same steering matrix. This property enables the operation of frequency smoothing while preserving the spatial information in the smoothed steering matrices. Frequency smoothing is then applied, averaging the aligned cross-spectrum matrices from the frequency band to which focusing has been applied. Denoting the averaging operation with an over-line, the smoothed cross-spectrum matrix $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau,\nu_0)$ is

$$\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau,\nu_0) = \mathbf{H}(\nu_0,\boldsymbol{\psi})\,\overline{\mathbf{S_s}}(\tau,\nu_0)\,\mathbf{H}^H(\nu_0,\boldsymbol{\psi}) + \overline{\mathbf{S}_{\tilde{\mathbf{n}}}}(\tau,\nu_0), \tag{5}$$

where $\mathbf{S}_{\tilde{\mathbf{n}}}(\tau,\nu_0) = \mathbf{T}(\nu,\nu_0)\mathbf{S_n}(\tau,\nu)\mathbf{T}^H(\nu,\nu_0)$. The frequency smoothing of $\mathbf{S_s}(\tau,\nu)$ aims to decorrelate the sources and to restore its rank such that $\overline{\mathbf{S_s}}(\tau,\nu)$ is of full rank [4]. Assuming that the smoothing operation successfully decorrelated the sources, and given that there are fewer sources than sensors, i.e. $L < I$, MUSIC can be applied with the smoothed cross-spectrum matrix $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau,\nu_0)$, leading to the DoA estimation of the $L$ sources.

## III. Current Approaches to Focusing

Several formulations for the focusing matrices have been previously proposed. One type of these approaches assumes knowledge of the DoAs. In the absence of such knowledge, an iterative scheme is typically applied, where in each step the focusing matrices are computed using the DoAs estimated in the previous step, as described in [4]. This process relies on an initial estimate of the DoAs. However, in reverberant environments with multiple coherent sources due to room reflections, an initial estimate of the DoAs may be difficult to obtain. To avoid the computational cost associated with this iterative process and the possible errors due to initial DoA estimation errors, different focusing approaches based on direction-independent focusing matrices were proposed. One approach is tailored to spherical arrays, and employs plane-wave decomposition (PWD), which can be viewed as the application of direction-independent focusing and completely removes the frequency dependence of the steering matrices [21]. Another approach is based on beamspace processing and is referred to as beamspace–invariance (BI). In this formulation, the beamspace matrices play the role of focusing matrices and are obtained by minimizing the focusing error over all directions [6]:

$$\mathbf{T}_{BI}\left(\nu,\nu_0\right)=\arg\left\{\min_{\mathbf{T}}\int_{S^2}\left\|\mathbf{T}\mathbf{h}\left(\nu,\psi\right)-\mathbf{T}_0\mathbf{h}\left(\nu_0,\psi\right)\right\|^2 w\left(\psi\right)d\psi\right\}, \quad (6)$$

where $w(\psi)$ is a generic weighting function, $S^2$ is a spherical surface of unit radius and $\mathbf{T}_0$ is a reference focusing matrix, which can be designed to achieve desired properties such as high SNR gain or low sidelobes, and to reduce focusing error. The solution to (6) is given by [6]

$$\mathbf{T}_{BI}\left(\nu,\nu_0\right)=\mathbf{T}_0\mathbf{Q}^H\left(\nu,\nu_0\right)\mathbf{R}^{-H}\left(\nu\right), \quad (7)$$

where

$$\mathbf{Q}\left(\nu,\nu_0\right)=\int_{S^2}\mathbf{h}\left(\nu,\psi\right)\mathbf{h}^H\left(\nu_0,\psi\right)w\left(\psi\right)d\psi \quad (8)$$

and

$$\mathbf{R}\left(\nu\right)=\int_{S^2}\mathbf{h}\left(\nu,\psi\right)\mathbf{h}^H\left(\nu,\psi\right)w\left(\psi\right)d\psi. \quad (9)$$

Another formulation, referred to here as unitary focusing, is proposed at the first stage of the iterative scheme presented in [5]. Unitary focusing matrices, originally formulated in [22], ensure that there is no SNR loss in the focusing process [22]. In [5], the unitary focusing matrices are reformulated as direction-independent matrices, similarly to the BI transformations:

$$\mathbf{T}_{\text{unitary}}\left(\nu,\nu_0\right)=\arg\left\{\min_{\mathbf{T}}\int_{S^2}\left\|\mathbf{T}\mathbf{h}\left(\nu,\psi\right)-\mathbf{h}\left(\nu_0,\psi\right)\right\|^2 \times w\left(\psi\right)d\psi \text{ subject to } \mathbf{T}^H\mathbf{T}=\mathbf{I}\right\}. \quad (10)$$

The solution to (10) is given by [5]

$$\mathbf{T}_{\text{unitary}}\left(\nu,\nu_0\right)=\mathbf{W}\left(\nu,\nu_0\right)\mathbf{U}^H\left(\nu,\nu_0\right), \quad (11)$$

where $\mathbf{W}(\nu,\nu_0)$ and $\mathbf{U}(\nu,\nu_0)$ are the matrices obtained from the singular value decomposition of $\mathbf{Q}(\nu,\nu_0)$.

Another focusing formulation, referred to as the WINGS transformations [7], is based on the representation of the steering function using spherical harmonics. Assuming that the steering function is order-limited, i.e. it can be represented by spherical harmonics coefficients up to a certain order $N$, its representation using spherical harmonics is given by [2]

$$\mathbf{h}\left(\nu,\psi\right)=\mathbf{V}\left(\nu\right)\mathbf{y}\left(\psi\right), \quad (12)$$

where $\mathbf{y}(\psi)=[Y_0^0(\psi),Y_1^{-1}(\psi),\ldots,Y_N^N(\psi)]^T$ is an $(N+1)^2\times 1$ vector of the spherical harmonics functions $Y_n^m(\psi)$ of order $n$ and degree $m$ and

$$\mathbf{V}\left(\nu\right)=\begin{bmatrix} v_{0,0}^1\left(\nu\right) & v_{1,-1}^1\left(\nu\right) & \cdots & v_{N,N}^1\left(\nu\right)\\ v_{0,0}^2\left(\nu\right) & v_{1,-1}^2\left(\nu\right) & \cdots & v_{N,N}^2\left(\nu\right)\\ \vdots & \vdots & \ddots & \vdots\\ v_{0,0}^I\left(\nu\right) & v_{1,-1}^I\left(\nu\right) & \cdots & v_{N,N}^I\left(\nu\right)\end{bmatrix} \quad (13)$$

is an $I\times(N+1)^2$ matrix of the spherical harmonics coefficients $v_{n,m}^i(\nu)$ of order $n$ and degree $m$ of the $i$-th microphone's steering function. For an order-limited steering function, the solution to (6), with constant weights $w(\psi)=1$, is given by [7]

$$\mathbf{T}_{\text{WINGS}}\left(\nu,\nu_0\right)=\mathbf{T}_0\boldsymbol{V}\left(\nu_0\right)\boldsymbol{V}^{\dagger}\left(\nu\right), \quad (14)$$

where $(\cdot)^{\dagger}$ is the pseudo-inverse operation. The matrix $\mathbf{V}(\nu)$ is computed via the spherical Fourier transform of the array's steering function

$$\mathbf{V}\left(\nu\right)=\int_{S^2}\mathbf{h}\left(\nu,\psi\right)\mathbf{y}^H\left(\psi\right)d\psi. \quad (15)$$

The above methods of direction-independent focusing use integrals over the steering functions $\mathbf{h}(\nu,\psi)$, that is, assuming that an analytical expression of the steering function is available, e.g. for sensors in free field. Therefore, these methods cannot be directly applied to arrays for which the steering function is available only for selected directions, and in a numerical form, such as arrays around complex scattering structures, e.g. a robot head.

## IV. Proposed Focusing Method

To apply the current approaches of direction-independent focusing to arrays for which the steering function is available only for selected directions, the integrals in (8), and (15) can be approximated using quadrature methods. A quadrature method aims to approximate the integral given a set of samples on the sphere, $\{\psi_l\}_{l=1}^L$, and sampling weights, $\alpha_l$, as follows [2]:

$$\int_{S^2}g\left(\psi\right)d\psi\approx\sum_{l=1}^{L}\alpha_l g\left(\psi_l\right). \quad (16)$$

The explicit expressions of the weights for some standard sampling schemes are given in [2]. It is shown that the number of samples depends on the spherical harmonics order of the sampled function. This implies that an insufficient number of samples may lead to an inaccurate approximation, and consequently to poor focusing performance.

In this work, a set of factors affecting focusing error is derived, providing insight into the focusing process and supporting the application of accurate focusing. These factors are presented in the following section. In the remainder of this section, the discrete form for the computation of the focusing matrix is developed. While the standard sampling sets can be used to sample the directions of the steering function, in this work a more generalized form is employed in which no sampling weights are required, and for which the sampling set does not necessarily follow a predefined sampling scheme. Given samples of the array's steering function from a set of directions $\boldsymbol{\psi}$, and assuming an order-limited steering function (see next section for further discussion), the steering matrix $\mathbf{H}(\nu, \boldsymbol{\psi})$ can be represented using (12) as

$$\mathbf{H}\left(\nu, \boldsymbol{\psi}\right) = \mathbf{V}\left(\nu\right)\mathbf{Y}\left(\boldsymbol{\psi}\right), \tag{17}$$

where $\mathbf{Y}(\boldsymbol{\psi}) = [\mathbf{y}(\psi_1) \cdots \mathbf{y}(\psi_L)]$. The matrix $\mathbf{V}(\nu)$ can be computed by the least-squares (LS) solution to (17), as follows [2]:

$$\mathbf{V}\left(\nu\right) = \mathbf{H}\left(\nu, \boldsymbol{\psi}\right)\mathbf{Y}^{\dagger}\left(\boldsymbol{\psi}\right), \tag{18}$$

where $\mathbf{Y}^{\dagger}(\boldsymbol{\psi}) = \mathbf{Y}^{H}(\boldsymbol{\psi})(\mathbf{Y}(\boldsymbol{\psi})\mathbf{Y}^{H}(\boldsymbol{\psi}))^{-1}$ in this case. Equation (18) is known as the discrete spherical Fourier transform (DSFT) and it provides an alternative way to compute $\mathbf{V}(\nu)$ where the steering function is available only from selected directions. Having computed matrices $\mathbf{V}(\nu)$, $\mathbf{V}(\nu_0)$, focusing can be applied by $\mathbf{T}_{\mathrm{WINGS}}(\nu, \nu_0)$ as in (14). $\mathbf{T}_{\mathrm{unitary}}(\nu, \nu_0)$ can also be computed for this case. Assuming an order-limited steering function, the minimization problem (6) with $\mathbf{T}_0 = \mathbf{I}$, $w(\psi) = 1$, and the additional unitary constraint, that ensures that there is no SNR-loss in the focusing process, can be rewritten using Parseval's identity as [7]

$$\mathbf{T}_{\mathrm{unitary}}\left(\nu, \nu_0\right) = \arg\left\{\min_{\mathbf{T}} \left\|\mathbf{T}\mathbf{V}\left(\nu\right) - \mathbf{V}\left(\nu_0\right)\right\|_F^2 \quad \text{subject to } \mathbf{T}^H\mathbf{T} = \mathbf{I}\right\}, \tag{19}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The problem in (19) is referred to as the orthogonal procrustes problem with a solution given by [22]

$$\mathbf{T}_{\mathrm{unitary}}\left(\nu, \nu_0\right) = \mathbf{W}\left(\nu, \nu_0\right)\mathbf{U}^{H}\left(\nu, \nu_0\right), \tag{20}$$

where $\mathbf{W}(\nu, \nu_0)$ and $\mathbf{U}(\nu, \nu_0)$ are the matrices obtained from the singular value decomposition of

$$\mathbf{Q}\left(\nu, \nu_0\right) = \mathbf{V}\left(\nu\right)\mathbf{V}^{H}\left(\nu_0\right). \tag{21}$$

The representation of the focusing transformations using matrices in the spherical harmonics domain leads to the formulation of factors affecting focusing error. These factors are presented in the next section.

The proposed focusing and frequency smoothing process is summarized in Algorithm 1.

**Algorithm 1:** Focusing and Frequency Smoothing.

**Input:**
- array signal $\mathbf{p}(\tau, \nu)$ for $J_\tau$ and $J_\nu$ time and freq. samples
- steering function $\mathbf{h}(\nu, \psi_l)$ at $J_\nu$ freq. and $L$ directions

**Compute:**
- spherical harmonics coeff. matrix $\mathbf{V}(\nu)$, $\forall\nu$ [use (18)]
- focusing trans. $\mathbf{T}(\nu, \nu_0)$, $\forall\nu$ [use (14) or (20)]

**Focusing:**
- compute $\tilde{\mathbf{p}}(\tau, \nu)$, $\forall(\tau, \nu)$ [use (3)]

**Cross-spectrum estimation:**
- compute $\mathbf{S}_{\tilde{\mathbf{p}}}(\tau, \nu)$, $\forall\nu$ [use (4), $E[\cdot]$ as time averaging]

**Frequency smoothing:**
- compute $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau, \nu_0)$ [use (5)]

**Output**:
- smoothed cross-spectrum matrix $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau, \nu_0)$

## V. Factors Affecting Focusing Error

In this section the factors affecting focusing error when using the proposed method are investigated. Substituting (17) into (1), the system model can be written as

$$\mathbf{p}\left(\tau, \nu\right) = \mathbf{V}\left(\nu\right)\mathbf{Y}\left(\boldsymbol{\psi}\right)\mathbf{s}\left(\tau, \nu\right) + \mathbf{n}\left(\tau, \nu\right). \tag{22}$$

However, this model requires that the steering function is of finite spherical harmonics order. Now, applying focusing to (22) leads to the formulation of an ideal focusing matrix:

$$\mathbf{T}\left(\nu, \nu_0\right)\mathbf{V}\left(\nu\right) = \mathbf{V}\left(\nu_0\right). \tag{23}$$

A set of focusing matrices that satisfy (23) for each frequency within the processed frequency band leads to ideal focusing for any set of sources. In practice, however, several factors may lead to errors in the focusing process; three major types of error are outlined below.

*1) Error 1 - Spherical harmonics order truncation:* Array steering functions, representing the transfer function from a source to a microphone, are typically of infinite spherical harmonics order, leading to an inevitable error in the representation of the steering matrix using a finite spherical harmonics order, as in (22). In particular, $v^i_{n,m}(\nu)$ is proportional to a spherical Bessel function $j_n(kr_i)$, for arrays in free-field, and to a combination of spherical Bessel functions and spherical Henkel functions, $h_n(kr_i)$ for arrays around scattering objects [2]. These functions decay for $n > kr_i$, where $r_i$ is the distance of the $i$-th microphone from the origin and $k$ is the wave number. Therefore, the steering function can be considered order-limited in practice, with a relatively small error for $kr_i \approx N$, and with a diminishing error for $kr_i \ll N$ [2]. The truncation error can be defined for a row vector $\mathbf{v}(\nu)$ in $\mathbf{V}(\nu)$ (for a given microphone), and is given by

$$\epsilon_{\mathrm{trun}}\left(\nu\right) = \frac{\left\|\tilde{\mathbf{v}}\left(\nu\right) - \begin{bmatrix}\mathbf{v}\left(\nu\right)\\ \mathbf{0}\end{bmatrix}\right\|^2}{\left\|\tilde{\mathbf{v}}\left(\nu\right)\right\|^2}, \tag{24}$$

where vector $\tilde{\mathbf{v}}(\nu)$ is of order $\tilde{N}$, which is assumed to be sufficiently high to guarantee a negligible error.

*2) Error 2 - Matrix inversion:* Accurate focusing requires a perfect solution to (23). Using the WINGS transformations as in (14) with $\mathbf{T}_0 = \mathbf{I}$, this translates to the inversion of matrix $\mathbf{V}(\nu)$.

Therefore, focusing error may arise due to an error in matrix inversion. This normalized matrix inversion error is given by

$$\epsilon_{inv}(\nu) = \frac{\|\mathbf{T}(\nu,\nu_0)\,\mathbf{V}(\nu) - \mathbf{V}(\nu_0)\|_F^2}{\|\mathbf{V}(\nu_0)\|_F^2}. \quad (25)$$

This error is zero if matrix $\mathbf{V}(\nu)$ is perfectly invertible. Since matrix $\mathbf{V}(\nu)$ is of size $I \times (N+1)^2$, in case where there are more microphones than coefficients, i.e. $I > (N+1)^2$, the system of equations in (23) (solving separately for each row in $\mathbf{T}(\nu,\nu_0)$), is under-determined, and any of its infinite solutions will satisfy (23). In the case where there are more coefficients than microphones, i.e. $I < (N+1)^2$, the system in (23) is over-determined, and can be solved in the least-square sense using (14), potentially leading to some error. Note that this over-determined case can be avoided by limiting the order $N$ such that $I \geq (N+1)^2$. This, in turn, can be achieved in practice by limiting the wave number (by limiting the operating frequency), or by truncating the order at the expense of increasing the truncation error.

*3) Error 3 - Spatial aliasing:* When an analytical expression for the steering function is not available, matrix $\mathbf{V}(\nu)$ can be computed from spatial samples (source directions) of the steering function, as in (18). However, an insufficient number of spatial samples relative to the spherical harmonics order of the steering function may lead to aliasing errors [2]. Denoting the steering matrix coefficients computed using (18) by $\hat{\mathbf{V}}(\nu)$, and then substituting (17) into (18) leads to

$$\hat{\mathbf{V}}(\nu) = \mathbf{V}(\nu)\,\mathbf{Y}(\boldsymbol{\psi})\,\mathbf{Y}_{\hat{\mathbf{V}}}^{\dagger}(\boldsymbol{\psi}) = \mathbf{V}(\nu)\left[\mathbf{I}\ \mathbf{A}\right]^T. \quad (26)$$

$\mathbf{Y}_{\hat{\mathbf{V}}}^{\dagger}(\boldsymbol{\psi})$ holds the sampled steering directions, with order $N_{\hat{\mathbf{V}}}$ which is determined by the number and distribution of the samples [2]. $\mathbf{A}$ is an $[(\tilde{N}+1)^2 - (N_{\hat{\mathbf{V}}}+1)^2] \times (N_{\hat{\mathbf{V}}}+1)^2$ matrix which reflects the way in which high order harmonics are aliased to the low orders when $N_{\hat{\mathbf{V}}} < \tilde{N}$, where $\tilde{N}$ is the true order of $\mathbf{V}(\nu)$, as defined in Error 1 above. Note that in this case the sampling process will lead to additional truncation, from $\tilde{N}$ to $N_{\hat{\mathbf{V}}}$. Aliasing error may be unavoidable because $\tilde{N}$ may be infinite in practice. Nevertheless, aliasing error can be reduced by selecting a sampling scheme leading to $N_{\hat{\mathbf{V}}}$ that is sufficiently high to faithfully represent $\mathbf{V}(\nu)$. This requires that the number of samples satisfies $L \geq (\tilde{N}+1)^2$ and that $\mathbf{Y}(\boldsymbol{\psi})$ is of full rank with a reasonably low condition number so that matrix $\mathbf{Y}(\boldsymbol{\psi})\mathbf{Y}^H(\boldsymbol{\psi})$ in (18) has a stable inverse. The exact number of samples (or directions), $L$, may depend on the sampling schemes, see e.g. [2] for some standard schemes. Because $\tilde{N}$ increases with frequency (see discussion in Error 1 above), it is expected that the number of required samples will increase with frequency. The aliasing error can be formulated as

$$\epsilon_{\text{alias}}(\nu) = \frac{\left\|\tilde{\mathbf{v}}(\nu) - \begin{bmatrix}\hat{\mathbf{v}}(\nu)\\ \mathbf{0}\end{bmatrix}\right\|^2}{\|\tilde{\mathbf{v}}(\nu)\|^2}, \quad (27)$$

where $\hat{\mathbf{v}}(\nu)$ is a row vector in $\hat{\mathbf{V}}(\nu)$ (for a given microphone) and $\tilde{\mathbf{v}}(\nu)$ is of order $\tilde{N}$ assumed to be sufficiently high to guarantee a negligible error. Note that this aliasing error also includes truncation, because sampling with an insufficient number of directions causes both types of error.

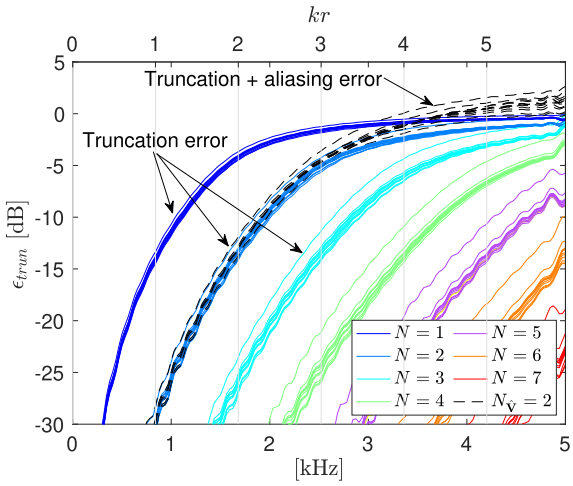


Fig. 1. Truncation error $\epsilon_{\text{trun}}(\nu)$ computed using (24) for orders $N$ from 1 to 7 (solid lines with different colors) and for all 12 microphones (different plots within the same color). The dashed lines denote aliasing (and added truncation) error, computed using (27) with $N_{\hat{\mathbf{V}}} = 2$ from a sampling set of $L = 9$ directions.

## VI. Focusing Analysis for a Robot Head

This section presents an analysis of focusing for a microphone array that is mounted on the Nao robot head, and investigates the factors that may lead to focusing error for this array example. The array is composed of 12 omnidirectional microphones arranged in a pseudo-spherical arrangement. This array was employed in the recent acoustic sources LOCalization And TrAcking (LOCATA) challenge and a detailed description of the array can be found in the challenge documents [23]. Simulated impulse responses from a source to each microphone were provided with a sampling frequency of $Fs = 10$ kHz from a set of $L = 240$ source directions that follow a nearly-uniform distribution. The impulse responses were up-sampled to a sampling frequency of $Fs = 16$ kHz and represented using $M = 512$ frequency points. Since the steering function is available only from a finite set of directions, the proposed method is employed to compute the focusing transformations. In the remainder of this section, the errors due the factors listed in Section V are evaluated, and next, overall focusing performance is studied.

### A. Spherical Harmonics Order Truncation

This section investigates the truncation error due to the representation of the array steering function by a finite order. The array is composed of microphones mounted around Nao's head, which act as a scattering object. Therefore, as discussed in the previous section, the steering function can be considered order-limited for $kr_i \approx N$, with a diminishing error for $kr_i \ll N$, for some finite $N$. For this array, $r_i$, the distance of the $i$-th microphone from the origin (center of the head) is fairly uniform and reaches about $r \approx 6.5$ cm for the furthest microphone. This leads to $kr = 6$ for 5 kHz, such that $N = 6$ chosen in the operating frequency range. Therefore, $\tilde{N} = 10$ was chosen as an accurate reference in the computation of the steering function.

Fig. 1 depicts $\epsilon_{\text{trun}}(\nu)$ for each microphone and for $N = 1, \ldots, 7$ with $\tilde{N} = 10$ as a reference. The corresponding values

of $kr$ are presented at the top of Fig. 1. Fig. 1 shows that for all $N$ the truncation error $\epsilon_{\text{trun}}(\nu)$ is similar between all microphones. This is because of the similar distance of the microphones from the origin. Fig. 1 also shows that, as expected, $\epsilon_{\text{trun}}(\nu)$ is relatively small for $kr \approx N$ and diminishes for $kr \ll N$, for all $N$. This verifies the theory on truncation error discussed in Section V. In particular, for this array and for a maximum operation frequency of 5 kHz, the steering function is approximately order-limited with an order $N = 6$, with a normalized truncation error that is lower than $-10$ dB. This leads to two conclusions: (i) the focusing matrices should be computed with order of $N = 6$, and (ii) the number and distribution of steering function samples (directions) should be sufficient to provide aliasing-free sampling with $N = 6$. This is further discussed in Subsection VI-C.

### B. Matrix Inversion

The error due to the inversion of the matrix $\mathbf{V}(\nu)$ is now examined. As outlined in Section V, an inversion error of zero requires that $I \geq (N+1)^2$, i.e. more microphones than coefficients. Because the array has $I = 12$ microphones this requirement is met for orders $N \leq 2$. This means that at frequencies that require higher orders for an accurate representation of the steering function, i.e. above about 1500 Hz (see Fig. 1), an ideal transformation matrix may not exist. In this case, some focusing error is expected. This will be studied further in Subsection VI-D.

### C. Spatial Aliasing

This subsection examines the aliasing error for the studied array. From Subsection VI-A it follows that for 0–5 kHz the steering function is approximately order-limited with order $N = 6$. Therefore, a small aliasing error will be obtained for sampling schemes that maintain the sampling conditions for this order, i.e. $L \geq 49$, and matrix $\mathbf{Y}(\boldsymbol{\psi})$ is of full rank with a reasonably low condition number. The given sampling set of $L = 240$ nearly-uniformly distributed directions implies oversampling in this case, and thus a negligible aliasing error is expected in this frequency range.

As this is not always the case, a sparser sampling scheme is employed to demonstrate the aliasing error in the case of insufficient sampling. A sampling scheme of $L = 9$ directions is used, which facilitates the computation of coefficients up to a maximal order of $N_{\hat{\mathbf{V}}} = 2$. Fig. 1 depicts in dashed lines the aliasing (and added truncation) error in this case. The figure shows that the aliasing error becomes significant in the range of $kr > 2$, where the number of directions is insufficient relative to the spherical harmonics order of the steering function.

### D. Focusing Performance

The purpose of this subsection is to analyze focusing performance with WINGS transformations, calculated using the proposed approach, and to examine the effects of the various factors on overall focusing error. From Subsection VI-B it was concluded that selecting $N \leq 2$ leads to a zero inversion error. On the other hand, according to Subsection VI-A, the selection of $N = 6$ leads to a small truncation error over the entire operating frequency range. This insight motivates the investigation of the focusing performance for various orders $N$.

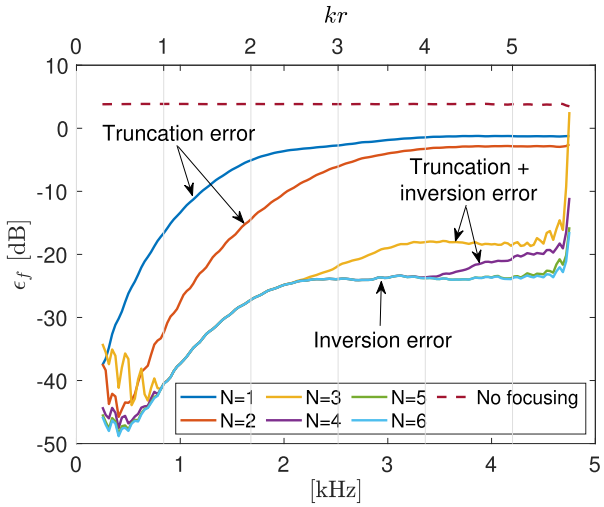


Fig. 2. Mean (over directions) normalized focusing error for a bandwidth of 469 Hz as a function of the central frequency $\nu_0$. The WINGS transformations were computed for orders $N$ from 1 to 6 as denoted on the figure.

The WINGS transformations were computed from different orders $N = 1, \ldots, 6$ according to (14), with $\mathbf{T}_0 = \mathbf{I}$, and for a smoothing bandwidth of 469 Hz. The considerations for choosing the smoothing bandwidth are discussed in the next section. The spherical harmonics coefficient matrices that were used in the computation of (14) are computed from (18) using the $L = 240$ steering function samples. This sampling implies that the aliasing error is negligible in this study.

The measure that was employed for assessing the focusing performance is the mean (over directions) of the normalized focusing error, which is defined as [6]

$$\epsilon_f(\nu_0) = \frac{1}{L}\sum_{l=1}^{L} \frac{\overline{\left\|\mathbf{T}(\nu,\nu_0)\,\mathbf{h}(\nu,\psi_l) - \mathbf{h}(\nu_0,\psi_l)\right\|^2}}{\left\|\mathbf{h}(\nu_0,\psi_l)\right\|^2}, \quad (28)$$

where $\overline{(\cdot)}$ denotes averaging over the frequencies within the processed frequency band. Fig. 2 depicts $\epsilon_f(\nu_0)$ as a function of the central frequency $\nu_0$ for the different orders $N = 1, \ldots, 6$. The dashed line denotes the values of $\epsilon_f(\nu_0)$ when the focusing matrices $\mathbf{T}(\nu,\nu_0)$ were chosen to be the identity matrix, i.e. no focusing was employed.

For $N = 1, 2$, the inversion error is expected to be very small (because $I > (N+1)^2$), so that the total error is likely to be dominated by truncation. Indeed, Fig. 2 shows that, for these orders, the overall focusing error is similar to the truncation error depicted in Fig. 1. For $N > 2$, the error in the two figures is no longer similar, due to the contribution of the matrix inversion error. For $N = 6$, for which truncation is minor, the error is attributed to matrix inversion, whereas for $N = 3, 4, 5$, truncation error is added to the matrix inversion error for $kr > N$, where the effect of truncation is no longer negligible. Fig. 2 also shows that for $kr \ll 1$, performance deteriorates. This can be explained

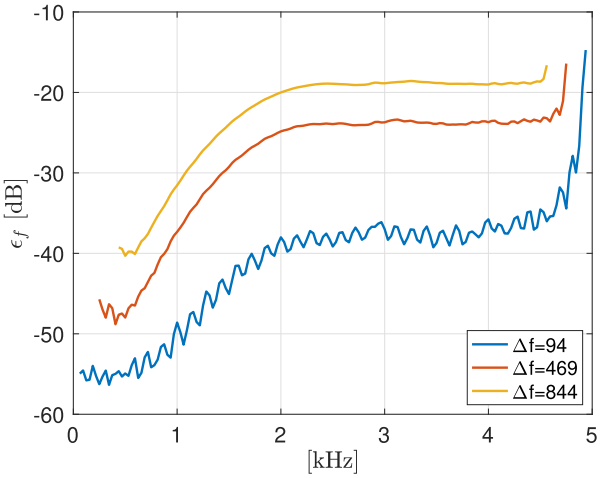


Fig. 3. Mean (over directions) normalized focusing error of the WINGS transformations with $N = 6$ as a function of the central frequency $\nu_0$ for various smoothing bandwidths $\Delta f$.

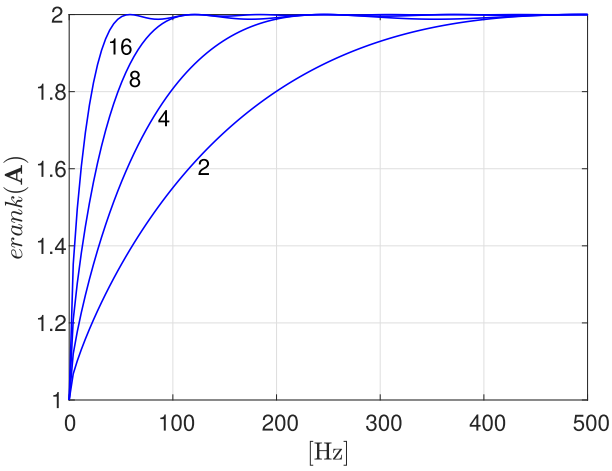


Fig. 4. $erank(\mathbf{A})$ as a function of the smoothing range $\Delta f$ for several delays $\tau_0 = 2, 4, 8, 16$ ms.

by the ill-conditioning of $\mathbf{V}(\nu)$ at low frequencies, leading to an unstable inversion.

## VII. Smoothing Bandwidth Selection

Another parameter that should be determined prior to focusing is the smoothing bandwidth. Fig. 3 depicts $\epsilon_f(\nu_0)$ for the Nao array with WINGS transformations that are computed using spherical harmonics of order $N = 6$, and for several smoothing bandwidths $\Delta f$. Fig. 3 suggests that focusing performance is degraded as the bandwidth increases. This behavior can be explained by noting that the larger the bandwidth the greater the difference between $\mathbf{H}(\nu_0, \psi)$ and $\mathbf{H}(\nu, \psi)$. Consequently, $\mathbf{V}(\nu)$ will differ more significantly from $\mathbf{V}(\nu_0)$ for larger smoothing ranges and the residuals in the focusing LS fitting process will be larger. In light of these results, it may be better to choose a narrow bandwidth to reduce the focusing error, and yet a wide enough bandwidth to successfully decorrelate the source signals and restore the rank of the source signals cross-correlation matrix.

As an example to study this tradeoff, consider a simple scenario of direct sound with unit amplitude, and a single reflection with the same magnitude, arriving after some propagation delay. For these two coherent sources, the source signal's cross-correlation matrix $\mathbf{S_s}(\nu)$ is of unit rank with its column space spanned by the characteristic vector $[1, e^{j\frac{2\pi Fs}{M}\nu\tau_0}]^T$, where $\tau_0$ is the delay between the signals. The frequency smoothing operation exploits the frequency diversity of the characteristic vectors within the bandwidth to restore the rank. It is therefore expected that a larger bandwidth will provide better frequency smoothing. The effective-rank [24] can be employed for measuring the frequency diversity of the characteristic vectors within a smoothing window. The effective-rank of

$$\mathbf{A} = \begin{bmatrix} 1 & \cdots & 1 \\ e^{j\frac{2\pi Fs}{M}(\nu_0 - \lfloor \frac{J_\nu}{2} \rfloor)\tau_0} & \cdots & e^{j\frac{2\pi Fs}{M}(\nu_0 + \lfloor \frac{J_\nu}{2} \rfloor)\tau_0} \end{bmatrix} \quad (29)$$

as a function of the bandwidth $\Delta f = \frac{Fs}{M} J_\nu$ was examined. The matrix $\mathbf{A}$ consists of the characteristic vectors from the smoothing bandwidth, where the $j_\nu$-th column of the matrix $\mathbf{A}$ holds the characteristic vector from frequency $\frac{Fs}{M}(\nu_0 + j_\nu)$, $j_\nu = -\lfloor \frac{J_\nu}{2} \rfloor, \ldots, \lfloor \frac{J_\nu}{2} \rfloor$. The effective-rank is a real-valued measure that generalizes the rank term and quantifies the dimensionality of a matrix. The effective-rank of $\mathbf{A}$ is defined as [24]

$$erank\,(\mathbf{A}) = \exp\left\{ -\sum_{k=1}^{2} \frac{\sigma_k}{\|\sigma_k\|_1} \log \frac{\sigma_k}{\|\sigma_k\|_1} \right\}, \quad (30)$$

where $\sigma_k$ is the $k$-th singular value of $\mathbf{A}$ and $\|\cdot\|_1$ denotes the $l_1$-norm. Since the number of rows of $\mathbf{A}$ is fixed to be 2, $erank(\mathbf{A})$ ranges between 1 and 2, where for bandwidths for which $erank(\mathbf{A}) = 1$ the vectors in $\mathbf{A}$ are linearly dependent, so that smoothing with this bandwidth will not restore its rank. The closer that $erank(\mathbf{A})$ is to 2, the larger the frequency diversity of the column in $\mathbf{A}$, and so smoothing will probably succeed in decorrelating the sources. Note that $erank(\mathbf{A})$ is not a function of the central frequency $\nu_0$, since applying the unitary transformation $\begin{bmatrix} 1 & 0 \\ 0 & e^{-j\frac{2\pi Fs}{M}\nu_0\tau_0} \end{bmatrix}$ to $\mathbf{A}$ eliminates its dependence on $\nu_0$ while not affecting its singular values.

Fig. 4 shows the values of $erank(\mathbf{A})$ as a function of $\Delta f$ for delays $\tau_0 = 2, 4, 8, 16$ ms, which represent the delay values of typical early reflections in a room. Fig. 4 shows that $erank(\mathbf{A})$ is lower for shorter delays, suggesting that decorrelating sources with short delays requires a wider bandwidth. Examining the results in Figs. 3 and 4, it appears that a bandwidth of about $\Delta f = 500$ Hz is a good choice for delays of $\tau_0 \geq 2$ ms since it is sufficiently wide for decorrelating sources and yet narrow enough to achieve accurate focusing. Because reflections in rooms typically have a larger delay, it is expected that this bandwidth should be sufficient for applications in the realm of sound in rooms. For different applications, additional factors may need to be taken into consideration when choosing the smoothing bandwidth. In particular, as the number of speakers increases,

the W-disjoint orthogonality assumption, which is exploited for multiple speaker localization in Section VIII, may be violated as the bandwidth broadens. However, the investigation of this assumption is beyond the scope of this work and will be left for further study.

## VIII. Application to Speaker Localization

In this section the focusing method developed in this article is integrated into a speaker localization method based on the DPD test [9], which has been developed for spherical microphone arrays. The proposed method therefore extends the DPD test based method to arrays with an arbitrary configuration. The DPD test is used to select time-frequency (TF) bins with one dominant source, assuming that these bins contain a significant contribution from the direct-sound and insignificant contributions from room reflections, and then only these bins are used for the estimation task [9]. In the DPD test based method [9], the plane-wave decomposition (PWD) operation, which is tailored to spherical arrays, is employed to remove the steering matrices' frequency dependence to support local frequency smoothing operation to decorrelate coherent sources. In the proposed extension, the direction-independent focusing process developed here is employed instead of the PWD operation, while the rest of the DPD test based algorithm remains unchanged [10], [11]. The development of the DPD test for arbitrary arrays is now presented.

To construct the smoothed cross-spectrum in bin $(\tau_0, \nu_0)$, local focusing is applied to a rectangular window of $J_\tau$ time frames and $J_\nu$ frequencies around $(\tau_0, \nu_0)$, followed by an averaging of the spectrum over the window. The averaging over time frames approximates the statistical expectation operation and the averaging over frequencies implements frequency smoothing. With ideal focusing, the smoothed cross-spectrum can be expressed as in (5). For clarity, (5) is rewritten here as

$$\begin{aligned}\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}\left(\tau_0,\nu_0\right) &= \overline{\mathbf{p}\left(\tau_0,\nu_0\right)\mathbf{p}^H\left(\tau_0,\nu_0\right)} \\ &= \mathbf{H}\left(\nu_0,\boldsymbol{\psi}\right)\overline{\mathbf{S}_{\mathbf{s}}}\left(\tau_0,\nu_0\right)\mathbf{H}^H\left(\nu_0,\boldsymbol{\psi}\right) + \overline{\mathbf{S}_{\tilde{\mathbf{n}}}}\left(\tau_0,\nu_0\right), \quad (31)\end{aligned}$$

where the over-line represents averaging over the rectangular window around $(\tau_0, \nu_0)$. After smoothing, the matrix $\overline{\mathbf{S}_{\mathbf{s}}}(\tau_0, \nu_0)$ is assumed to be of full rank with a low condition number. This enables the identification of bins with one source by examining the numerical rank of $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau_0, \nu_0)$. The set of bins selected by the DPD test is

$$\mathcal{A}_{\text{DPD}} = \left\{(\tau_0,\nu_0) : \frac{\sigma_1\left(\tau_0,\nu_0\right)}{\sigma_2\left(\tau_0,\nu_0\right)} > \mathcal{TH}_{\text{DPD}}\right\}, \quad (32)$$

where $\sigma_1(\tau_0, \nu_0)$ and $\sigma_2(\tau_0, \nu_0)$ are the largest and second largest singular values of $\overline{\mathbf{S}_{\tilde{\mathbf{p}}}}(\tau_0, \nu_0)$ and $\mathcal{TH}_{\text{DPD}}$ is a chosen threshold.

Several other measures for the dominance of the direct-path have been proposed. The measure proposed in [25] quantifies the similarity of the first eigenvector to a specific steering vector. Another measure is based on the sound field directivity, which is a computationally-efficient alternative to the aforementioned singular values ratio measure [26]. However, the use of these measures is limited to spherical arrays since they are based on processing in the spherical harmonics formulation. Contrary to these measures, the measure proposed in [14] can be applied to arbitrary arrays.

Several approaches for estimating the speaker DoA from the selected bins have been proposed, including MUSIC with coherent and incoherent integration of the signal subspaces from the different bins [9], and bin-wise DoA estimation followed by statistical analysis to fuse the estimates [27]–[30].

TABLE I
The Number of Sources, Their Distance From the Array and Their DoAs (Elevation and Azimuth Angles) for the Five Recordings

| | | Array distance | True DoAs |
|---|---|---|---|
| Recording 1 | Speaker #1 | 2.6 m | $(92.4^\circ, 58.4^\circ)$ |
| Recording 2 | Speaker #1 | 1.72 m | $(84.5^\circ, -55^\circ)$ |
| Recording 3 | Speaker #1 | 1.34 m | $(93^\circ, -13^\circ)$ |
| Recording 4 | Speaker #1 | 1.36 m | $(92.2^\circ, -25.8^\circ)$ |
| | Speaker #2 | 1.5 m | $(94.6^\circ, -10.4^\circ)$ |
| Recording 5 | Speaker #1 | 2.62 m | $(92.8^\circ, 21.7^\circ)$ |
| | Speaker #2 | 2.22 m | $(89.9^\circ, 48.6^\circ)$ |
| | Speaker #3 | 1.69 m | $(90.2^\circ, -53.6^\circ)$ |
| | Speaker #4 | 2.49 m | $(94.3^\circ, -7^\circ)$ |

## IX. Experimental Verification

This section aims to evaluate and compare the performance of the proposed method, the local space-domain distance (LSDD)-DPD test proposed in [14] and the DPD test [9] in estimating the DoAs of both single and multiple speakers in real-world conditions. The LSDD-DPD test selects TF bins based on signals' similarity to a steering vector, suggesting the existence of a single source. Unlike the proposed method, the LSDD-DPD test does not use either time or frequency smoothing, and therefore does not require focusing. The underlying assumption is that successful focusing and frequency smoothing with the proposed method will facilitate good DoA estimation performance. For this purpose, the real-world recordings with the Nao array, obtained as a part of the LOCATA challenge, are employed [23]. Data was recorded in a laboratory of the Department of Computer Science at Humboldt University Berlin, of size 7.1 m $\times$ 9.8 m $\times$ 3 m with an approximate reverberation time of $T_{60} = 0.55$ s. Speech segments were played through stationary loudspeakers and perceived by the array. The number of sources, their DoAs and the distance of each source from the array for each of the recorded scenes are presented in Table I. The Nao array steering vectors available for the challenge were interpolated in space to 7442 directions that follow a Gaussian sampling scheme to obtain a spatial resolution of $3^\circ$.

The recorded data was down-sampled from 44 kHz to 16 kHz before it was processed by the tested methods to align with the sampling frequency of the steering vectors. The recorded signal was transformed to the STFT domain using a 512 samples (32 ms) Hann window with an overlap of 16 ms. For the proposed method, the cross-spectrum was computed according to (31) with averaging over 3 time frames and 15 frequencies. The WINGS transformations were computed as described in Subsection VI-D using a spherical harmonics order of $N = 6$. The LSDD-DPD test was implemented according to [14]. The

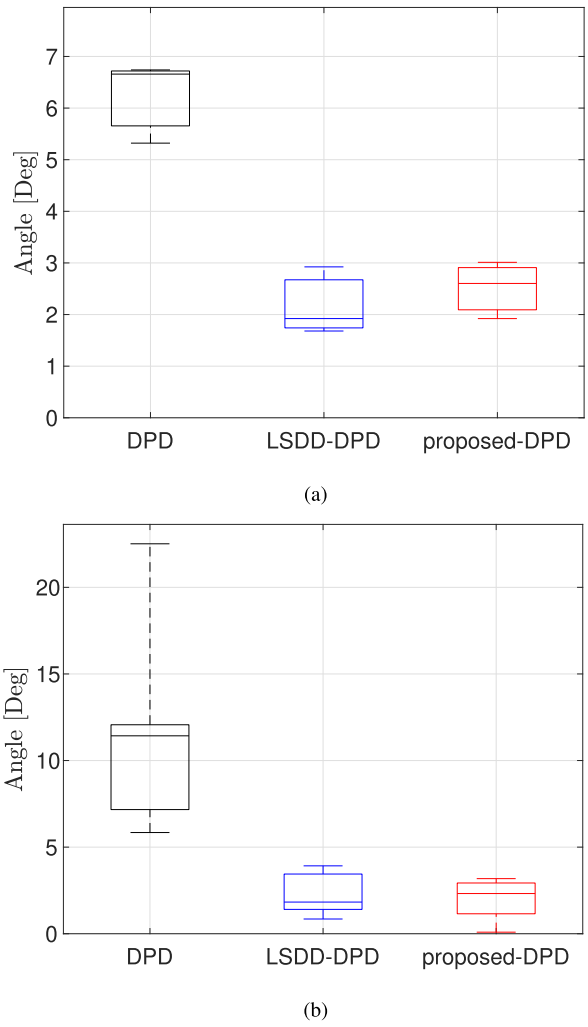


Fig. 5. DoA estimate errors for the examined methods for (a) single speaker and (b) multiple speakers.

DPD test method was implemented according to [31], with the spherical harmonics coefficients of the plane-wave density estimated up to spherical harmonics order $N = 1$ and with averaging over 3 time frames and 15 frequencies to construct the correlation matrix. For all tested methods the minimal operating frequency was limited to 1 kHz due to the array aperture. The maximal operating frequency of the proposed method and of the LSDD-DPD test was selected to be 5 kHz, while the maximal operating frequency of the DPD test was limited to 2.5 kHz due to spatial aliasing. The threshold for each test was chosen such that 5% of all available bins will pass the test. For all methods, MUSIC with a source subspace of a single dimension is applied to the bins that passed the tests. A fusion phase is applied, for all methods, after estimating the DoA from each of the selected bins. The DoAs are first represented as unit vectors in Cartesian coordinates, and then MATLAB implementation of the K-means algorithm (`kmeans()`) with the cosine distance metric is employed to classify the estimates from the different bins. The number of clusters was equal to the number of speakers, which is assumed to be known. The final DoA estimates were chosen to be the mean of the obtained clusters.

The angular distance (in the $\phi\theta$ plane) between an estimate and the true source direction was employed as a measure for DoA estimation error. For multiple speakers, all permutations of the DoA estimation vector were compared with the true DoAs and the permutation that minimizes the MSE was used to associate each estimate with a source. Fig. 5 shows the estimation error bars for each method and for a single speaker scenario (a) and for a multiple speakers scenario (b). Fig. 5 shows that the proposed extension of the DPD test achieved performance comparable with the LSDD-DPD test and with a maximal error smaller than $5^\circ$ for both the single and the multiple speakers. This result demonstrates the feasibility of the proposed method in estimating the DoAs in real-world scenarios. The inferior performance of the DPD test is due to a relatively narrow operational frequency band and also due to data reduction caused by transforming the 12-dimensional microphone signal to a 4 harmonics PWD domain.

The proposed method achieved good performance that is comparable with the state-of-the-art LSDD-DPD test [14]. This result verifies that the focusing and smoothing technique works well. However, the LSDD-DPD test [14] is strictly used for DoA estimation while the proposed method can be used to compute smoothed cross-spectrum matrices for other applications, such as speech enhancement [17], blind source separation [18], and beamforming [19]. Moreover, the measure proposed in [14] is based on a steered beam response; thus, it may have resolution limitations, especially when applied to arrays with a small number of microphones, such as a binaural array. However, the investigation of the latter is left for future study.

## X. Conclusion

Direction-independent focusing methods facilitate the computation of the coherently smoothed spatial spectrum matrix, in particular in reverberant environments. In this article, a direction-independent focusing method has been developed. The proposed method extends the current methods to arrays with a steering function that is available only for a set of selected directions. Spherical harmonics decomposition of the steering function was employed to formulate factors that affect the focusing error and to assess the number of required directions. A case of two coherent sources has been studied, leading to the conclusion that a smoothing bandwidth of about 500 Hz is wide enough for successful frequency smoothing in typical scenes of speech in rooms. Finally, an experimental study that employed recordings with the Nao robot array from the LOCATA challenge showed that the proposed focusing based extension of the DPD test method achieves comparable performance to that of the state-of-the-art method called the LSDD-DPD test. This result implies that the proposed focusing and smoothing process works well and can be used for applications other than DoA.

## References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Berlin, Germany: Springer Science & Business Media, 2008.
[2] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8, Berlin, Germany: Springer, 2015.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[4] 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. TASSP-33, no. 4, pp. 823–831, Aug. 1985.
[5] F. Sellone, "Robust auto-focusing wideband DOA estimation," *Signal Process.*, vol. 86, no. 1, pp. 17–37, 2006.
[6] T.-S. Lee, "Efficient wideband source localization using beamforming invariance technique," *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 1376–1387, Jun. 1994.
[7] M. A. Doron and A. Nevet, "Robust wavefield interpolation for adaptive wideband beamforming," *Signal Process.*, vol. 88, no. 6, pp. 1579–1594, 2008.
[8] V. Tourbabin and B. Rafaely, "Optimal design of microphone array for humanoid-robot audition," (a) in *Proc. Israeli Conf. Robot.*, 2016.
[9] 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.
[10] H. Beit-On and B. Rafaely, "Speaker localization using the direct-path dominance test for arbitrary arrays," in *Proc. IEEE Int. Conf. Sci. Elect. Eng.* Israel, 2018, pp. 1–4.
[11] H. Beit-On and B. Rafaely, "Binaural direction-of-arrival estimation in reverberant environments using the direct-path dominance test," in *Proc. 23nd Int. Congr. Acoust.*, Aachen, Germany, Sep. 2019, pp. 3305–3312.
[12] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2171–2186, Nov. 2016.
[13] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.
[14] V. Tourbabin, D. L. Alon, and R. Mehra, "Space domain-based selection of direct-sound bins in the context of improved robustness to reverberation in direction of arrival estimation," in *Proc. 11th Eur. Congr. Expo. Noise Control Eng.*, 2018, pp. 2589–2596.
[15] H. W. Löllmann *et al.*, "The Locata challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 410–414.
[16] C. Evers *et al.*, "The Locata challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1620–1643, 2020.
[17] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497–507, Sep. 2000.
[18] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 204–215, May 2003.
[19] Y. Bucris, I. Cohen, and M. A. Doron, "Bayesian focusing for coherent wideband beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1282–1296, May 2011.
[20] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[21] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, 2004.
[22] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 8, pp. 1272–1281, Aug. 1988.
[23] "LOCATA website," 2018. [Online]. Available: www.locata-challenge.org. Accessed on: Oct. 15, 2019.
[24] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *Proc. 15th Eur. Signal Process. Conf.*, 2007, pp. 606–610.
[25] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 131–142, Mar. 2019.
[26] B. Rafaely and K. Alhaiany, "Speaker localization using direct path dominance test based on sound field directivity," *Signal Process.*, vol. 143, pp. 42–47, 2018.
[27] B. Rafaely, C. Schymura, and D. Kolossa, "Speaker localization in a reverberant environment using spherical statistical modeling," *J. Acoust. Soc. Amer.*, vol. 141, no. 5, pp. 3523–3523, 2017.
[28] B. Rafaely and D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6120–6124.
[29] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DoA histograms," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 1473–1477.
[30] B. Rafaely, D. Kolossa, and Y. Maymon, "Towards acoustically robust localization of speakers in a reverberant environment," in *Proc. Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 96–100.
[31] V. Tourbabin and B. Rafaely, "Speaker localization by humanoid robots in reverberant environments," in *Proc. IEEE 28th Conv. Elect. Electron. Eng.*, Israel, 2014, pp. 1–5.