



# Direction of Arrival Estimation for Reverberant Speech Based on Enhanced Decomposition of the Direct Sound

Lior Madmoni , *Student Member, IEEE*, and Boaz Rafaely , *Senior Member, IEEE*

**Abstract**—Direction of arrival (DOA) estimation for speech sources is an important task in audio signal processing. This task becomes a challenge in reverberant environments, which are typical to real scenarios. Several methods of DOA estimation for speech sources have been developed recently, in an attempt to overcome the effect of reverberation. One effective approach aims to identify time-frequency bins in the short time Fourier transform domain that are dominated by the direct sound. This approach was shown to be particularly adequate for spherical arrays, with processing in the spherical harmonics domain. The direct-path dominance (DPD) test, and a method which is based on the directivity of the sound field are recent examples. While these methods seem to perform well, high reverberation conditions may degrade their performance. In this paper, the structure of the spatial correlation matrix is comprehensively studied, showing that under some well-defined conditions, the DOA of the direct sound can be correctly extracted from its dominant eigenvector, even when contaminated by reflections. This new insight leads to the development of a new test, performing an enhanced decomposition of the direct sound (EDS), denoted the DPD-EDS test. The proposed test is compared to previous DPD tests, and to other recently proposed reverberation-robust methods, using computer simulations and an experimental study, demonstrating its potential advantage. The studies include multiple speakers in highly reverberant environments, therefore representing challenging real-life acoustics scenes.

**Index Terms**—Direction of arrival estimation, reverberation, spherical arrays, plane wave decomposition.

## I. INTRODUCTION

**D**IRECTION-OF-ARRIVAL (DOA) estimation is an important task in the field of audio signal processing, employed extensively in applications such as signal enhancement, video conferencing, and robot audition. Some common DOA estimation algorithms include beamforming [1], subspace methods such as Multiple Signal Classification (MUSIC) [2], and time-delay estimation based methods [3].

DOA estimation becomes more challenging in reverberant environments, where room reflections mask the direct sound. This may degrade the performance of the aforementioned DOA estimation methods in practical, real-life acoustic scenes. Therefore,

Manuscript received July 1, 2018; revised October 31, 2018; accepted November 6, 2018. Date of publication December 10, 2018; date of current version April 11, 2019. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Peter K. Willett. (*Corresponding author: Lior Madmoni.*)

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

Digital Object Identifier 10.1109/JSTSP.2018.2885930

new methods for DOA estimation have been recently developed to overcome the effect of reverberation.

One approach to overcome reverberation has been proposed in the context of time-delay estimation. In this work, linear regression of the time delay estimates has been computed, with a cost function that is robust to outliers due to reverberation [4]. However, this method was only designed and studied for relatively low levels of reverberation. In [4], and later in [5], [6] a moderate reverberation time of 0.4 s was employed. Furthermore, time-delay estimation methods assume that the microphones are positioned in free space, therefore preventing their use in arrays mounted around rigid bodies, for example.

A different set of methods that overcome reverberation while facilitating a flexible array configuration, are based on relative transfer function estimation [7]–[9]. The direct-path component of the relative transfer-functions (DP-RTF) is typically estimated from the measured data. Therefore, the performance of these methods strongly depends on the accuracy of the DP-RTF estimation [7]. Moreover, acoustic parameters of the environment, such as reverberation time, direct-to-reverberation ratio (DRR) and noise characteristics, should be available for a reliable estimation of the DP-RTF. In practice, however, these may not be available or may be challenging to estimate.

Methods that do not require acoustic information related to the environment, and do not rely on estimation of transfer functions, include [10], and [11]. The former is designed for acoustic vector sensors, which are not commonly used in practice, and which may suffer from excessive noise at the low frequency range of the speech signal.

The latter method [11] is designed for spherical arrays of commonly-used pressure microphones, and is based on processing in the spherical harmonics (SH) domain. This domain supports a signal model with a frequency-independent array manifold, such that frequency smoothing can be directly applied to decorrelate coherent sources and reduce the effect of room reflections [12], [13]. The latter leads to the formulation of the direct-path dominance (DPD) test, designed to select time-frequency (TF) bins in the short time Fourier transform (STFT) domain that are dominated by a single source, typically representing the direct sound. Only these selected bins are then used for DOA estimation, providing accurate estimates even under reverberation.

Several extensions have been developed for this method, including Gaussian mixture modeling (GMM) to eliminate

outliers in the DOA estimation [14], and improve the robustness to challenging acoustic environments [15]. The computational cost of these methods is relatively high, since they require eigendecomposition of the spatial spectrum matrix at each TF bin. Hence, an alternative method has been proposed which is based on a sound field directivity measure that does not require eigendecomposition [16], and in this work will be referred to as DPD-DIR. While these methods have been shown to perform well, under high reverberation or highly challenging acoustic conditions, a large proportion of TF bins that pass these tests, may hold incorrect DOA information. Although these wrong DOAs may be eliminated to some extent by identifying them as outliers using GMM clustering [14], in some applications, the remaining percentage of bins that hold correct DOA information may not be sufficient for a reliable estimate of the source DOA, e.g. when only a very short signal is available or when tracking a moving speaker [17], [18].

This work aims to overcome the limitations of previous methods, offering several contributions.

- 1) An improved DPD test is developed, following a theoretical analysis showing that under some well-defined conditions, the information on the direct-path is accurately maintained by the significant eigenvector of the spatial correlation matrix, even when contaminated by room reflections. This leads to the development of a new test based on enhanced decomposition of the direct sound (EDS for short), and in this paper it is referred to as the DPD-EDS test.
- 2) Computer simulations showing that the DPD-EDS test is significantly more accurate in selecting TF bins that contain correct DOA information, compared to previous methods: the DPD test, the DPD-DIR test, the coherence test [19], and a DRR based test [20]. The study specifically shows that compared to previous tests, the new test can extract more TF bins, and with a greater accuracy, from the same data.
- 3) Experimental study with multiple speakers using data taken from the recent challenge on acoustic source localization and tracking (LOCATA) [21] validates the conclusions from the computer simulation study.

While the work describing the DPD-EDS test has been accepted for publication in a conference paper [22], this new paper presents a more comprehensive and extensive theoretical development of the new test, a more extensive simulation study and an experimental study with multiple speakers, that include a comparison to previous work.

## II. SYSTEM MODEL

This section presents the spherical microphone array based system model, that will be utilized throughout this work. The standard spherical coordinate system is used, denoted by  $(r, \theta, \phi)$ , where  $r$  is the distance from the origin,  $\theta$  is the angle measured downwards from the positive  $z$  axis to the  $xy$  plane, and  $\phi$  is the angle measured between the positive  $x$  axis towards the positive  $y$  axis. Consider an array comprised of  $Q$  omni-directional microphones, arranged in an arbitrary

configuration, with  $\{\mathbf{r}_q \equiv (r_q, \theta_q, \phi_q)\}_{q=1}^Q$  denoting the microphones positions. Consider also a sound field which is comprised of  $L$  far field sources at wavenumber  $k$ , and with DOAs of  $\{\Psi_l \equiv (\theta_l, \phi_l)\}_{l=1}^L$ , such that the sound pressure measured by the microphone array is described by the following narrowband model [23]:

$$\mathbf{p}(k) = \mathbf{V}(k, \Psi)\mathbf{s}(k) + \mathbf{n}(k), \quad (1)$$

where  $Q \times 1$  vector  $\mathbf{p}(k) = [p(k, \mathbf{r}_1), p(k, \mathbf{r}_2), \dots, p(k, \mathbf{r}_Q)]^T$  holds the noisy sound pressure measured by the microphones, the  $L \times 1$  source signal amplitudes vector is given by  $\mathbf{s}(k) = [s_1(k), s_2(k), \dots, s_L(k)]^T$ ,  $\mathbf{V}(k, \Psi)$  is a  $Q \times L$  steering matrix describing the propagation between each source and microphone [23], the  $Q \times 1$  noise vector is given by  $\mathbf{n}(k) = [n_1(k), n_2(k), \dots, n_Q(k)]^T$ , and  $(\cdot)^T$  denotes the transpose operator. The  $L$  sources can represent, in the context of this work, the direct sound from speakers in a room, and reflections from room boundaries.

The general model presented in (1), can also be used for an array which is of a spherical geometry, such that  $r_q = r$  for all  $q = 1, \dots, Q$ . This spherical array can facilitate the processing of signals in the SH domain [24]–[26]. Next, plane wave decomposition can be performed, and assuming that the sound field is comprised of far field sources (e.g., distant speakers) leads to the following [27]:

$$\mathbf{a}_{\text{nm}}(k) = \mathbf{Y}^H(\Psi)\mathbf{s}(k) + \tilde{\mathbf{n}}(k), \quad (2)$$

where the  $(N+1)^2 \times 1$  vector  $\mathbf{a}_{\text{nm}}(k) = [a_{00}(k), a_{1(-1)}(k), a_{10}(k), \dots, a_{NN}(k)]^T$  holds the noisy plane wave density (PWD) coefficients in the SH domain, the steering matrix in this domain is denoted by the  $(N+1)^2 \times L$  matrix  $\mathbf{Y}^H(\Psi) = [\mathbf{y}^*(\Psi_1), \mathbf{y}^*(\Psi_2), \dots, \mathbf{y}^*(\Psi_L)]$ , with columns  $\mathbf{y}(\Psi_l) = [Y_0^0(\Psi_l), Y_1^{-1}(\Psi_l), \dots, Y_N^N(\Psi_l)]^T$ , where  $Y_n^m(\cdot)$  are the SH functions of order  $n$  and degree  $m$ , with  $N$  denoting the maximal SH order, usually set to  $N = \lceil kr \rceil$  and satisfying  $(N+1)^2 \leq Q$  [26], [28].  $\tilde{\mathbf{n}}(k)$  is an  $(N+1)^2 \times 1$  vector holding the measurement noise components in the SH domain,  $(\cdot)^*$  and  $(\cdot)^H$  are the complex conjugate and the Hermitian operators, respectively.

Next, the PWD measurements in (2) are transformed to the STFT domain, where the W-disjoint orthogonality [29] can be utilized:

$$\mathbf{a}_{\text{nm}}(\tau, \omega) = \mathbf{Y}^H(\Psi)\mathbf{s}(\tau, \omega) + \tilde{\mathbf{n}}(\tau, \omega), \quad (3)$$

where  $\tau$  is the time index and  $\omega$  is the frequency index.

The local TF correlation matrices can now be computed for every  $(\tau, \omega)$  by averaging the PWD measurements over  $J_\tau$  time frames and  $J_\omega$  frequency bins:

$$\begin{aligned} \tilde{\mathbf{R}}_a(\tau, \omega) &= \frac{1}{J_\tau J_\omega} \sum_{j_\omega=0}^{J_\omega-1} \sum_{j_\tau=0}^{J_\tau-1} \mathbf{a}_{\text{nm}}(\tau - j_\tau, \omega - j_\omega) \\ &\quad \times \mathbf{a}_{\text{nm}}^H(\tau - j_\tau, \omega - j_\omega). \end{aligned} \quad (4)$$

Note that since the steering vectors are frequency-independent in this domain, frequency smoothing can be performed directly as in (4) [13], without the need for focusing matrices. This

frequency smoothing process is necessary to de-correlate coherent signals which are present in a multi-path environment [12], therefore clearly separating the contributions of the direct sound and room reflections within the spatial spectrum matrix.

### III. OVERVIEW OF PREVIOUS DOA ESTIMATION METHODS

In this section, several previous DOA estimation methods are presented that have been developed for speakers in reverberant environments. These methods aim to identify TF bins that are dominated by the direct sound, such that correct DOA estimation can be performed with these bins even under reverberation. Specifically, the coherence test [19], the DPD test [11], the DPD-DIR test [16] and a method based on the DRR measure [20], [30], are presented here. These are compared to the proposed method in the simulation study and the experimental study, in Sections VI and VII, respectively.

#### A. The Coherence Test

The coherence test described in [19], aims to identify TF bins with contribution from only a single source. In the development of this test, the environment is assumed to be anechoic, but with multiple uncorrelated sources. This test does not require a specific array geometry, and was developed for a space domain system model, as in (1). The model is then transformed to the STFT domain, from which the local spatial correlation matrices are computed by averaging over  $J_t$  time frames:

$$\tilde{\mathbf{R}}_p(\tau, \omega) = \frac{1}{J_t} \sum_{j=0}^{J_t-1} \mathbf{p}(\tau-j, \omega) \mathbf{p}^H(\tau-j, \omega). \quad (5)$$

The coherence test, suggested in [19], searches for TF bins with  $\tilde{\mathbf{R}}_p(\tau, \omega)$  of a unit rank, representing the contribution from a single source, as follows:

$$\mathcal{A}_{\text{coh}} = \left\{ (\tau, \omega) : \text{rank}1(\tilde{\mathbf{R}}_p(\tau, \omega)) > \mathcal{TH}_{\text{coh}} \right\}, \quad (6)$$

where  $\mathcal{A}_{\text{coh}}$  is the set of bins that pass the test, and  $\text{rank}1(\cdot)$  is one if the matrix has a unit rank, such that the threshold  $\mathcal{TH}_{\text{coh}}$  should be set relatively close to 1. A measure based on the computationally efficient pair-wise magnitude-squared coherences (MSC) was proposed in [19], [31], but in this work, a measure based on the eigenvalues ratio as in (8) has been adopted.

DOA estimation is now performed using MUSIC with a signal subspace of a single dimension, representing a single source:

$$\Omega_{\text{coh}} = \left\{ \Omega : \arg \max_{\Omega} \frac{1}{\|\tilde{\mathbf{U}}_n^H(\tau, \omega) \mathbf{v}(\Omega)\|^2}, \forall (\tau, \omega) \in \mathcal{A}_{\text{coh}} \right\}, \quad (7)$$

where  $\tilde{\mathbf{U}}_n(\tau, \omega)$  is the noise subspace of size  $Q \times (Q-1)$  [2], and  $\mathbf{v}(\Omega)$  is a  $Q \times 1$  steering vector corresponding to arrival direction  $\Omega$ . In the case of a single source, the final DOA estimate can be computed as the mean of  $\Omega_{\text{coh}}$ . Alternatively, clustering the DOAs in  $\Omega_{\text{coh}}$  can be applied to eliminate outliers, or, in the case of multiple speakers, to estimate the DOA of each speaker [11], [32]. Because room reflections are coherent with the direct sound, bins that contain direct sound and reflections may still

have a rank close to one, potentially degrading the performance of the coherence test, leading to errors under reverberation [11].

#### B. The DPD Test

Similar to the coherence test, the DPD test also measures the rank of the local TF correlation matrix to determine if a TF bin is dominated by a single source. Unlike the coherence test, the correlation matrix,  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , as in (4), is computed by applying frequency smoothing to de-correlate coherent sources (i.e., direct sound and reflections), therefore reducing the harmful effect of room reflections on the test [11]. Now, a unit-rank test is applied by computing the ratio of the two largest eigenvalues,  $\sigma_1(\tau, \omega)$  and  $\sigma_2(\tau, \omega)$  of matrix  $\tilde{\mathbf{R}}_a(\tau, \omega)$ . This leads to the definition of the DPD test [11]:

$$\mathcal{A}_{\text{DPD}} = \left\{ (\tau, \omega) : \frac{\sigma_1(\tau, \omega)}{\sigma_2(\tau, \omega)} > \mathcal{TH}_{\text{DPD}} \right\}, \quad (8)$$

where  $\mathcal{A}_{\text{DPD}}$  is the set of TF bins that pass the test, and  $\mathcal{TH}_{\text{DPD}}$  is a threshold chosen to be sufficiently larger than 1.

Finally, MUSIC is applied to compute the DOAs:

$$\Omega_{\text{DPD}} = \left\{ \Omega : \arg \max_{\Omega} \frac{1}{\|\tilde{\mathbf{U}}_n^H(\tau, \omega) \mathbf{y}^*(\Omega)\|^2}, \forall (\tau, \omega) \in \mathcal{A}_{\text{DPD}} \right\}, \quad (9)$$

where  $\tilde{\mathbf{U}}_n(\tau, \omega)$  is the noise subspace of  $\tilde{\mathbf{R}}_a(\tau, \omega)$  [2]. A single DOA estimate can be computed as in Subsection III-A, by the mean of  $\Omega_{\text{DPD}}$  or by performing a further step of clustering, such as GMM [14].

#### C. The DPD-DIR Test

The DPD-DIR test also aims to identify bins dominated by the direct sound, but unlike the DPD test, it is applied directly on the PWD measurements in (3). The test uses the following directivity measure of the sound field [16]:

$$\mathcal{DIR}(\tau, \omega) = \frac{\max_{\Omega} |\mathbf{y}(\Omega)^T \mathbf{a}_{\text{nm}}(\tau, \omega)|^2}{\mathbf{a}_{\text{nm}}(\tau, \omega)^H \mathbf{a}_{\text{nm}}(\tau, \omega)}. \quad (10)$$

The maximum directivity is  $(N+1)^2$  [16], achieved when the sound field is composed of a single plane wave. Hence, the DPD-DIR test is defined as [16]:

$$\mathcal{A}_{\text{DIR}} = \left\{ (\tau, \omega) : \mathcal{DIR}(\tau, \omega) > \mathcal{TH}_{\text{DIR}} \right\}, \quad (11)$$

where  $\mathcal{TH}_{\text{DIR}} \in [1, (N+1)^2]$ , but typically set to a value close to  $(N+1)^2$ , such that the set  $\mathcal{A}_{\text{DIR}}$  is comprised of TF bins dominated by a single plane wave. Note that no decomposition of matrix  $\tilde{\mathbf{R}}_a(\tau, \omega)$  is necessary; the computation complexity of this test is therefore significantly lower than that of the DPD test.

The DOA estimation for each bin is given by the argument  $\Omega$  that maximizes  $|\mathbf{y}(\Omega)^T \mathbf{a}_{\text{nm}}(\tau, \omega)|^2$ , already computed in (10),

$$\Omega_{\text{DIR}} = \left\{ \Omega : \arg \max_{\Omega} |\mathbf{y}(\Omega)^T \mathbf{a}_{\text{nm}}(\tau, \omega)|^2, \forall (\tau, \omega) \in \mathcal{A}_{\text{DIR}} \right\}. \quad (12)$$

Similarly to the coherence and DPD tests, a single DOA estimate can be computed as the mean of  $\Omega_{\text{DIR}}$  or by performing a further step of clustering, as described in [16].

#### D. A DRR Based Test

Another test that can be used to identify TF bins dominated by the direct sound, is based on the DRR, generally given by:

$$\text{DRR} = \frac{P_D}{P_R}, \quad (13)$$

where  $P_D$  and  $P_R$  are the energy of the direct sound and the reverberant components of the sound field, respectively. Therefore, DRR with a high value means dominance of the direct sound over the reverberant component [30]. Several methods have been presented to estimate the DRR measure for each TF bin using spherical arrays, [20], [33]. In this work, the DRR was computed according to the method in [20], which is based on the decomposition of  $\tilde{\mathbf{R}}_a(\tau, \omega)$  into direct and reverberant components. Since this method requires an initial DOA of the direct sound for DRR estimation, the DOAs from all TF bins were computed prior to DRR estimation. For further details on this method, the reader is referred to [20]. Finally, the DRR based test can be applied as follows:

$$\mathcal{A}_{\text{DRR}} = \left\{ (\tau, \omega) : \text{DRR}(\tau, \omega) > \mathcal{TH}_{\text{DRR}} \right\}, \quad (14)$$

where  $\mathcal{A}_{\text{DRR}}$  is the set of TF bins that pass the test, and  $\mathcal{TH}_{\text{DRR}}$  is a test threshold which should be sufficiently high (typically satisfying  $\mathcal{TH}_{\text{DRR}} \gg 1$ ) to indicate that the TF bin is dominated by the direct sound. Similarly to previous methods, a single DOA estimation can then be computed only from bins that pass the test.

#### IV. ANALYSIS OF DIRECT SOUND INFORMATION IN THE CORRELATION MATRIX

In this section, the eigendecomposition of the local TF correlation matrix,  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , is investigated. It is shown that under some well-defined conditions, accurate direct sound information is available in the first eigenvector corresponding to the largest eigenvalue, even when the sound field is composed of room reflections in addition to the direct sound. This important understanding will be used later to develop the DPD-EDS test.

Following the formulation of the PWD coefficients in (3), consider the simple case where the sound field is comprised of  $L = 2$  far field sources. Equation (3) in this case is described by

$$\mathbf{a}_{\text{nm}}(\tau, \omega) = \mathbf{y}^*(\Psi_1)s_1(\tau, \omega) + \mathbf{y}^*(\Psi_2)s_2(\tau, \omega) + \tilde{\mathbf{n}}(\tau, \omega), \quad (15)$$

where  $\Psi_1$  and  $\Psi_2$  are the DOAs of the sources. In this case, the correlation matrix, computed using expectation, (estimated in (4) using summations), takes the following form:

$$\begin{aligned} \mathbf{R}_a(\tau, \omega) &= \mathbb{E}[\mathbf{a}_{\text{nm}}(\tau, \omega)\mathbf{a}_{\text{nm}}^H(\tau, \omega)] \\ &= [\mathbf{y}^*(\Psi_1) \ \mathbf{y}^*(\Psi_2)] \mathbf{R}_s(\tau, \omega) \begin{bmatrix} \mathbf{y}^T(\Psi_1) \\ \mathbf{y}^T(\Psi_2) \end{bmatrix} \\ &\quad + \mathbf{R}_{\tilde{\mathbf{n}}}(\tau, \omega), \end{aligned} \quad (16)$$

where  $\mathbf{R}_a(\tau, \omega)$  is the correlation matrix of size  $(N+1)^2 \times (N+1)^2$  at TF  $(\tau, \omega)$ ,  $\mathbf{R}_s(\tau, \omega)$  is the  $2 \times 2$  source correlation matrix, and  $\mathbf{R}_{\tilde{\mathbf{n}}}(\tau, \omega)$  is the  $(N+1)^2 \times (N+1)^2$  noise correlation matrix. Next, it is assumed that the signals  $s_1(\tau, \omega)$  and  $s_2(\tau, \omega)$  are uncorrelated, such that

$$\mathbf{R}_s(\tau, \omega) = \begin{bmatrix} \sigma_1^2(\tau, \omega) & 0 \\ 0 & \sigma_2^2(\tau, \omega) \end{bmatrix}, \quad (17)$$

where  $\sigma_1^2(\tau, \omega)$  and  $\sigma_2^2(\tau, \omega)$  are the variances of signals  $s_1(\tau, \omega)$  and  $s_2(\tau, \omega)$ , respectively.

Assuming further that the noise is white, leads to

$$\mathbf{R}_{\tilde{\mathbf{n}}}(\tau, \omega) = \sigma_n^2 \mathbf{I}, \quad (18)$$

where  $\sigma_n^2$  is the noise variance and  $\mathbf{I}$  is the identity matrix of size  $(N+1)^2$ . Substituting (17) and (18) in (16) yields:

$$\begin{aligned} \mathbf{R}_a(\tau, \omega) &= \sigma_1^2(\tau, \omega) \mathbf{y}^*(\Psi_1) \mathbf{y}^T(\Psi_1) \\ &\quad + \sigma_2^2(\tau, \omega) \mathbf{y}^*(\Psi_2) \mathbf{y}^T(\Psi_2) \\ &\quad + \sigma_n^2 \mathbf{I}. \end{aligned} \quad (19)$$

Note that while (19) was derived for two uncorrelated sources, it will also approximately hold for a direct sound and its reflection after frequency smoothing, which has been shown to perform decorrelation [12], [13], and which is performed when  $\mathbf{R}_a(\tau, \omega)$  is estimated as in (4).

Next, (19) is multiplied from the right by  $\mathbf{y}^*(\Psi_1)$ :

$$\begin{aligned} \mathbf{R}_a(\tau, \omega) \mathbf{y}^*(\Psi_1) &= \sigma_1^2(\tau, \omega) \mathbf{y}^*(\Psi_1) \mathbf{y}^T(\Psi_1) \mathbf{y}^*(\Psi_1) \\ &\quad + \sigma_2^2(\tau, \omega) \mathbf{y}^*(\Psi_2) \mathbf{y}^T(\Psi_2) \mathbf{y}^*(\Psi_1) \\ &\quad + \sigma_n^2 \mathbf{y}^*(\Psi_1). \end{aligned} \quad (20)$$

Note that the terms  $\mathbf{y}^T(\Psi_1) \mathbf{y}^*(\Psi_1)$  and  $\mathbf{y}^T(\Psi_2) \mathbf{y}^*(\Psi_1)$  are scaled versions of samples of the maximum directivity beam-pattern in the SH domain [26]. Thus, if  $\Psi_1$  and  $\Psi_2$  are sufficiently spatially separated with respect to the beamwidth of the maximum directivity beampattern, the following will typically hold:

$$\mathbf{y}^T(\Psi_2) \mathbf{y}^*(\Psi_1) \ll \mathbf{y}^T(\Psi_1) \mathbf{y}^*(\Psi_1). \quad (21)$$

Now, the spatial angle  $\Theta$  between  $\Psi_1$  and  $\Psi_2$ , is defined as

$$\cos \Theta = \cos \theta_1 \cos \theta_2 + \cos(\phi_1 - \phi_2) \sin \theta_1 \sin \theta_2 \quad (22)$$

and the beamwidth of the maximum directivity beampattern has been shown to approximately equal to  $\frac{\pi}{N}$ , [34], which provides an approximated criterion for the inequality in (21) to hold:  $\Theta > \frac{\pi}{N}$ .

Next, substituting (21) in (20) yields

$$\mathbf{R}_a(\tau, \omega) \mathbf{y}^*(\Psi_1) \approx [\sigma_1^2(\tau, \omega) \|\mathbf{y}^*(\Psi_1)\|^2 + \sigma_n^2] \mathbf{y}^*(\Psi_1), \quad (23)$$

which means that  $\mathbf{y}^*(\Psi_1)$  is an eigenvector of  $\mathbf{R}_a(\tau, \omega)$  and its eigenvalue is  $[\sigma_1^2(\tau, \omega) \|\mathbf{y}^*(\Psi_1)\|^2 + \sigma_n^2]$ . Therefore, according to (23), the eigendecomposition of  $\mathbf{R}_a(\tau, \omega)$  may include a normalized version of  $\mathbf{y}^*(\Psi_1)$  as its eigenvector. While eigendecomposition is not unique, the vector space comprised of eigenvectors corresponding to identical eigenvalues is unique [35]. Therefore, if  $\sigma_1^2(\tau, \omega)$  is larger than  $\sigma_2^2(\tau, \omega)$ , then  $\mathbf{y}^*(\Psi_1)$

and  $\mathbf{y}^*(\Psi_2)$  correspond to different eigenvalues (note that (23) also holds for  $\mathbf{y}^*(\Psi_2)$ ). Hence, in this case, the first eigenvector corresponding to the largest eigenvalue, denoted  $\mathbf{u}_1(\tau, \omega)$ , will satisfy

$$\mathbf{u}_1(\tau, \omega) \approx c\mathbf{y}^*(\Psi_1), \quad (24)$$

where  $c$  is some complex constant, such that  $\mathbf{u}_1(\tau, \omega)$  is of unity norm. Now, assuming  $\mathbf{y}^*(\Psi_1)$  represents the direct sound, and  $\mathbf{y}^*(\Psi_2)$  a reflection, and assuming they are sufficiently spatially separated, the direct sound information is accurately maintained in the first eigenvector of  $\mathbf{R}_a(\tau, \omega)$ , even when  $\mathbf{a}_{\text{nm}}(\tau, \omega)$  is corrupted by the reflection.

Although the analysis leading to (24) was presented for only two sources, for simplicity, this theoretical formulation can be extended to the more general case of multiple sources, representing a direct sound with multiple reflections, assuming the reflections are sufficiently separated from the direct sound. Nevertheless, a rigorous formulation and proof are left for future work.

Finally, although the condition related to the spatial separation between the direct sound and room reflections leading to (24) may not always hold in practice, it may still approximately hold for a significant number of bins, potentially leading to a more accurate decomposition of the direct sound component from the spatial spectrum matrix. Equation (24) will be used in the following section for the development of the DPD-EDS test.

## V. THE PROPOSED DPD-EDS TEST

In this section, a new DPD test is developed. The new test aims to be more robust to reverberation than previous tests, based on the new insights into the decomposition of the spatial correlation matrix, as presented in Section IV. These new insights led to the result that the first eigenvector,  $\mathbf{u}_1(\tau, \omega)$ , of  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , may be approximately proportional to the steering vector of the direct sound. Motivated by this important result, the extent to which  $\mathbf{u}_1(\tau, \omega)$  in a given TF bin represents a direct sound, can now be measured. One way is to quantify the similarity between  $\mathbf{u}_1(\tau, \omega)$  and a vector of the form  $\mathbf{y}^*(\Omega)$ , using a MUSIC-based measure

$$\mathcal{EDS}(\tau, \omega) = \max_{\Omega} \frac{1}{\left\| \mathbf{P}_{\mathbf{u}_1(\tau, \omega)}^{\perp} \mathbf{y}^*(\Omega) \right\|^2}, \quad (25)$$

where

$$\mathbf{P}_{\mathbf{u}_1(\tau, \omega)}^{\perp} = \mathbf{I} - \frac{\mathbf{u}_1(\tau, \omega)\mathbf{u}_1^H(\tau, \omega)}{\|\mathbf{u}_1(\tau, \omega)\|^2} \quad (26)$$

is the projection into the subspace which is orthogonal to  $\mathbf{u}_1(\tau, \omega)$ . This measure is invariant to scaling, necessary to compensate for the unknown scalar  $c$  in (24). Note that other scale-invariant measures for the proximity between  $\mathbf{u}_1(\tau, \omega)$  and a single plane wave,  $\mathbf{y}^*(\Omega)$ , can also be used.

Finally, the DPD-EDS test can be formulated as

$$\mathcal{A}_{\text{EDS}} = \left\{ (\tau, \omega) : \mathcal{EDS}(\tau, \omega) > \mathcal{TH}_{\text{EDS}} \right\}, \quad (27)$$

where  $\mathcal{A}_{\text{EDS}}$  is the set of bins that pass the DPD-EDS test, and  $\mathcal{TH}_{\text{EDS}}$  is a threshold value which should be chosen sufficiently

larger than 1. This is because the measure suggested in (25) returns zero in the denominator for perfect similarity to a single plane wave, and so, an expected high score for the measure should satisfy  $\mathcal{TH}_{\text{EDS}} \gg 1$ .

Now, bin-wise DOA estimation is given by the argument  $\Omega$  that maximizes  $\mathcal{EDS}(\tau, \omega)$ ,

$$\Omega_{\text{EDS}} = \left\{ \Omega : \arg \max_{\Omega} \mathcal{EDS}(\tau, \omega), \forall (\tau, \omega) \in \mathcal{A}_{\text{EDS}} \right\}, \quad (28)$$

already computed in (25). A single DOA estimate can then be computed as the mean of  $\Omega_{\text{EDS}}$  or by a further step of clustering, similarly to as described in Section III.

It is important to address the differences between the DPD-EDS test and the previous DPD and DPD-DIR tests, in order to understand the novelty of the new test. Unlike the original DPD test, which only measures the ratio of the first two dominant eigenvalues of  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , the DPD-EDS test explicitly examines the presence of a direct sound information in the dominant eigenvector of  $\tilde{\mathbf{R}}_a(\tau, \omega)$ . Thus, it is less likely for TF bins dominated by reverberation to pass the DPD-EDS test, even if their correlation matrix has a dominant first eigenvalue. The original DPD test, on the other hand, may pass such bins, as will be demonstrated in the next section. In addition, note that both the original DPD test and the DPD-EDS test use similar expressions for the noise subspace, as described in (9) for the DPD test, and in (25) for the DPD-EDS test. However, the latter uses this noise subspace for identifying direct TF bins, as in (27) and for DOA estimation as in (28), while the DPD test uses the noise subspace only for the DOA estimation stage, as in (9). While the DPD-DIR test also examines the proximity to a single plane wave sound field, by using the directivity measure, it does so directly for the measured PWD coefficients,  $\mathbf{a}_{\text{nm}}(\tau, \omega)$ . If they are distorted by reflections, the DPD-DIR test will score low, not passing these TF bins. However, as argued in Section IV, the direct sound information may be present at the dominant eigenvector, even in the case of contribution from reflections. This is a potential advantage for the DPD-EDS test, which may extract accurate direct sound information from more TF bins.

The computation complexity of the proposed method may be of importance when only limited computing resources are available. Hence, the computation complexity of the proposed test is presented in Table I, also compared to the coherence test, the DPD test, and the DPD-DIR test. The table shows the complex multiplications necessary for each test at each TF bin. Each step within the test is presented separately: correlation matrix estimation, eigendecomposition, and the grid search to estimate the DOAs. It is clear that the DPD-DIR is the most efficient among the four tests, since no correlation matrix estimation nor eigendecomposition are necessary. The DPD-EDS test is the most computationally complex in the grid search stage, since the DPD and coherence tests use grid search only for bins that pass the test, while the DPD-EDS test uses grid search for all TF bins. Recall from Subsection III-D, that the DRR based test also calculates matrix  $\tilde{\mathbf{R}}_a(\tau, \omega)$  and uses grid search for all TF bins. In addition, the method uses matrix multiplication that requires  $\mathcal{O}((N+1)^8)$  complex multiplications for each TF bin

TABLE I  
COMPUTATIONAL COMPLEXITY OF THE COHERENCE TEST, THE DPD TEST, THE DPD-DIR TEST, AND THE DPD-EDS TEST, SHOWN IN TERMS OF COMPLEX MULTIPLICATIONS NEEDED FOR EACH STEP AND FOR EACH TF BIN

Method	Correlation matrix estimation	Eigendecomposition	Grid search
Coherence test	$J_t \times Q(5)$	$\mathcal{O}(Q^3)$	$(Q - 1) \times Q$ (7)
DPD test	$J_\tau \times J_\omega \times (N + 1)^2$ (4)	$\mathcal{O}((N + 1)^6)$	$((N + 1)^2 - 1) \times (N + 1)^2$ (9)
DPD-DIR test	0	0	$(N + 1)^2$ (12)
DPD-EDS test	$J_\tau \times J_\omega \times (N + 1)^2$ (4)	$\mathcal{O}((N + 1)^6)$	$((N + 1)^2 - 1) \times (N + 1)^2$ (28)

[20]. Hence, this method is more computationally complex than the DPD-EDS test.

## VI. SIMULATION STUDY

In this section, a simulation study is presented, aiming to evaluate the performance of the proposed DPD-EDS test, and compare it to the other tests described in Section III.

A spherical array is simulated using MATLAB [36], consisting of  $Q = 32$  microphones arranged around a rigid sphere of radius  $r = 4.2$  cm (similar to the Eigenmike [37]) facilitating SH processing with an order of  $N = 3$ . The array was positioned in a room with dimensions  $8 \times 5 \times 3$  m simulated using the image method [38], and with its center located at (4, 3, 0.8) m. The walls reflection coefficients were chosen to lead to reverberation time of  $T_{60} = 1$  s, and a critical distance of 0.6 m. Three point sources were positioned in the room with the following distances from the array center: 1.87 m, 1.78 m and 1.81 m, and with the following DOAs:  $\Psi_1 = (61.3^\circ, 267.5^\circ)$ ,  $\Psi_2 = (66.9^\circ, 192.4^\circ)$ , and  $\Psi_3 = (56.5^\circ, 57.6^\circ)$ , respectively. Speech signals from the TIMIT [39] database were used for the three source signals, with lengths of approximately 4 s, and a sampling frequency of 16 kHz. A diffuse noise was superimposed on the sound field generated by the three sources, with a signal-to-noise ratio (SNR) of 20 dB at the microphones. Sensor noise with an SNR of 40 dB was also added to the microphone signals. A transformation of the signals to the STFT domain was performed with a Hanning window of size 512 samples, and an overlap of 50%. An analysis frequency range of [400, 5000] Hz was employed, which leads to a total of approximately 38,000 TF bins for a 4 s recording. The PWD coefficients, as in (2), were computed from the pressure measurements in (1), with a similar method to the R-PWD described in [40] (equation (2.27)). The calculation of the spatial correlation matrix for the coherence test,  $\tilde{\mathbf{R}}_p(\tau, \omega)$  as in (5) was performed with  $J_t = 7$  time frames. The calculation of the local TF correlation matrix,  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , for the DPD test, the DRR based test, and the DPD-EDS test was performed as in (4) with  $J_\tau = 2$  time frames, and  $J_\omega = 15$  frequency bins. Finally, each of the tests described in this work, namely the coherence test [19], the DPD test [11], the DPD-DIR test [16], the DRR based test [20], and the newly developed DPD-EDS test, was computed for the recorded microphone signals. The simulation parameters described above were used in all of the following subsections of this simulation study.

An example of typical running times for these methods is presented in Table II. The running times were measured on a MacBook Pro laptop computer, 2015 model, with 8 GB of RAM, and 2.7 GHz Intel Core i5 processor. For this running

TABLE II  
TYPICAL RUNNING TIMES FOR THE TESTS UNDER STUDY FOR A 4 S RECORDING, ON A MACBOOK PRO LAPTOP. THE TEST THRESHOLDS FOR EACH TEST HAS BEEN SET SUCH THAT 5000 BINS PASS THE TEST

Method	Running time
Coherence test	15.72 s
DPD test	7.27 s
DPD-DIR test	3.91 s
DRR based test	53.66 s
DPD-EDS test	33 s

time comparison only, the thresholds for each test were chosen such that 5000 TF bins pass the tests, which are approximately 13% of all bins. Table II shows that the DRR based test has the longest running time, since it requires the largest number of complex multiplications, as described in Section V. The second longest running time is for the DPD-EDS test, since it requires the eigendecomposition of the correlation matrix for all TF bins. In addition, the running time for the coherence test is significantly longer than the running time of the DPD test, which may be due to the fact that  $Q > (N + 1)^2$  in this case, see (2), leading to correlation matrices of much higher dimension. Nevertheless, in [19] a computationally efficient alternative was proposed which was not implemented here. The shortest running time is achieved by the DPD-DIR test, because it does not require matrix decomposition.

### A. Performance Analysis - DOA Histograms

In this subsection, 3D histograms of the DOA estimates from TF bins that passed each test are presented, with the aim of studying the DOA distribution compared to the true direction of the sources. The threshold for each test was chosen such that 4000 bins pass the test, which are approximately 10.5% of all available TF bins. This approach to threshold selection was chosen so that the different tests, which use a diverse set of measures, could be evaluated on a common basis.

Fig. 1 presents the 3D histogram of the DOA estimates from TF bins that passed the coherence test. The horizontal green bars indicate the true DOAs. It seems that a significant number of TF bins hold DOA information which is far from the true DOAs. This is somewhat expected, because for this test, bins that are contaminated by coherent reflections may still have a rank close to one and pass the test (see Section III-A). Figs. 2, 3, and 4 present the histogram of the DOA estimates from TF bins that passed the DPD test, the DPD-DIR test, and the DRR based test, respectively. Note that unlike the histogram of the coherence test in Fig. 1, there are clearly visible three dominant

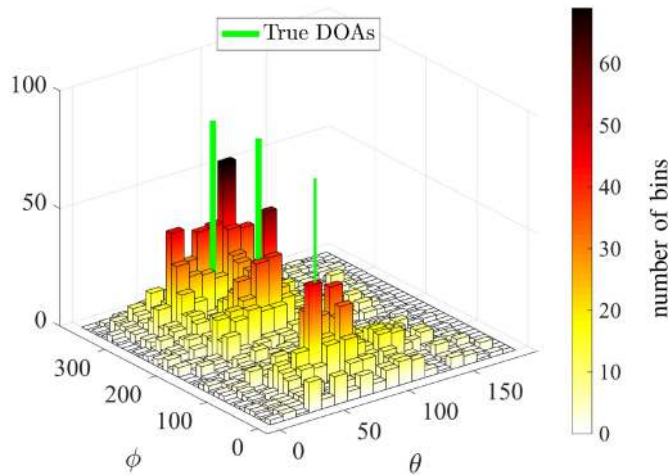


Fig. 1. Histogram of DOA estimates from TF bins that passed the coherence test. The threshold was chosen such that 4000 bins passed the test. The horizontal green bars indicate the true DOAs.

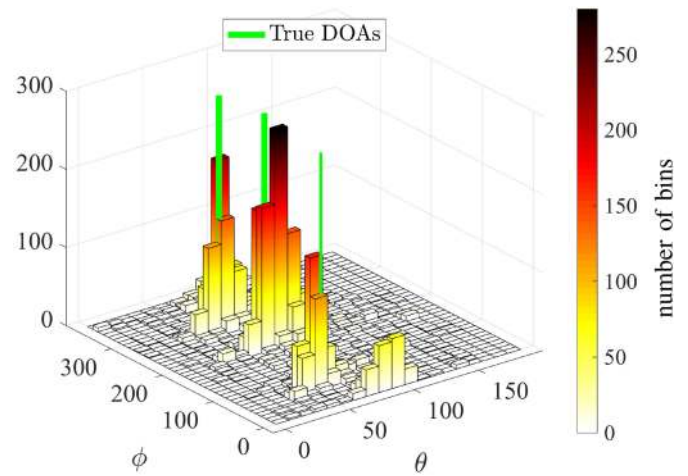


Fig. 4. Same as Fig. 1, but for TF bins that passed the DRR test.

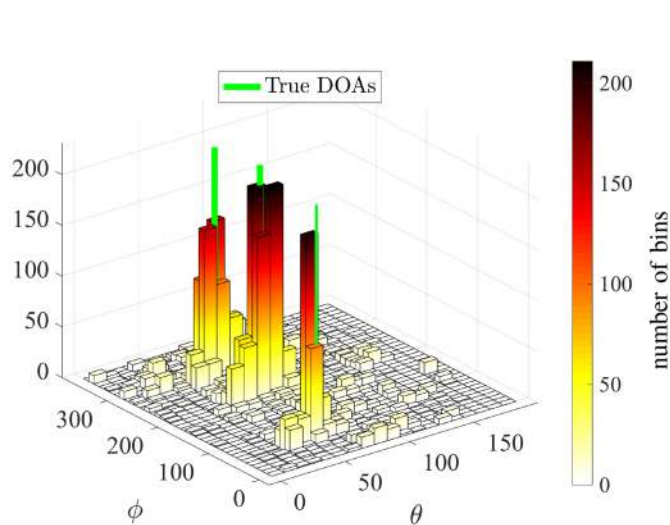


Fig. 2. Same as Fig. 1, but for TF bins that passed the DPD test.

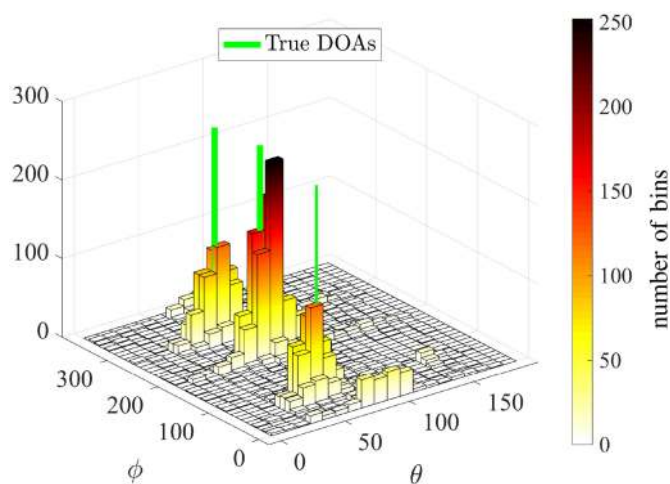


Fig. 3. Same as Fig. 1, but for TF bins that passed the DPD-DIR test.

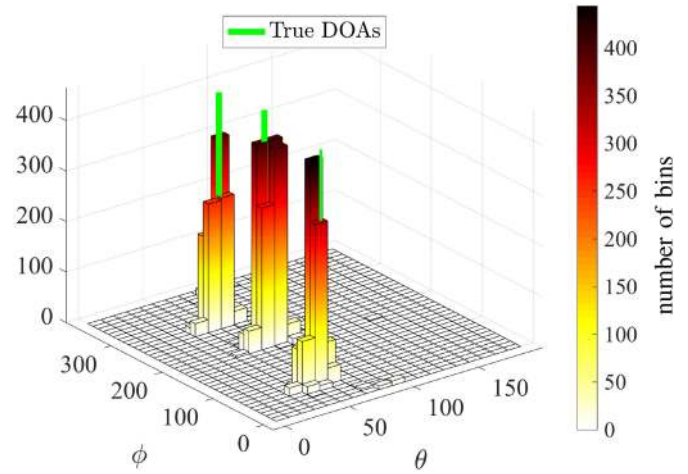


Fig. 5. Same as Fig. 1, but for TF bins that passed the DPD-EDS test.

clusters corresponding to TF bins that hold a relatively correct DOA information on the three speakers. In addition, other  $(\theta, \phi)$  regions, which are far from the true DOAs, also contain contribution from TF bins, although, they seem less dominant compared to the coherence test. This is expected as all three tests are designed for reverberant speech.

Finally, Fig. 5 presents the histogram of the DOAs estimated from bins that passed the DPD-EDS test. In this case, the TF bins located around the true DOAs seem the largest in number, compared to all other tests, as evident from the histogram height scale. In addition,  $(\theta, \phi)$  regions far from the true DOAs seem to have relatively few TF bins. This is a first evidence to validate the theoretical development leading to the DPD-EDS test - the test manages to extract accurate direct sound information even under high reverberation.

The analysis presented in this subsection is somewhat qualitative. In the following subsection, the performance of DOA estimation will be investigated with several quantitative measures.

TABLE III

THE THRESHOLD FOR EACH TEST, REQUIRED TO ACHIEVE THE PERCENTAGE OF TF BINS PASSING EACH TEST, AS DENOTED IN FIG. 6. THE TOTAL NUMBER OF BINS IS 38,000

TF bins [%]	1	5	10	15	20	25	30	35	40
DPD-EDS test	8.72	4.24	2.82	2.25	2.02	1.85	1.74	1.62	1.57
DPD test	5.11	3.21	2.60	2.29	2.07	1.94	1.83	1.74	1.66
DPD-DIR test	10.33	8.02	6.95	6.39	5.97	5.66	5.40	5.19	5.00
DRR based test	20.90	9.10	6.40	5.19	4.47	3.96	3.58	3.27	3.03
Coherence test	34.69	12.30	6.45	4.37	3.48	2.88	2.59	2.39	2.19

### B. Performance Analysis - DOA Estimation

In this subsection, the extent to which the TF bins that passed a test hold correct DOA information will be investigated quantitatively. As a first step in this analysis, TF bins that passed each test were assigned to one of the three speakers. Let  $\Psi_{TF} = (\theta_{TF}, \phi_{TF})$  denote the DOA estimated from a given TF bin. The error between  $\Psi_{TF}$  and the true DOA  $\Psi_i = (\theta_i, \phi_i)$  for  $i = 1, 2, 3$ , i.e., the three speakers, is calculated as

$$\Delta\theta_i = \cos^{-1} \left( \cos \theta_{TF} \cos \theta_i + \cos(\phi_{TF} - \phi_i) \sin \theta_{TF} \sin \theta_i \right), \quad (29)$$

which is the relative spatial angle, similar to (22). Next, each TF bin was associated with the nearest speaker, based on  $\Delta\theta_i$ . The spatial angle between the DOA for that bin and the DOA of the selected speaker is denoted as  $\Delta\theta$ .

It should be noted that this informed assignment of speakers to TF bins was performed to avoid limitations of practical clustering methods [11], [14], [16], and focus the analysis in this work on performance of the tests to provide accurate DOAs from individual bins.

The performance of bin-wise DOA estimation is studied next for all tests, with respect to the following measures:

- 1)  $\Theta_{ERR}$ , defined as the mean of  $\Delta\theta$  averaged over all TF bins that passed the test and all three speakers.
- 2)  $\sigma_{ERR}$ , defined as the standard deviation (STD) of  $\Delta\theta$  with respect to the true directions, averaged over all TF bins that passed the test and all three speakers.
- 3)  $P_{10^\circ}$ , defined as the percentage of TF bins with DOA estimates satisfying  $\Delta\theta < 10^\circ$  relative to all TF bins that passed the test [30].

Fig. 6(a) presents  $\Theta_{ERR}$  as a function of the percentage of TF bins that passed each test. Control over the percentage of bins that pass each test was achieved through the test threshold. Typical threshold values are presented in Table III. Fig. 6(a) shows that the errors of the TF bins that passed the DPD-EDS test are significantly lower than all other methods, in particular when 5%–25% of the TF bins pass the test. The TF bins that passed the DPD-DIR test also perform well but mainly when only a relatively small percentage of bins pass the test. The bins that passed the DPD test and the DRR based test have relatively similar errors at the lower percentage range, while at the higher range, the DRR based test has lower errors. As expected, the bins that passed the coherence test achieve the largest errors of all tests.

Fig. 6(b) presents  $\sigma_{ERR}$  as a function of the percentage of TF bins that passed each test. Bins that passed the DPD-EDS test have the lowest STDs among all tests for percentage values up to 30%. At 30% and higher, the coherence test has the lowest

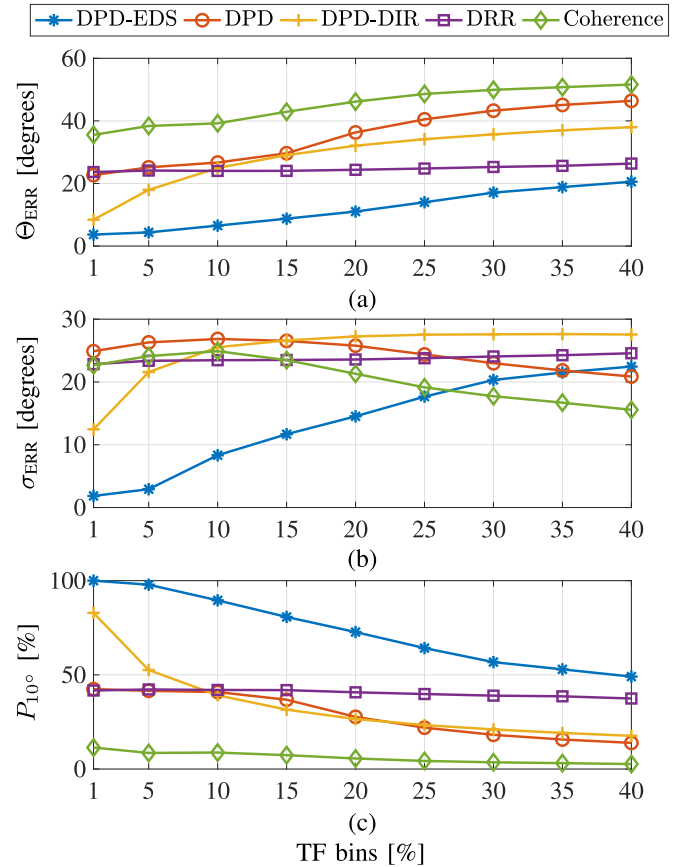


Fig. 6. (a)  $\Theta_{ERR}$ , (b)  $\sigma_{ERR}$  and (c)  $P_{10^\circ}$  as a function of the percentage of TF bins that passed each test with the simulated data of three speakers. The total number of bins is 38,000.

STDs. However, at this range, the mean error  $\Theta_{ERR}$  is relatively high, and so this may not be a useful operating point for all tests. Nevertheless, it may serve to demonstrate the limits of performance of the tests in terms of percentage of useful TF bins.

Finally, Fig. 6(c) shows the measure  $P_{10^\circ}$  as a function of the percentage of TF bins that passed each test. The figure shows that TF bins that passed the DPD-EDS test achieve very high values of  $P_{10^\circ}$ , when the percentage of bins is 10% or less. The  $P_{10^\circ}$  remains higher than all the other methods for all percentage values. However, at 30% and higher, this measure is low for all tests, suggesting that the threshold may be too low, as discussed above. The bins that passed the DPD-DIR test also achieve a relatively high percentages of  $P_{10^\circ}$ , for percentage values around 1%. The remaining tests, and in particular the coherence test, have relatively low values of  $P_{10^\circ}$  at the entire percentage range.

To summarize this study, the DPD-EDS test has been shown to have superior performance over previous tests. In the scenario analyzed here, the DOA estimates from TF bins that passed the DPD-EDS test showed the lowest DOA estimation errors, the lowest STDs at the majority of the TF bins percentage range, and the highest  $P_{10^\circ}$  percentage values. These results validate the theoretical basis of the proposed method - accurate DOAs can be estimated from TF bins even if contaminated by reverberation. The improved level of accuracy is reflected by the fact that



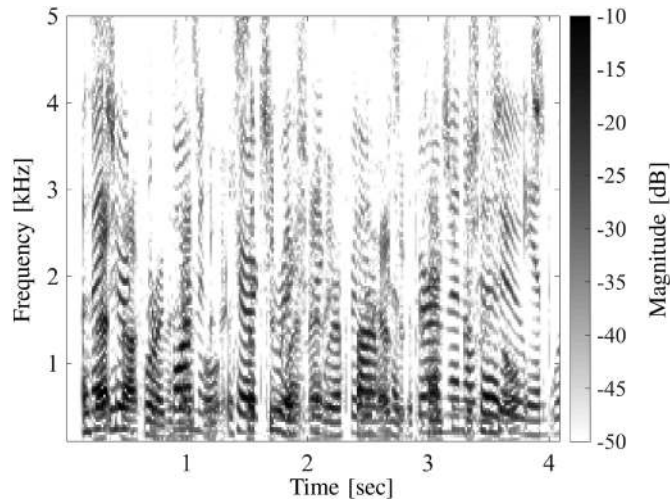


Fig. 7. A spectrogram of the three speech signals.

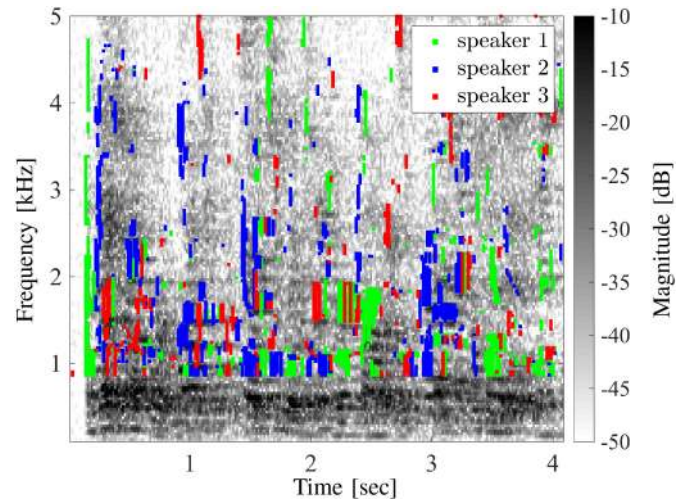


Fig. 9. Same as Fig. 8, but for the DPD test.

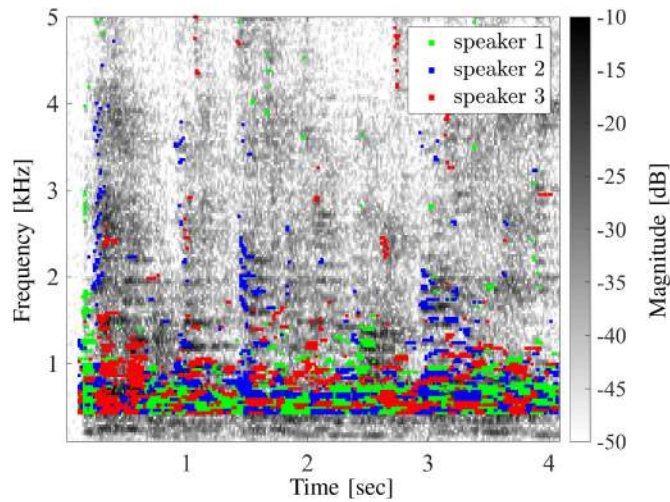


Fig. 8. A spectrogram of the recorded reverberant speech from the three speakers. The TF bins that passed the coherence test are marked on the spectrogram, with green, blue and red colors corresponding to the three speakers. The test thresholds was set such that 4000 TF bins passed the test, out of a total of 38,000 bins.

many more bins pass the proposed test, and with a smaller DOA estimation error, compared to previous tests.

### C. Speech Spectrograms and DPD Maps

The aim of the analysis presented in this subsection is to provide further insight into the proposed test, by presenting the selected TF bins on top of speech spectrograms, using maps referred to as DPD maps [32]. Once again, the threshold for each test was chosen such that the percentage of TF bins that pass each test will be 10.5%, to support a common basis for comparison. The clean speech spectrogram is presented in Fig. 7, for reference, while in the following figures, the reverberant speech spectrogram is highlighted with TF bins that passed each test. Bins are assigned to speakers using the method described in the previous subsection.

TF bins that passed the coherence test are marked on Fig. 8, with the three colors representing the three speakers. Note that

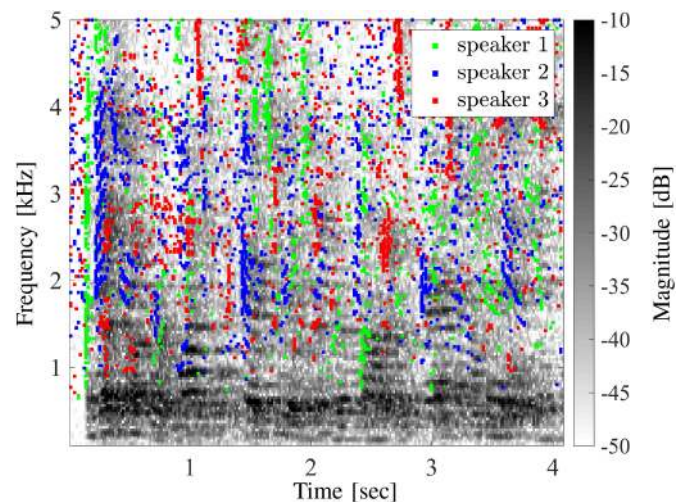


Fig. 10. Same as Fig. 8, but for the DPD-EDS test.

the majority of TF bins that passed the coherence test are located at the low frequency range. One possible explanation is the relatively higher spatial correlation of reverberant or diffuse sound fields at low frequencies [41], leading to higher coherence scores between microphones at low frequencies.

A similar spectrogram is presented in Figs. 9, 10, and 11, with TF bins that passed the DPD test, the DPD-EDS test, and the DRR based test, respectively. Note that most TF bins that passed these tests are located at frequencies above about 800 Hz. This could be the result of the processing in the SH domain, to robustly compute PWD coefficients, with poor estimation of the coefficients at low frequencies due to sensor noise, leading to an ineffective test. The DPD-EDS test does not perform averaging of TF bins, and as a result, the TF bins that passed this test are more sparsely scattered, compared to other tests.

Finally, Fig. 12 presents the same spectrogram with TF bins that passed the DPD-EDS test. Note that the bins that passed this test are located above 1.3 kHz, a range that is higher than that of the other tests. This may suggest that the DPD-EDS test is slightly more sensitive to errors in the computation of the PWD

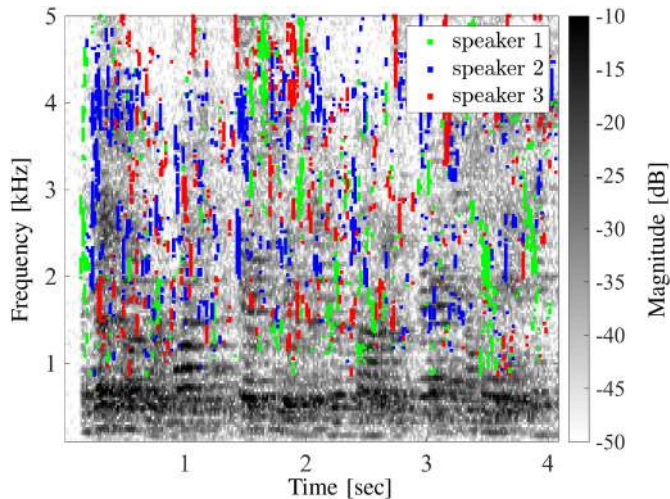


Fig. 11. Same as Fig. 8, but for the DRR test.

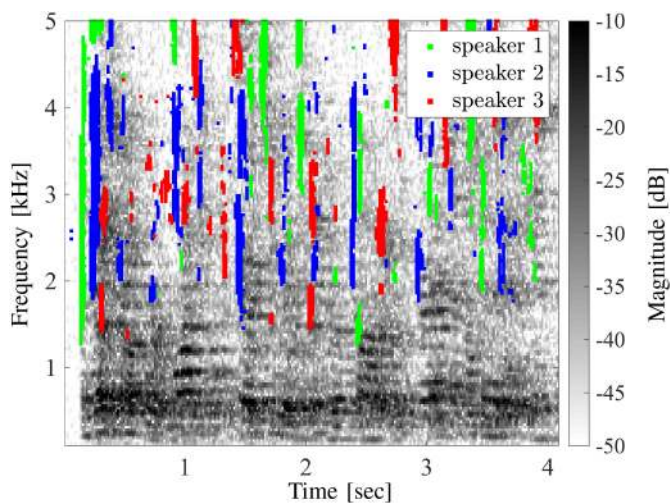


Fig. 12. Same as Fig. 8, but for the DPD-DES test.

coefficients, which are more distorted at the lower frequencies. Furthermore, note that the bins that passed the DPD-EDS test seem to be the most tightly grouped in the spectrogram, and located around speech segments that start after pauses in speech and which may hold mainly direct sound information. This is another possible indication to the accuracy of the test in correctly identifying direct sound regions in the spectrogram.

## VII. EXPERIMENTAL STUDY

In this section, a performance study of the different tests with real recorded data is presented. This data is taken from the LOCATA challenge [21]. In this challenge, several acoustic scenes of sound sources in a laboratory room of the Department of Computer Science at Humboldt University Berlin and of dimensions  $7.1 \times 9.8 \times 3$  m, were presented, with sound recorded using several arrays. The approximate reverberation time of the room is  $T_{60} = 0.55$  s. In this work, five scenarios from the LOCATA challenge with stationary sources and the Eigenmike [37] microphone array were used for the analysis. The number of sources, their distance from the array and their DOAs for each

TABLE IV  
THE NUMBER OF SOURCES, THEIR DISTANCE FROM THE ARRAY AND THEIR DOAs FOR THE FIVE RECORDINGS IN THE EXPERIMENTAL STUDY (TAKEN FROM THE LOCATA CHALLENGE)

Recording	Sources	Array distance	DOAs
1	1	2.26 m	(87.5°, 144.5°)
2	1	2.04 m	(86°, 28.5°)
3	1	1.46 m	(90.2°, 62.5°)
4	2	1.55 m 1.6 m	(90.5°, 51.5°) (91.5°, 66.5°)
5	4	2.49 m 1.93 m 2 m 2.55 m	(87.5°, 103.5°) (84.3°, 131.3°) (91°, 29.2°) (90.4°, 74.2°)

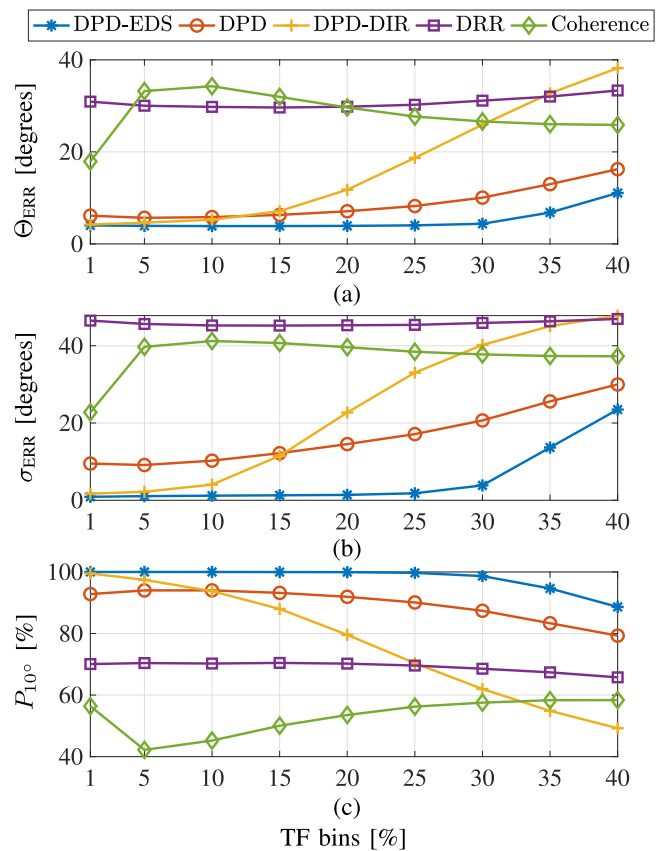


Fig. 13. (a)  $\Theta_{ERR}$ , (b)  $\sigma_{ERR}$  and (c)  $P_{10^\circ}$  as a function of the percentage of TF bins that passed each test for recordings 1–3 from the LOCATA challenge (see Table IV). These recordings include a single speaker. The results are averaged over the three recordings.

recording are presented in Table IV. The recorded signals have been down-sampled from 44 kHz to 16 kHz, and transformed to the STFT domain, from which the PWD coefficients, matrices  $\tilde{\mathbf{R}}_p(\tau, \omega)$  and  $\tilde{\mathbf{R}}_a(\tau, \omega)$ , were computed as described in Section VI. Array calibration was performed using the reference DOAs, to reduce potential measurement errors of the array and source positions. The measures  $\Theta_{ERR}$ ,  $\sigma_{ERR}$  and  $P_{10^\circ}$  were calculated for each recording and for each test, as in Section VI.

Recordings 1–3 that contain a single source are analyzed first. Fig. 13 presents the performance measures of all tests averaged over the three recordings, as a function of the percentage of TF bins that passed each test. The percentage values were controlled by adjusting the test thresholds, to support a common basis for

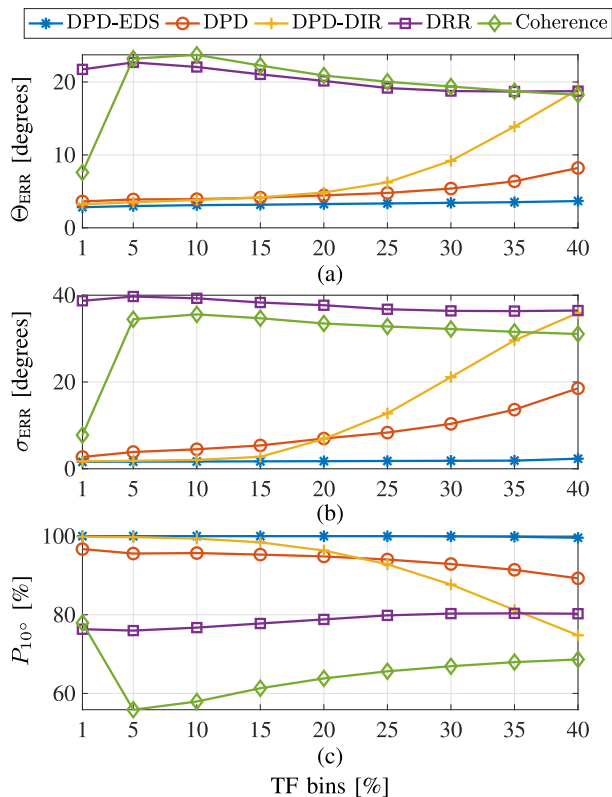


Fig. 14. (a)  $\Theta_{ERR}$ , (b)  $\sigma_{ERR}$  and (c)  $P_{10^\circ}$  as a function of the percentage of TF bins that passed each test for recording 4 from the LOCATA challenge (see Table IV). This recording includes two speakers.

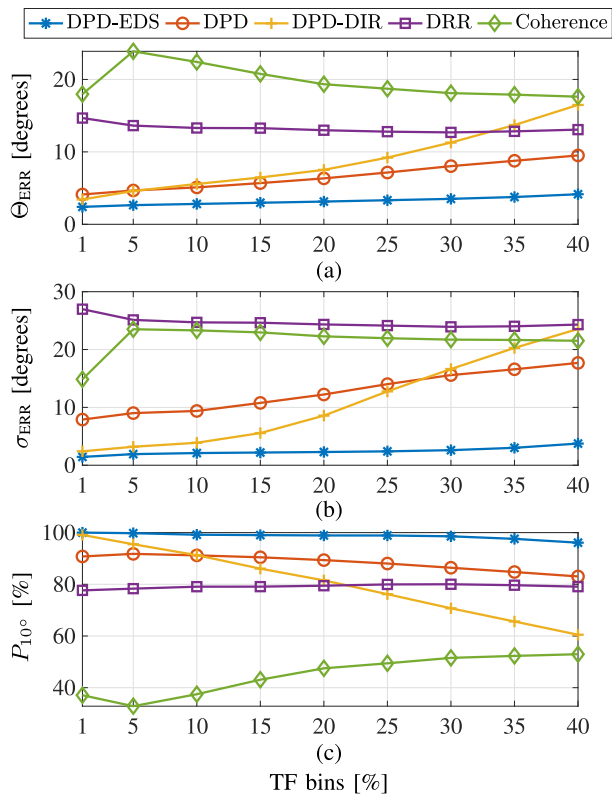


Fig. 15. (a)  $\Theta_{ERR}$ , (b)  $\sigma_{ERR}$  and (c)  $P_{10^\circ}$  as a function of the percentage of TF bins that passed each test for recording 5 from the LOCATA challenge (see Table IV). This recording includes four speakers.

comparison, as in the simulation study. The figure shows that the coherence test and the DRR based test achieve the worst results at the lower percentage range. The DPD and the DPD-DIR tests perform better, but the performance of the latter seems to degrade more significantly as the threshold decreases and more bins pass the test. The DPD-EDS test has the best performance, with the lowest  $\Theta_{ERR}$  and  $\sigma_{ERR}$ , and the highest  $P_{10^\circ}$  measure, in particular at the 10%–30% range.

Next, performance with recordings 4 and 5 that contain multiple speakers is analyzed. Figs. 14 and 15 present the performance measures of all tests for recording 4 and 5, respectively. Performance of all tests for recordings 4 and 5 seems similar to the performance presented in Fig. 13. Note that most tests achieve better performance measures in recording 4, compared to recordings 1–3 and 5. This may be explained by the relative proximity of the sources to the array in recording 4 compared to the other recordings. Thus, the direct sound may be more dominant in this recording. The DPD-EDS test achieves the best performance measures also for recordings 4 and 5. The results presented for the five recordings validate, on experimental data, the simulation studies and the theoretical foundations for the potentially high accuracy of the proposed method in estimating source DOAs. Finally, note that most tests achieve better performance measures in this experimental study compared to the simulation study, as in Fig. 6. This may be due to the fact that the reverberation time is significantly lower in this experimental study compared to the simulation study.

## VIII. CONCLUSIONS

In this work, a new method has been developed, based on the enhanced decomposition of the sound field, to identify TF bins that are dominated by the direct sound, and extract accurate source DOA estimates from these bins. The performance of the method is compared to the coherence test, the original DPD test, the DPD-DIR test, and a DRR based test. Extensive multiple-speaker computer simulations and an experimental study verified the superiority of the new DPD-EDS test, under different test conditions.

## REFERENCES

- [1] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
- [2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
- [3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. Berlin, Germany: Springer, 2008.
- [4] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, vol. 1, pp. 375–378.
- [5] E. D. D. Claudio, R. Parisi, and G. Orlandi, "Multi-source localization in reverberant environments by ROOT-MUSIC and clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 2, pp. II921–II924.
- [6] V. C. Raykar, B. Yegnanarayana, S. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 751–761, Sep. 2005.
- [7] 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.

- [8] X. Li, R. H. L. Girin, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1997–2012, Oct. 2017.
- [9] X. Li, B. Mourgue, L. Girin, S. Gannot, and R. Horaud, "Online localization of multiple moving speakers in reverberant environments," in *Proc. 10th IEEE Workshop Sensor Array Multichannel Signal Process.*, 2018, pp. 405–409.
- [10] M. Aktas and H. Ozkan, "Acoustic direction finding using single acoustic vector sensor under high reverberation," *Digit. Signal Process.*, vol. 75, pp. 56–70, 2018.
- [11] 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.
- [12] 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. ASSP-33, no. 4, pp. 823–831, Aug. 1985.
- [13] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 221–224.
- [14] 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.
- [15] 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.
- [16] B. Rafaely and K. Alhaiyani, "Speaker localization using direct path dominance test based on sound field directivity," *Signal Process.*, vol. 143, pp. 42–47, 2018.
- [17] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (A-SLAM) of a moving microphone array and its surrounding speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6–10.
- [18] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, Jul. 2015, pp. 1206–1210.
- [19] 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. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.
- [20] P. N. Samarasinghe, T. D. Abhayapala, H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "Estimating the direct-to-reverberant energy ratio using a spherical harmonics-based spatial correlation model," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 2, pp. 310–319, Feb. 2017.
- [21] 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*, Sheffield, U.K., Jul. 2018, pp. 410–414.
- [22] L. Madmoni and B. Rafaely, "Improved direct-path dominance test for speaker localization in reverberant environments," in *Proc. 26th Euro. Signal Process. Conf.*, 2018, pp. 2438–2442.
- [23] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation and Modulation Theory*, vol. 1. Hoboken, NJ, USA: Wiley, 2002.
- [24] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. II–1781.
- [25] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. II–1949.
- [26] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. Berlin, Germany: Springer, 2015.
- [27] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 221–224.
- [28] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, Sep. 2001.
- [29] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 5, pp. 2985–2988.
- [30] A. Brendel, C. Huang, and W. Kellermann, "STFT bin selection for localization algorithms based on the sparsity of speech signal spectra," in *Proc. Euronoise*, 2018, pp. 2561–2568.
- [31] S. Mohan, M. L. Kramer, B. C. Wheeler, and D. L. Jones, "Localization of nonstationary sources using a coherence test," in *Proc. IEEE Workshop Statist. Signal Process.*, 2003, pp. 470–473.
- [32] 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.
- [33] H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and W. Zhang, "Estimation of the direct-to-reverberant energy ratio using a spherical microphone array," *CoRR*, vol. abs/1510.08950, 2015.
- [34] 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.
- [35] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
- [36] *MATLAB and Statistics Toolbox Release 2018a*, The MathWorks, Inc., Natick, MA, USA, 2018.
- [37] *EM32 Eigenmike Microphone Array Release Notes (v17.0)*, mh acoustics LLC, Summit, NJ, USA, 2013.
- [38] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
- [39] J. S. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, vol. 10, no. 5, 1993.
- [40] D. L. Alon and B. Rafaely, "Spatial decomposition by spherical array processing," in *Parametric Time-Frequency Domain Spatial Audio*, V. Pulkki, S. Delikaris-Manias, and A. Politis, Eds. Hoboken, NJ, USA: Wiley, 2017.
- [41] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, 4th ed. Berlin, Germany: Wiley, Dec. 1999, p. 560.



**Lior Madmoni** (S'18) received the B.Sc. degree (*cum laude*) and the M.Sc. degree in electrical and computer engineering in 2016 and 2018, respectively, from Ben-Gurion University of the Negev, Beer-Sheva, Israel, where he is currently working toward the Ph.D. degree in electrical and computer engineering.

His current research interests include speech processing with microphone arrays under real-life conditions.

Mr. Madmoni is a recipient of the Ben-Gurion University High-Tech fellowship.



**Boaz Rafaely** (SM'01) received the B.Sc. degree (*cum laude*) in electrical engineering from Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 1986, the M.Sc. degree in biomedical engineering from Tel-Aviv University, Tel Aviv, Israel, in 1994, and the Ph.D. degree from the Institute of Sound and Vibration Research (ISVR), Southampton University, Southampton, U.K., in 1997.

At the ISVR, he was appointed Lecturer in 1997 and a Senior Lecturer in 2001, working on active control of sound and acoustic signal processing. In 2002, he spent six months as a Visiting Scientist with the Sensory Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, investigating speech enhancement for hearing aids. He then joined the Department of Electrical and Computer Engineering, Ben-Gurion University as a Senior Lecturer in 2003, and appointed as an Associate Professor in 2010, and a Professor in 2013. He is currently heading the acoustics laboratory, investigating methods for audio signal processing and spatial audio.

Prof. Rafaely was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, during 2010–2014, and since 2013, has been a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He is currently an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, *Acta Acustica United With Acustica*, and *IET Signal Processing*. During 2013–2016, he was the Chair of the Israeli Acoustical Association, and is currently chairing the Technical Committee on Audio Signal Processing in the European Acoustical Association. He was awarded the British Councils Clore Foundation Scholarship.