Perpendicular Cross-Spectra Fusion for Sound Source Localization with a Planar Microphone Array

Nikolaos Stefanakis, Despoina Pavlidi, *Student Member, IEEE* and Athanasios Mouchtaris, *Member, IEEE*

***Abstract*—Multiple sound source localization in reverberant environments stands as one of the most difficult challenges for many applications related to microphone array signal processing. In this paper, we describe Perpendicular Cross-Spectra Fusion (PCSF), a new Direction of Arrival (DOA) estimation algorithm which utilizes an analytic formula for direction estimation in the time-frequency (TF) domain. Inherent to this technique is the presence of multiple direction estimation subsystems which operate in parallel, producing a multiplicity of candidate DOAs at each TF point. We define a metric of coherence, based on the property of divergence of the different DOA estimators, for assessing the reliability of different signal portions, so that only TF bins with a high quality of directional information are exploited for local DOA estimation. The resulting collection of local DOAs is provided as input to a recently proposed histogram processing approach which is based on matching pursuit. Results based on simulation and real recordings illustrate the advantages of PCSF compared to other DOA estimation techniques subjected to the same histogram based processing, in the context of real-time multiple source localization and counting; improved performance in reverberant conditions and high tolerance to diffuse and common mode noise.**

***Index Terms*—Direction of arrival estimation, multiple source localization, source counting, information fusion**

EDICS: AUD-ASAP:Acoustic Sensor Array Processing

I. INTRODUCTION

RELYING on acoustic signals for estimating the Direction of Arrival (DOA) of one or more sound sources is a central process to many sensor array processing applications, as for example in surveillance, source separation, speech enhancement, teleconferencing and hearing aids. Over the years, many different techniques have been proposed for sound source localization, yet it is difficult to say that one particular approach guarantees the best performance in all cases. Different techniques may be more or less suitable than others, in accordance to the characteristics of the radiating sound sources (e.g., musical sources or speech sources), of the environment (noisy, reverberant), of the sensor array geometry, or in accordance to the limitations introduced by the available computational resources.

Localizing multiple sound sources whose locations and number may change arbitrarily in time has been efficiently tackled by techniques which exploit the property of disjointness of the sound sources in the time-frequency (TF) domain. In several occasions this is fulfilled by the fact that the sources are sparse in this domain, and that, as a consequence, one source is dominant over the others in some time-frequency windows or "zones". A regularly adopted assumption in accordance to this principle is the so-called W-Disjoint Orthogonality (WDO) assumption [1], which states that in each TF component, at most one source is active. As a consequence, the DOA estimation problem can be written independently at each TF component, and the resulting collection of estimated local DOAs can be processed in terms of a histogram [2]–[5] or by using a clustering approach [6], [7].

WDO is approximately satisfied by speech signals in anechoic environments, but not in reverberant conditions. Adopting a more relaxed assumption about the source disjointness than WDO, a large number of techniques propose to collect DOA information only from portions of the signal where there is evidence that the overlap between different sound sources is minimal. In this direction, Mohan et al in [8] detects single source TF bins by observing the effective rank of the time-averaged covariance matrix, an approach also followed by the authors in [9], [10]. In a slightly different manner, the authors in [11]–[13] propose the use of a single source confidence measure which is defined across zones of consecutive frequency bins. Finally, discarding of erroneous local DOAs based on a measure of the consistency error between the steering vector corresponding to the most likely DOA and the actual measurement has been proposed in [14].

Although not treated as a separate case in many of the previous works, robustness to noise and reverberation is an important prerequisite for sound source localization. In order to better handle these problems, the localization process can be improved by incorporating an estimation of noise or reverberation at each TF point, as shown for example in [15]. On the other hand, onset detection and noise floor tracking have been proposed as supplementary processes to assist the selection of signal portions which are less contaminated by noise and/or reverberation than others [9], [14].

Additional works addressing localization of multiple, simultaneous active sound sources are based on the well known MUltiple SIgnal Classification (MUSIC) algorithm, with its narrowband [16], [17] and wideband variations [18], [19]. Recently, it was shown that MUSIC performance in diffuse noise conditions may be improved by exploiting the symmetries in particular types of array geometries, with the scope to denoise

the spatial covariance matrix [20]. Similar to this work, the technique proposed in this paper exploits symmetries of the microphone array with the purpose to derive a second-order statistical measure which is neutralized with respect to the signature of diffuse isotropic noise. However, our approach relies on a completely different way of exploiting the local signal covariance matrix to that of MUSIC.

Other than accurate and efficient DOA estimation, an extremely important issue in sound source localization is estimating the number of active sources at each time instant, known as source counting. Many methods in the literature propose estimating the intrinsic dimension of the recorded data, i.e., for an acoustic problem, they perform source counting at each time instant. Most of them are based on information theoretic criteria (see [21] and the references within). In other methods, the estimation of the number of sources is derived from a large set of DOA estimates that need to be clustered. In classification, some approaches to estimating both the clusters and their number have been proposed (e.g., [22]), while several solutions specially dedicated to DOAs have been tackled in [23]–[26].

In this paper, the problem of multiple sound source localization and counting is addressed using a novel DOA estimation technique named Perpendicular Cross Spectra Fusion (PCSF). PCSF operates on a smoothed – in time and frequency – observation of the local signal covariance matrix to establish an analytic relation between two different cross-spectra terms and the incident acoustic direction. Inherent to this technique is the presence of multiple direction estimation subsystems which operate in parallel, producing a multiplicity of candidate DOAs at each TF point. We illustrate that the different subsystems are complementary to one another and furthermore, their outputs tend to diverge for signal portions conveying poor directional information. We define a metric of coherence, based on the property of divergence of the different DOA estimators, for assessing the reliability of different signal portions, so that only TF bins with a high directional information gain are exploited for local DOA estimation. Ultimately, as the result of fusion may be a void DOA, the proposed process may be seen as an efficient TF point selection method with significant potential to improve the performance of a multiple sound source localization and counting system deployed on a planar microphone array with a square canonical configuration.

While the proposed algorithm for constructing the local collection of DOAs is essentially novel, the additional steps required for completing the process are based on a recently proposed method presented in [13]. Specifically, we apply a matching pursuit-based approach to the histogram of DOA estimates, which allows for joint estimation of the sound sources' number and their corresponding directions. The particular method has demonstrated excellent performance in adverse conditions, outperforming several other approaches, both in terms of accuracy and computational complexity [13], [27]. In this paper it becomes useful as a common framework to process the DOAs which are provided by the proposed method, as well as those provided by other methods that we use for comparison.

Additional value to this paper is given by the fact that the previously described histogram processing method is evaluated in the context of additional DOA estimation algorithms. In particular, in the original paper [13], the authors used the method of the Circular Integrated Cross Spectrum (CICS) [28] in combination with the single source zone (SSZ) confidence measure described in [11] to produce the required collection of direction estimates. In comparison to this approach, we illustrate that a DOA estimation algorithm based on Directional Audio Coding (DIRAC) [29] results to a competitive performance and may even lead to improvements under certain conditions, at least for the type of microphone array which is considered here. In parallel, it is demonstrated that the diffuseness estimation process inherent to a two-dimensional implementation of DIRAC provides an efficient metric for selecting reliable TF points for DOA estimation in 2D.

The structure of this paper is as follows; in Section II, we present the basic principles and the requirements that need to be fulfilled for the method to be applicable to the given microphone array topology. In Section III, we describe four additional state-of-the-art methods for DOA estimation with the scope to use them for comparison with the proposed technique. The common framework for sound source localization and counting using an available collection of local DOAs is presented in Section IV, results based on simulated and real data are presented in Section V and finally, we conclude in Section VI.

II. PROPOSED METHOD

A. Assumptions and Notations

Throughout the rest of the paper, $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ denote complex conjugation, transposition and Hermitian transposition respectively, while $\Im[\cdot]$ and $\Re[\cdot]$ denote the imaginary and real part of a complex number. Signals are presented in the TF domain with $\omega \in \mathbb{R}$ denoting the angular frequency and $\tau \in \mathbb{Z}$ the time-frame index.

Consider a planar array of M sensors and let $\mathbf{r}_m \in \mathbb{R}^2$, $m = 1, ..., M$ denote the vector with the coordinates of each sensor. Now, let $\mathbf{x}$ be a vector in $\mathbb{R}^2$ and $l_{\mathbf{x}} = \|\mathbf{x}\|_2$ be its length, with $\|\cdot\|_2$ denoting the Euclidean norm. Considering now all pairwise sensor combinations, ij, we define the set $\Omega(\mathbf{x}) = \{ij : \mathbf{r}_j - \mathbf{r}_i = \mathbf{x}\}$ and let $N_{\mathbf{x}} = |\Omega(\mathbf{x})|$ denote the cardinality of that set. This set contains all sensor combinations which form line segments that have the same direction and length as $\mathbf{x}$. In practice $\mathbf{x}$ is not arbitrarily chosen, but in accordance to the array topology. Now let $p_m(\tau, \omega)$ denote the signal received at the mth sensor at time τ and radial frequency ω. Assuming that the set $\Omega(\mathbf{x})$ is not empty, we define the complex cross-spectra relevant to $\mathbf{x}$ as

$$\Phi_{\mathbf{x}}(\tau, \omega) = \frac{1}{N_{\mathbf{x}}} E\{ \sum_{ij \in \Omega(\mathbf{x})} p_i(\tau, \omega) p_j^*(\tau, \omega) \}, \qquad (1)$$

where $E\{\cdot\}$ denotes expectation. It is easy to observe that if there is a non-empty set $\Omega(\mathbf{x})$, then neither $\Omega(-\mathbf{x})$ is empty and moreover, the property $\Phi_{\mathbf{x}}(\tau, \omega) = \Phi^*_{-\mathbf{x}}(\tau, \omega)$ holds.

The presented approach can be applied on two-dimensional microphone arrays of specific geometry; it is required that among all the line segments that connect any sensor pair, there

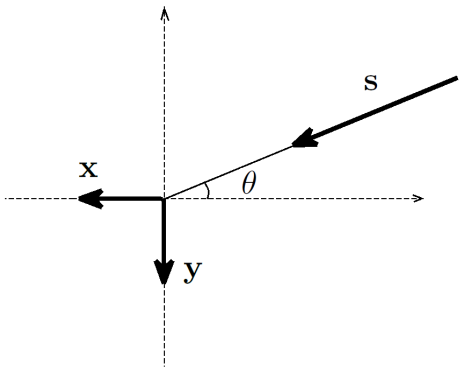

Fig. 1. Coordinate system and vector notation.

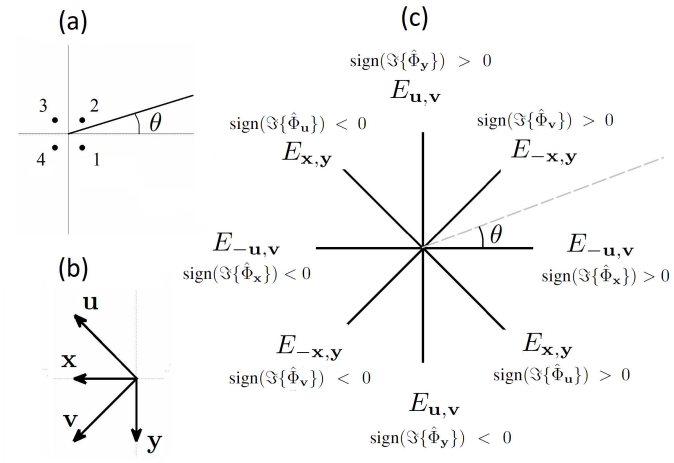

Fig. 2. Square sensor configuration in (a) and corresponding vectors in (b). Maximum Robustness Axes and corresponding disambiguation criteria for each one of the four estimators rising from the case of a square array in (c)

are line segments which are perpendicular to one another and have the same length. Based on the previous definitions, this requirement can be expressed in terms of two sets $\Omega(\mathbf{x})$ and $\Omega(\mathbf{y})$ such that $\mathbf{x} \perp \mathbf{y}$ and $l_{\mathbf{x}} = l_{\mathbf{y}}$. A square microphone array of four sensors represents the simplest configuration required for deploying the method in its full extent, and it will be the basic configuration which will be used to demonstrate the technique in this paper. Additional requirements are that the distance of the sound sources with respect to the center of the array is large enough so that the far field assumption holds and that the maximum frequency of analysis is within the limits imposed by spatial aliasing.

B. Perpendicular Cross Spectra Difference Model

Assume now that the array receives a signal from a single acoustic source at an unknown direction in the presence of additive isotropic noise. Let $s(\tau,\omega)$ be the source signal and $\mathbf{s}$ be the unit norm vector pointing from the source to the center of the sensor array (we remind that the far field assumption is required to hold). The observed signal at the mth sensor can be written as

$$p_m(\tau,\omega) = s(\tau,\omega)d_m(\omega,\theta) + h_m(\tau,\omega), \tag{2}$$

where $d_m(\omega,\theta) = e^{-j\omega\delta_m}$ is the transfer function for the mth microphone with δ_m denoting the time of flight from the source to that microphone and $h_m(\tau,\omega)$ is the noise component at the same microphone. Now, assuming that the previous requirements are fulfilled for two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^2$ such that $\mathbf{x} \perp \mathbf{y}$ and $l_{\mathbf{x}} = l_{\mathbf{y}}$, we may derive a model for the cross-spectra along $\mathbf{x}$

$$\Phi_{\mathbf{x}}(\tau,\omega) = \Phi_{ss}(\tau,\omega)e^{jk\mathbf{x}^T\mathbf{s}} + \Psi_{l_{\mathbf{x}}}(\tau,\omega), \tag{3}$$

and along the perpendicular direction of $\mathbf{y}$ as

$$\Phi_{\mathbf{y}}(\tau,\omega) = \Phi_{ss}(\tau,\omega)e^{jk\mathbf{y}^T\mathbf{s}} + \Psi_{l_{\mathbf{y}}}(\tau,\omega), \tag{4}$$

where $k = \omega/c$ is the wavenumber with c denoting the speed of sound, $\Phi_{ss}(\tau,\omega) = E\{s(\tau,\omega)s^*(\tau,\omega)\}$ is the signal power spectrum and $\Psi_{l_{\mathbf{x}}}$, $\Psi_{l_{\mathbf{y}}}$ are the diffuse noise components of the cross-spectra. The subscript $l_{\mathbf{x}}$ and $l_{\mathbf{y}}$ are used here to denote the well known fact that in an isotropic noise field, the second-order statistics between two measurement points are only dependent on the distance between the two points [30]. This means that the noise components are equal in both cross-spectra terms since $l_{\mathbf{x}} = l_{\mathbf{y}}$. Of particular interest in this paper is the Perpendicular Cross-Spectra Difference (PCSD) defined as

$$\begin{aligned} \Delta\Phi_{\mathbf{x},\mathbf{y}}(\tau,\omega) &= \Phi_{\mathbf{x}}(\tau,\omega) - \Phi_{\mathbf{y}}(\tau,\omega), \\ &= \Phi_{ss}(\tau,\omega)(e^{jk\mathbf{x}^T\mathbf{s}} - e^{jk\mathbf{y}^T\mathbf{s}}), \end{aligned} \tag{5}$$

which has the very interesting property that the isotropic noise components vanish. Considering now that a spherical-isotropic noise field represents well the second-order statistics of the reverberant part of the sound signals in a reverberant environment [30], [31], we expect that a sound source localization method relying on an approximation of PCSD should exhibit increased robustness to reverberation, which is actually one of the most valuable properties of the proposed method.

C. Relating the Measured PCSD to the DOA

In this subsection, we provide an analytical relation between PCSD and the incident acoustic direction. While PCSD is defined on a specific TF point, in what follows, we omit the time-frame, τ, and frequency index, ω, for convenience. Without loss of generality, we may assume that the $\mathbf{x}$ and $\mathbf{y}$ vectors are parallel to the x and y axes of our coordinate system, as shown in Figure 1. Let's also symbolize with θ the angle of the incident wave and let $d = l_{\mathbf{x}} = l_{\mathbf{y}}$ be the length of $\mathbf{x}$ and $\mathbf{y}$ respectively. We may then define a model of the PCSD as a function of the incident angle as

$$\begin{aligned} \Delta\Phi_{\mathbf{x},\mathbf{y}} &= \Phi_{ss} \cdot (\cos(kd\cos(\theta)) - \cos(kd\sin(\theta))) \\ &+ j\Phi_{ss} \cdot (\sin(kd\cos(\theta)) - \sin(kd\sin(\theta))). \end{aligned} \tag{6}$$

Consider now the ratio between the real and imaginary part of Eq. (6); this may be expressed as

$$\begin{aligned} \frac{\Re\{\Delta\Phi_{\mathbf{x},\mathbf{y}}\}}{\Im\{\Delta\Phi_{\mathbf{x},\mathbf{y}}\}} &= \frac{\cos(kd\cos(\theta)) - \cos(kd\sin(\theta))}{\sin(kd\cos(\theta)) - \sin(kd\sin(\theta))} \\ &= -\tan(\tfrac{kd}{2}(\cos(\theta) + \sin(\theta))), \\ &= -\tan(\tfrac{kd}{\sqrt{2}}\sin(\theta + \tfrac{\pi}{4})). \end{aligned} \tag{7}$$

Equation (7) relates the incident angle θ to the cross-spectra terms $\Phi_{\mathbf{x}}$ and $\Phi_{\mathbf{y}}$. We can thus exploit Eq. (7) in order to derive a closed-form solution for the unknown angle θ based on an estimation of $\Phi_{\mathbf{x}}$ and $\Phi_{\mathbf{y}}$ through Eq. (1). In practice, some additional information and effort is required for reaching this point because this operation involves inverse trigonometric functions which are multivalued and highly nonlinear.

For the needs of DOA estimation, the first step required is to define the auxiliary observation

$$z_{\mathbf{x},\mathbf{y}} = \frac{\sqrt{2}}{kd} \tan^{-1}\left(\frac{\Re\{\Delta\hat{\Phi}_{\mathbf{x},\mathbf{y}}\}}{\Im\{\Delta\hat{\Phi}_{\mathbf{x},\mathbf{y}}\}}\right), \tag{8}$$

where $\tan^{-1}(\cdot)$ is the inverse tangent function and $\Delta\hat{\Phi}_{\mathbf{x},\mathbf{y}} = \hat{\Phi}_{\mathbf{x}} - \hat{\Phi}_{\mathbf{y}}$ involves the local observed cross-spectra along $\mathbf{x}$ and $\mathbf{y}$, obtained by averaging in the neighborhood of a particular TF point. Observe now from Eq. (7) that the auxiliary observation defined in Eq. (8) can be associated to the incident direction θ through

$$-\sin(\theta + \frac{\pi}{4}) \leftarrow z_{\mathbf{x},\mathbf{y}}, \tag{9}$$

which brings us one step closer to an estimation of θ. Now, some issues need to be clarified because in practice, the actual observations might not be consistent with respect to the model. For example, when using the function $\tan^{-1}(\cdot)$, most computer programs will return a solution q such that $q \in (-\pi/2, \pi/2)$, and depending on the value of $\frac{\sqrt{2}}{kd}$, the auxiliary observation might not lie in the range $[-1, 1]$ which is meaningful to us according to Eq. (9). A simple way for treating this problem is to completely disregard auxiliary observations which are not within $[-1, 1]$. Furthermore, when using $\tan^{-1}(\cdot)$ to derive a value q, we ignore the additional solutions $q + k\pi$, $k \in \mathbb{Z}$, although these values may still lie in the $[-1, 1]$ range. The considered model actually dictates that cases corresponding to $k \neq 0$ can be ignored as long as the maximum frequency of investigation is not higher than $f_A = \frac{c}{2\sqrt{2}d}$, which may be seen as an upper limit related to spatial aliasing.

The auxiliary observation $z_{\mathbf{x},\mathbf{y}}$ together with function $-\sin(\theta + \frac{\pi}{4})$ in Eq. (9) define an ambiguous closed-form solution to the unknown direction θ, the ambiguity resulting from the fact that function $\sin^{-1}(\cdot)$ returns two possible angles in $[-\pi, \pi)$. Interestingly, an additional closed-form solution may be constructed, if instead of the cross-spectra measure $\hat{\Phi}_{\mathbf{x}}$ [resp. $\hat{\Phi}_{\mathbf{y}}$] we employ $\hat{\Phi}_{-\mathbf{x}}$ [resp. $\hat{\Phi}_{-\mathbf{y}}$]. We let the reader verify that had we replaced $\mathbf{x} = [-1, 0]$ with its opposite one, $-\mathbf{x} = [1, 0]$, to measure $\Delta\hat{\Phi}_{-\mathbf{x},\mathbf{y}}$, we would end up with an additional ambiguous closed-form solution of the form

$$\sin(\theta - \pi/4) \leftarrow z_{-\mathbf{x},\mathbf{y}}, \tag{10}$$

with the auxiliary observation in this case defined as

$$z_{-\mathbf{x},\mathbf{y}} = \frac{\sqrt{2}}{kd} \tan^{-1}\left(\frac{\Re\{\Delta\hat{\Phi}_{-\mathbf{x},\mathbf{y}}\}}{\Im\{\Delta\hat{\Phi}_{-\mathbf{x},\mathbf{y}}\}}\right). \tag{11}$$

As it is shown in the next subsection, the two systems $\{\mathbf{x}, \mathbf{y}\}$ and $\{-\mathbf{x}, \mathbf{y}\}$ are not equivalent, and they lead to angle estimators with very different properties.

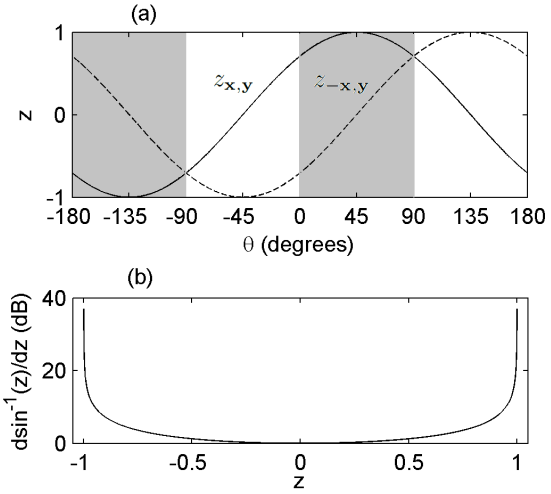

Fig. 3. Theoretical relation between DOA and auxiliary observation for system $\mathbf{x}, \mathbf{y}$ (solid line) and system $-\mathbf{x}, \mathbf{y}$ (dashed line) in (a). Error sensitivity of any of the two estimators as a function of the actual value of the auxiliary observation is shown in logarithmic scale in (b).

D. Behaviour of the Estimators

In this subsection, we illustrate that for each system $\{\mathbf{x}, \mathbf{y}\}$, a complementary ambiguous estimator based on $\{-\mathbf{x}, \mathbf{y}\}$ (or $\{\mathbf{x}, -\mathbf{y}\}$) can be constructed and moreover, the two estimators have very different properties. Based on Eqs. (9) and (10), the theoretical relation between the direction θ and the value of the auxiliary variable can be seen for each one of the two approaches in Figure 3. The plot in (a) reflects the ambiguity associated with each one of the two angle estimation systems as for each value of $z \in (-1, 1)$ there are always two possible angles θ. In principle, if we treat $z_{\mathbf{x},\mathbf{y}}$ and $z_{-\mathbf{x},\mathbf{y}}$ as the independent variables, the unknown direction $\hat{\theta}$ can be estimated as

$$\begin{array}{rcl} \hat{\theta}_i & = & \sin_i^{-1}(-z_{\mathbf{x},\mathbf{y}}) - \pi/4, \\ & \text{and} & \\ \hat{\theta}_i & = & -\sin_i^{-1}(-z_{-\mathbf{x},\mathbf{y}}) + \pi/4, \end{array} \tag{12}$$

for the first and second system respectively, with subscript $i \in \{1, 2\}$ here denoting that $\sin^{-1}(\cdot)$ is actually a multivalued function. In the range $[-\pi, \pi)$, two possible directions $\hat{\theta}_1$ and $\hat{\theta}_2$ will rise for any of the two approaches, related through $\hat{\theta}_2 = \pi/2 - \hat{\theta}_1$ for $\{\mathbf{x}, \mathbf{y}\}$ and through $\hat{\theta}_2 = -\pi/2 - \hat{\theta}_1$ for $\{-\mathbf{x}, \mathbf{y}\}$. It will be shown in the next subsection that this ambiguity may be easily resolved by exploiting additional information from the observed cross-spectra terms.

At this stage, it is worth performing a basic analysis in order to quantify the reliability of each angle estimation system with respect to the actual value of the auxiliary variable when using Eqs. (9) and (10). We treat here θ as the dependent variable considering that the auxiliary observation z is perturbed from its actual values by zero-mean random errors. The sensitivity of any of the two estimators in Eqs. (9) and (10) to these errors can be quantified in terms of the derivative

$$\frac{d(\sin^{-1}(z))}{dz} = \frac{1}{\sqrt{1 - z^2}}, \quad -1 \leq z \leq 1, \tag{13}$$

	$E_{\mathbf{x},\mathbf{y}}$	$E_{-\mathbf{x},\mathbf{y}}$	$E_{\mathbf{u},\mathbf{v}}$	$E_{-\mathbf{u},\mathbf{v}}$
MRA	$y=-x$	$y=x$	$x=0$	$y=0$
Solution+	$\theta_1=\text{asin}(-z_{\mathbf{x},\mathbf{y}})-\frac{\pi}{4}$	$\theta_1=-\text{asin}(-z_{-\mathbf{x},\mathbf{y}})+\frac{\pi}{4}$	$\theta_1=\text{asin}(-z_{\mathbf{u},\mathbf{v}})-\frac{\pi}{2}$	$\theta_1=-\text{asin}(-z_{-\mathbf{u},\mathbf{v}})$
Solution-	$\theta_2=-\text{asin}(-z_{\mathbf{x},\mathbf{y}})+\frac{3\pi}{4}$	$\theta_2=\text{asin}(-z_{-\mathbf{x},\mathbf{y}})-\frac{3\pi}{4}$	$\theta_2=-\text{asin}(-z_{\mathbf{u},\mathbf{v}})+\frac{\pi}{2}$	$\theta_2=\text{asin}(-z_{-\mathbf{u},\mathbf{v}})-\pi$
Disamb. crit.	$\text{sign}(\Im\{\hat{\Phi}_{\mathbf{u}}\})$	$\text{sign}(\Im\{\hat{\Phi}_{\mathbf{v}}\})$	$\text{sign}(\Im\{\hat{\Phi}_{\mathbf{y}}\})$	$\text{sign}(\Im\{\hat{\Phi}_{\mathbf{x}}\})$

TABLE I
MAXIMUM ROBUSTNESS AXIS, SOLUTION AND DISAMBIGUATION CRITERIA FOR EACH ONE OF THE FOUR ESTIMATORS CORRESPONDING TO THE SQUARE SENSOR ARRAY.

where the characteristic subscripts $(\cdot)_{\mathbf{x},\mathbf{y}}$ and $(\cdot)_{-\mathbf{x},\mathbf{y}}$ are here removed for convenience. The error sensitivity as a function of the actual value z is shown in logarithmic scale in Fig. 3(b), where it can be seen that a system is most reliable when its auxiliary observation is close to 0. This basically reflects nothing else than the fact that function $\sin^{-1}(z)$ has the largest gradient when z is close to ± 1. Combining information from both Figs. 3(a) and 3(b) indicates that Eq. (9) is completely unreliable for estimating the impinging angle when the source is at $\theta = 45^o$ or at -135^o, while at the same time, Eq. (10) is least sensitive to errors at these angles. Vice-versa, Eq. (10) is completely unreliable for estimating the direction when the source is at $\theta = -45^o$ or at 135^o, angles for which the first estimator shows maximum robustness. In this sense, the line $y = -x$ defines a Maximum Robustness Axis (MRA) for system $\{\mathbf{x},\mathbf{y}\}$ while the line $y = x$ is a MRA for system $\{-\mathbf{x},\mathbf{y}\}$. Extending these observations to the entire angle range, the white regions in Fig. 3(a) indicate the angular sectors where system $\{\mathbf{x},\mathbf{y}\}$ is more robust than $\{-\mathbf{x},\mathbf{y}\}$, while the gray regions indicate the angular sectors where system $\{-\mathbf{x},\mathbf{y}\}$ is more robust than $\{\mathbf{x},\mathbf{y}\}$. Observe that in the gray [resp. white] regions $|z_{\mathbf{x},\mathbf{y}}| < |z_{-\mathbf{x},\mathbf{y}}|$ [resp. $|z_{-\mathbf{x},\mathbf{y}}| < |z_{\mathbf{x},\mathbf{y}}|$] holds.

E. Application With a Square Array

In this subsection, we adapt the presented methodology to the case of a planar array comprised of four sensors placed at the vertices of a square, as shown in Fig. 2(a). Letting d be the side length of the square, the particular configuration allows the definition of four vectors; $\mathbf{x} = [-d, 0]$, $\mathbf{y} = [0, -d]$, $\mathbf{u} = [-d, d]$ and $\mathbf{v} = [-d, -d]$, all of which are shown in Figure 2(b). The orientations of these vectors are such that they allow us to construct four DOA estimation systems; the pre-defined requirements are fulfilled in the sense that $\mathbf{x} \perp \mathbf{y}$, $l_{\mathbf{x}} = l_{\mathbf{y}} = d$, and $\mathbf{u} \perp \mathbf{v}$, $l_{\mathbf{u}} = l_{\mathbf{v}} = \sqrt{2}d$ hold. Along these directions, we may define four cross-spectra terms

$$\Phi_{\mathbf{x}} = 0.5(E\{p_1 p_4^* + p_2 p_3^*\}), \quad (14)$$

$$\Phi_{\mathbf{y}} = 0.5(E\{p_2 p_1^* + p_3 p_4^*\}), \quad (15)$$

$$\Phi_{\mathbf{u}} = E\{p_1 p_3^*\}, \quad (16)$$

$$\Phi_{\mathbf{v}} = E\{p_2 p_4^*\}, \quad (17)$$

and use them to compose the two PCSDs $\Delta\Phi_{\mathbf{x},\mathbf{y}}$ and $\Delta\Phi_{\mathbf{u},\mathbf{v}}$ together with their mirror-symmetric ones $\Delta\Phi_{-\mathbf{x},\mathbf{y}}$ and $\Delta\Phi_{-\mathbf{u},\mathbf{v}}$. The link between the four auxiliary observations $z_{\mathbf{x},\mathbf{y}}$, $z_{-\mathbf{x},\mathbf{y}}$, $z_{\mathbf{u},\mathbf{v}}$ and $z_{-\mathbf{u},\mathbf{v}}$ and their corresponding closed-form solutions is shown in Table 1. While the lines $y = -x$ and $y = x$ correspond once more to the MRAs of systems $\{\mathbf{x},\mathbf{y}\}$ and $\{-\mathbf{x},\mathbf{y}\}$, we now additionally have the lines $y = 0$ and $x = 0$ defining the MRAs of systems $\{\mathbf{u},\mathbf{v}\}$ and $\{-\mathbf{u},\mathbf{v}\}$. Furthermore, the $\Phi_{\mathbf{x}}$ and $\Phi_{\mathbf{y}}$ terms in Eqs. (14) and (15) are averaged along two sensors pairs, which may potentially result to some increased robustness for systems $\{\mathbf{x},\mathbf{y}\}$ and $\{-\mathbf{x},\mathbf{y}\}$ as opposed to systems $\{\mathbf{u},\mathbf{v}\}$ and $\{-\mathbf{u},\mathbf{v}\}$, whose cross-spectra terms are calculated along single sensor pairs.

A problem that still needs to be resolved is the ambiguity related to the $\sin^{-1}(\cdot)$ function. Assume, for example, that the actual direction of the incident plane wave is at 5^o, then both 5^o and 175^o are possible solutions for system $\{-\mathbf{u},\mathbf{v}\}$. One straightforward way to decide which one is the true angle would be to use acoustic beamformering, by observing for which of the two possible directions the beamformer response is greater. For the case of the square array however, we propose another approach which is based on the observation that the two possible solutions indicate flow of acoustic energy along almost opposite directions (this is actually valid for any estimator, as long as the incident angle is close to its corresponding MRA). Due to this property, a straightforward way to detect the correct direction is by observing the sign of the cross-spectra term which is aligned with respect to the MRA of each estimator. In this same example, $\Im[\hat{\Phi}_{\mathbf{x}}]$ should be positive [resp. negative] if 5^o [resp. 175] is the true angle. In fact, for the case of the square sensor array, we may associate one disambiguation cross-spectra term to each system and use its sign in order to disambiguate as shown in Fig. 2 and in the third row of Table I. This approach exploits the well known advantage that the imaginary parts of the cross-spectra terms are immune to isotropic noise [30].

Using all the available systems, it is obvious that we may obtain up to four different angle estimations at each TF point. For system $\{-\mathbf{x},\mathbf{y}\}$ as an example, provided that $|z_{-\mathbf{x},\mathbf{y}}| < 1$, the full process for finding a DOA according to Table I is

$$\hat{\theta}_{-\mathbf{x},\mathbf{y}} = \begin{cases} -\text{asin}(-z_{-\mathbf{x},\mathbf{y}}) + \pi/4, & \text{if } \text{sign}(\Im[\hat{\Phi}_{\mathbf{v}}]) > 0 \\ \text{asin}(-z_{-\mathbf{x},\mathbf{y}}) - 3\pi/4, & \text{if } \text{sign}(\Im[\hat{\Phi}_{\mathbf{v}}]) < 0. \end{cases} \quad (18)$$

It should be noted that in Eq. (18), function $\text{asin}(\cdot)$ is different from $\sin^{-1}(\cdot)$ in the sense that it returns only one solution in $[-\pi/2, \pi/2]$. This convention implies that the final estimate returned by each system will be a real number in $(-\pi, \pi]$.

F. Property of Divergence of the Estimators

Following the analysis in Section II-D, it is interesting at this point to observe certain facts associated with the ensemble of the four DOA estimators rising from the case of a square array. In Fig. 4 we illustrate the distribution of the candidate DOAs, for each one of the four estimators, when the microphone signals are spherical isotropic noise (top row) and spatially

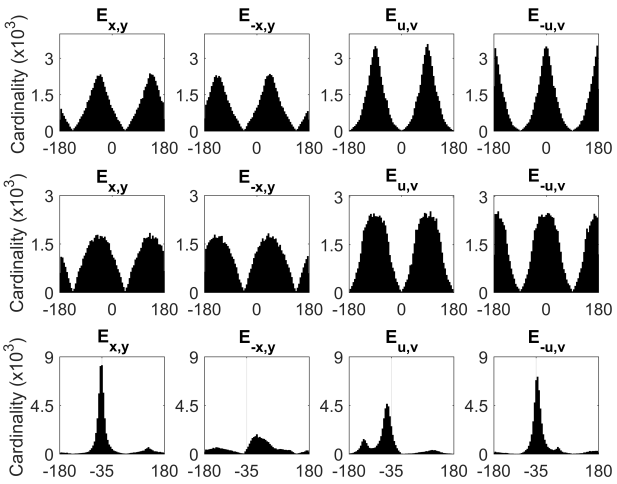

Fig. 4. Histogram with the candidate DOAs as obtained from for each one of the four estimators rising from the case of the square array for spherical isotropic noise (top row), spatially white noise (middle row) and single acoustic source at -35^o inside a reverberant room (bottom row).

white noise (middle row), while in the bottom row we illustrate each estimator's response for a single speech source at -35^o at noiseless conditions inside a reverberant room. The histograms are obtained by accumulating all DOAs up to 4500 Hz. A first thing to observe from the first two rows of Fig. 4 is that the highest cardinality values appear at directions which coincide with the MRA of each estimator. By first glance, this seems to be a negative attribute, implying that the estimators are biased, favouring the selection of certain angle regions than others. Observe however that for each estimator, the DOAs tend to concentrate at different parts of the x-axis, meaning that the estimators responses diverge under noisy input. Intuitively, this property can be linked to the case of an ensemble of classifiers, where divergence is a important prerequisite for improving classification and decision making in general [32], [33].

In contrast to the case of noisy input, observe that the candidate DOAs for three out of the four estimators concentrate near the actual acoustic direction in the lower row of Fig. 4. In line with the sensitivity analysis presented in Section II-D, it can be visually verified that $E_{\mathbf{x},\mathbf{y}}$ exhibits the clearest cardinality peak at -35^o as the actual acoustic source direction is close to the MRA of this estimator. On the contrary, system $E_{-\mathbf{x},\mathbf{y}}$ induces large errors as the true DOA is at an angle which is almost perpendicular to its corresponding MRA. Finally, the majority of the local DOAs for systems $E_{\mathbf{u},\mathbf{v}}$ and $E_{-\mathbf{u},\mathbf{v}}$ also concentrate on the true angle. These observations dictate that the estimators' outputs are expected to diverge for signal components lacking direction but at the same time, two or three of the estimators' outputs are expected to converge to the true acoustic direction in the presence of a dominant directional component. This contradiction is exploited in the fusion process which is described in more detail below, by using a metric of convergence for assessing the reliability of different signal portions, so that only TF bins with a high quality of directional information are exploited for local DOA estimation.

G. DOA Fusion

Extending the analysis to array topologies with a minimum of four sensors, in what follows we assume that we have a coincident sensor array generating $J \geq 4$ PCSDs corresponding to J unambiguous DOA estimation subsystems. It is obvious that such a system generates a great redundancy of information as it may produce up to J candidate DOAs at each TF point. From all this quantity of information however, only a small portion is required for inferring the sound source locations as well as their activities along time.

Exploiting this redundancy, we define a fusion process which operates on the collection of candidate DOAs to calculate a consistency metric. Depending on this metric, we decide whether to output a local DOA specific to the TF point of analysis. We then collect local DOAs into groups of consecutive frequency bins and apply an almost similar fusion process on each group in order to estimate a group-specific DOA, which represents the highest-level DOA information.

1) Angular Consistency Metric: Assume that we have a set of $N_E \geq 2$ DOA estimations, $\Phi = \{\phi_1, ..., \phi_i, ..., \phi_{N_E}\}$ with $\phi_i \in [-\pi, \pi), \forall i$. We define the *overall consistency* of a collection of estimates Φ as

$$C(\Phi) = \frac{1}{N_E} \left| \sum_{i=1}^{N_E} e^{j\phi_i} \right|, \tag{19}$$

and the *pairwise consistency* as

$$\tilde{C}(\Phi) = \max_{i,j} \frac{1}{2} \left| e^{j\phi_i} + e^{j\phi_j} \right|, \ i \neq j, \ i,j \in \{1, ..., N_E\}. \tag{20}$$

It is easy to observe that $C(\Phi), \tilde{C}(\Phi) \in [0, 1]$ and a value close to 1 is obtained when the estimates converge to the same direction. When $N_E = 2$, the pairwise and the overall consistency metrics are identical.

2) Candidate DOA Fusion: At each TF point, we calculate the auxiliary variable z_j, $j = 1, ..., J$ where sub-index j is arbitrarily associated to each one of the angle estimation systems under consideration. The auxiliary variables are ordered in terms of their absolute value and the one with the largest absolute value (and thus least reliable for DOA estimation according to the analysis in Section II-D) is disregarded. If among the remaining auxiliary variables there is any z_j for which $|z_j(\tau, \omega)| \geq 1$, then this is also disregarded, as explained in Section II-C. Let $\Phi_z(\tau, \omega)$ denote the set with the remaining auxiliary variables. The procedure to decide whether a local DOA $\alpha(\tau, \omega)$ will be assigned to the particular TF point or not is as follows (time-frequency index is omitted for convenience);

1) If the cardinality of Φ_z is greater or equal to 2, then proceed to the next step, otherwise, completely disregard the particular TF point, assigning an empty local DOA $\alpha = \emptyset$
2) For each available auxiliary variable calculate the corresponding candidate DOA $\hat{\theta}_j = E_j(z_j)$ and store these DOAs in the set Φ_θ
3) Calculate the pairwise consistency metric using Eq. (20) and let k, l be the index of the two direction estimates

exhibiting the maximum pairwise consistency. Assign a direction α to the particular TF point (or not) as

$$\alpha = \begin{cases} \angle(e^{j\theta_k} + e^{j\theta_l}), & \text{if } \tilde{C}(\Phi_\theta) \geq 1 - \epsilon_\alpha \\ \emptyset, & \text{otherwise} \end{cases} \quad (21)$$

where $\epsilon_\alpha << 1$ is a predefined positive threshold. This type of processing reflects the belief that TF points which are contaminated with noise and/or reverberation will lead to irrelevant candidate direction estimates, inevitably failing the proposed consistency test. On the contrary, there will be at least one case of strong pairwise agreement at TF points characterized by a dominant directional part, and the average of the two candidate DOAs in that pair is regarded as the DOA specific to the particular TF point of analysis.

3) Local DOA Fusion: We have observed that, especially for speech sources, the estimation accuracy can be significantly improved by applying the previously defined selection process based on angular coherence not only across different candidate estimates obtained at the same TF point, but also across different neighbour TF points. This condition is in accordance with the belief that the dominance of a source in the TF domain appears not on unique/isolated frequency points but on a neighborhood of consecutive frequency bins which we call "zones". Assume that the nth zone comprises G consecutive frequency bins, then by applying the previous step in this zone we will end up with a collection of local DOAs $\Theta_n = [\alpha_1, ..., \alpha_{G'}]$, with $G' \leq G$. The condition for assigning a DOA to the particular zone is that we have $G' \geq N_Z$ entries in the nth set and that the overall consistency metric of that set is greater than a threshold. More specifically, this two step process can be written as follows;

1) If the cardinality of set Θ_n is equal or greater to N_Z go to the next step, otherwise set $\beta(\tau, n) = \emptyset$ and proceed to the next zone
2) Assign a DOA to the nth zone based on the output of the overall consistency metric as

$$\beta(\tau, n) = \begin{cases} \angle(\frac{1}{G'} \sum_{i=1}^{G'} e^{j\alpha_i}), & \text{if } C(\Theta_n) \geq 1 - \epsilon_\beta \\ \emptyset, & \text{otherwise} \end{cases} \quad (22)$$

and proceed to the next zone.

Similar to Eq. (21), ϵ_β stands for a predefined threshold to decide upon the reliability of the specific zone. The zone specific DOA $\beta(\tau, n)$ provides the highest-level of local DOA information which is required for building the final histogram, upon which multiple source localization and counting relies. The entire procedure for DOA estimation is summarized in Algorithm 1, with K_n denoting the set with the frequency indices comprising the nth frequency zone.

To illustrate the efficiency of the proposed approach, we have plotted the histogram with the resulting DOA collections, before and after local DOA fusion, in Fig. 5(a) and (b) respectively, for the case of three simultaneous speakers at -115, 60 and 90 degrees inside a reverberant environment. The histogram after candidate DOA fusion, is shown in Fig. 5(a). Applying local DOA fusion reduces the number of available estimates but at the same time makes the peaks in the histogram more easily identifiable, as it can be seen for the two sources at 60 and 90 degrees in Fig. 5(b).

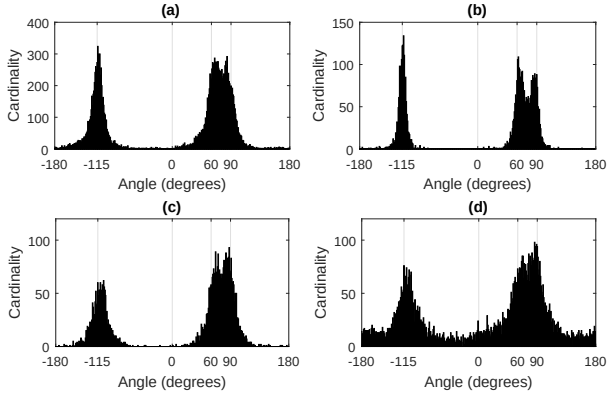

Fig. 5. Histogram with the DOAs obtained from PCSF after candidate DOA fusion in (a) and after local DOA fusion in (b). Histogram obtained for the same conditions using DIRAC is shown in (c) and using CICS in (d). Results are shown for a simulated reverberant environment of $RT_{60} = 300$ ms with three simultaneous sources at -115, 60 and 90 degrees.

Algorithm 1 PCSF

Input: microphone signals: $\mathbf{p}(\tau, \omega)$
Output: zone specific DOA collection: $\beta(\tau, n)$
for $\tau = 1$ to $Nframes$ **do**
  for $n = 1$ to $Nzones$ **do**
    $\Theta_n \leftarrow \emptyset$
    for $\omega \in K_n$ **do**
      $\Phi_z \leftarrow \emptyset$
      $\hat{\mathbf{\Phi}}(\tau, \omega) \leftarrow$update_cov_matrix$(\hat{\mathbf{\Phi}}(\tau - 1, \omega), \mathbf{p}(\tau, \omega))$
      for $j = 1$ to J **do**
        $z_j \leftarrow$calculate_auxiliary_var$(\hat{\mathbf{\Phi}}(\tau, \omega), d_j)$
        $\Phi_z \leftarrow \Phi_z \cup z_j$
      end for
      $\Phi_\theta \leftarrow$calculate_candidate_DOAs$(\Phi_z, \hat{\mathbf{\Phi}}(\tau, \omega))$
      $\alpha(\tau, \omega) \leftarrow$candidate_DOA_fusion$(\Phi_\theta, \epsilon_\alpha)$
      $\Theta_n \leftarrow \Theta_n \cup \alpha(\tau, \omega)$
    end for
    $\beta(\tau, n) \leftarrow$local_DOA_fusion$(\Theta_n, \epsilon_\beta, N_Z)$
  end for
end for

III. ADDITIONAL METHODS FOR DOA ESTIMATION

In order to compare our technique with other algorithms, we describe four additional well-studied methods for DOA estimation. Implemented in the TF domain, the selected algorithms, together with PCSF, are subjected to the same histogram-based framework for sound source localization and counting in Section V.

A. DIRAC

Directional Audio Coding (DIRAC) is a very well known technique for capturing and reproducing spatial audio events [34]. DIRAC has been originally designed for B-format microphone signals, which are suitable for 3D sound field analysis. However, with some minor modifications, the method can be easily adapted to planar microphone arrays, such as the square microphone array previously described [29], [35]. In

this paper, the method described in [35] is implemented in the TF domain providing two kinds of parameters which are useful for the scope of our investigation; a potential DOA $\hat{\theta}(\tau,\omega)$ and a two-dimensional approximation of the diffuseness $\Psi(\tau,\omega)$. It is reasonable to expect that portions of the signal characterized by a small diffuseness value are less contaminated with noise and/or reverberation than others. We thus propose to use the diffuseness value as a metric for assessing the reliability of a certain TF point.

Let $\Psi(\tau,\omega)$ denote the diffuseness at a particular TF point and $\hat{\theta}(\tau,\omega)$ be the direction associated to that TF point, both of which are provided by DIRAC. The process for assigning a local DOA $\beta(\tau,\omega)$ to that TF point reads

$$\beta(\tau,\omega) = \begin{cases} \hat{\theta}(\tau,\omega), & \Psi(\tau,\omega) \leq T_{\Psi} \\ \emptyset, & \text{otherwise} \end{cases} \tag{23}$$

where T_{Ψ} is set equal to 0.2 in this work. The local DOAs are accumulated across multiple frequency points and time-frames and processed in the form of the histogram using matching pursuit as described in a following section.

B. CICS With SSZ Detection

A method for multiple sources DOA estimation was presented by the authors in [13]. It relies on the sparsity of audio signals in the TF domain, by detecting areas in the TF domain where only one source is active. Those areas are called single source zones (SSZs) and their detection is accomplished via the estimation of the average correlation coefficient between all pairs of microphones comprising a uniform circular array.

A frequency domain single source DOA estimation method is applied over the strongest TF bins of each detected SSZ, which is based on the estimation of the circular integrated cross spectrum (CICS) [28]. The CICS function is estimated for every possible direction ϕ in the xy-plane and the index of its highest value reveals the DOA of the active source at the SSZ under consideration,

$$\beta(\tau,\omega) = \arg \max_{0\leq\phi<2\pi} |\text{CICS}^{(\omega)}(\phi)|, \tag{24}$$

where ω belongs in a SSZ. In this paper, we have used non-overlapping zones with band corners in the set $\{$100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2340, 2680, 3020, 3360, 3700, 4040, 4380$\}$ in Hz. Furthermore, the zone-specific correlation coefficient is averaged between diagonal sensor pairs only, e.g., pairs $\{1,3\}$ and $\{2,4\}$ with respect to the numbering of Fig. 2(a). When the correlation coefficient exceeds the value of 0.8, a frequency zone is considered to be a SSZ and a DOA specific to that zone is calculated, otherwise the empty set is returned.

C. Circular Harmonics Beamforming

The next DOA estimation method that we intend to examine was presented in [4] and relies on beamforming in the circular harmonics domain at each TF bin, adopting the WDO assumption [36] for the sparsity of the source signals. The estimated DOA at each TF point is obtained as the angle where the circular harmonics beamformer output power gets maximized, i.e.,

$$\beta(\tau,\omega) = \arg \max_{0\leq\phi<2\pi} |Y(\omega,\phi)|, \tag{25}$$

where $Y(\omega,\phi)$ is the output of the beamformer for a steering direction ϕ. Similarly to the proposed and the aforementioned methods, DOA estimates resulting from the Circular Harmonic Beamforming (CHB) method at each TF point are collectively processed in order to form a histogram, as explained in more detail in the coming section.

D. MUSIC

MUSIC is the most well known subspace method for DOA estimation and we employ this technique here by using the rank-1 approximation presented in [15]. At each TF bin, we calculate the pseudospectrum and we keep only the DOA corresponding to the maximum value in order to build a histogram, in the same spirit as in [37]. Similar to PCSF, we have observed that DOA estimation performance improves by operating on a temporally and spectrally averaged version of the local signal covariance matrix. More details about how the local covariance matrix is estimated at each TF point are given in Section V.

IV. SOURCE LOCALIZATION AND COUNTING ALGORITHM

We describe a common framework for joint localization and counting of the sound sources based on the approach proposed in [13]. The approach is based on the formation of a histogram from all the DOA estimations in a block of L consecutive time-frames. Let $\beta(\tau,\kappa)$ denote the DOA specific to a particular time-frame τ and subband region or frequency point, in general indexed here by κ. We note here that if a selection method is used, $\beta(\tau,\kappa)$ may be the empty set, meaning that the number of available DOAs at time τ may vary significantly from one technique to the other. To formulate the block-based estimation procedure, let the time-frame specific collection $B(\tau)$ be constructed as

$$B(\tau) = \cup_{\kappa}\beta(\tau,\kappa). \tag{26}$$

We extend the collection, not only across many subband regions or frequency bins, as Eq. (26) implies, but also across multiple time-frames to construct the global DOA estimation collection as

$$\Gamma(\tau) = \cup_{t=\tau-L+1}^{\tau} B(t), \tag{27}$$

where L is an integer denoting the History Length (HL). Note that $\Gamma(\tau)$ is updated at each time-frame, making it possible to handle situations in which the sound sources' number and locations vary dynamically in time [13]. Accumulating the DOAs at the current and the past time-frames may significantly improve the accuracy of localization, reducing however the responsiveness of the method. Furthermore, this action requires a period of L time-frames after the onset of the sound source in order to reach to its full effect. Periods of time of 1 up to $L-1$ time-frames after the onset of the sound source will be referred as "transition periods" from now on.

As shown in [13] and depicted in Fig. 5, the local DOAs are expected to cluster smoothly around the true source directions, while erroneous estimates due to noise and/or reverberation will appear with a low cardinality in the histogram and therefore will not severely affect the final decision. The natural approach for jointly estimating the number Q and location of the sound sources is to apply Matching Pursuit (MP) [38], using source atoms modelled as smooth pulses centred at different points on a grid of A possible angles. However, as stated in [13], it is advantageous to consider a "narrow" source atom for detecting the peak in the histogram and a "wide" atom for removing the contribution of that source. The use of two fixed atom widths in MP stands as an interesting alternative to more sophisticated fitting approaches, such as those in [2], [3], which rely on Expectation Maximization for estimating not only the mean, but also the variance of the curve at each local peak in the DOA histogram.

Let $\mathbf{B}^n = [\mathbf{b}_1^n, ..., \mathbf{b}_A^n]$ and $\mathbf{B}^w = [\mathbf{b}_1^w, ..., \mathbf{b}_A^w]$ denote the $A \times A$ dictionaries with the narrow and wide source atoms respectively and $\mathbf{h}_\tau$ be the $A \times 1$ vector with the histogram cardinality values ordered from 0 to 360 degrees, at time-frame τ. This histogram is a smoothed version of the original histogram which is constructed from the global collection $\Gamma(\tau)$. Smoothing is simply achieved by convolving the original histogram with a rectangular window of N_w degrees length. Our source counting and DOA estimation algorithm in slightly simpler in comparison to that proposed in [13] and proceeds as follows:

1) Calculate the Euclidean norm of the histogram $E_0 = \|\mathbf{h}_\tau\|_2$, set ϵ_0 to a very small value and set the loop index $q = 1$
2) Form the product $\mathbf{c} = \mathbf{B}^{n\,T}\mathbf{h}_{\tau,q}$
3) Let the elements of $\mathbf{c}$ be given by c_i, find $i^* = \arg\max_i c_i$
4) The DOA of this source is given by $(i^* - 1) \times 360/A$ degrees
5) Calculate the contribution of this source as

$$\epsilon_q = \frac{\mathbf{b}_{i^*}^{n\,T}\mathbf{h}_{\tau,q}}{\mathbf{b}_{i^*}^{n\,T}\mathbf{b}_{i^*}^n}$$

6) If $\frac{\epsilon_q}{\epsilon_{q-1}} < T_\epsilon$ or $\epsilon_q/E_0 < \gamma$ go to step 10
7) Remove the contribution of this source as

$$\mathbf{h}_{\tau,q+1} = \mathbf{h}_{\tau,q} - \mathbf{b}_{i^*}^w c_{i^*}$$

8) Increment q
9) If $q \leq Q_{MAX}$ go to step 2
10) $\hat{Q}_\tau = q - 1$ and the corresponding DOAs are those estimated in step 4

The main difference with the previously proposed algorithm in [13] is that we account for one only threshold γ (rather than Q_{MAX} such thresholds) and that we stop when there is a great difference in the estimated contributions from one iteration to the next one, based on the ratio $\frac{\epsilon_q}{\epsilon_{q-1}}$ and the corresponding threshold T_ϵ. Finally, an even simpler variant of this method rises if $\mathbf{b}_i^n = \mathbf{b}_i^w, \forall i$, corresponding to the case that a single atom width is used for implementing MP.

V. EVALUATION AND DISCUSSION

Results are presented in terms of real and simulated data using a uniform circular array of 4 omnidirectional sensors and radius $R = 0.02$ m. The values of several parameters were kept the same for both experiments based on simulated and real data and were as follows; for the STFT we use a squared Hanning window of 1024 samples length and a hop size of 512 samples (50% overlap) at a sampling rate of 22050 Hz, while the maximum frequency of interest was set to 4500 Hz for all techniques. Similar to [13], we used a Blackman sequence for constructing the source atoms required for MP; their centres span -180^o to 179^o with a resolution of 1^o. The history length is set to $L = 43$ time-frames, corresponding to a duration of 1 s and finally, the length of the rectangular window used for smoothing the histograms was set equal to $N_w = 3^o$.

Cross-spectra terms required for PCSF and MUSIC are exported from the observed covariance matrix $\hat{\mathbf{\Phi}}(\tau, \omega)$, which is estimated by exploiting the observed data in several TF bins instead of just one. In general, a spectrally smoothed estimation of the cross-covariance matrix is obtained as

$$\tilde{\mathbf{\Phi}}(\tau, \omega) = \sum_{i=-\lambda}^{\lambda} w_i \mathbf{p}(\tau, \omega + i)\mathbf{p}^H(\tau, \omega + i), \qquad (28)$$

where $\mathbf{p}(\tau, \omega) = [p_1(\tau, \omega), ..., p_4(\tau, \omega)]^T$ represents the signal at the microphones and $\lambda = 1$ with $w_{-1} = w_1 = 0.5$ and $w_0 = 1$ for PCSF, while $\lambda = 2$ with $w_{-2} = w_2 = 0.25$, $w_{-1} = w_1 = 0.5$ and $w_0 = 1$ for MUSIC. Then, this matrix is smoothed in time using the simple recursive formula

$$\hat{\mathbf{\Phi}}(\tau, \omega) = (1 - a)\hat{\mathbf{\Phi}}(\tau - 1, \omega) + a\tilde{\mathbf{\Phi}}(\tau, \omega), \qquad (29)$$

with $a = 0.7$. For candidate DOA fusion with the proposed technique we have used $\epsilon_\alpha = 0.0038$ while local DOA fusion is implemented based on a segmentation of the spectrum from 150 to 4500 Hz into 57 bands of 75 Hz width and 50% overlap. The process of Section II-G3 is applied with $N_Z = 2$ and $\epsilon_\beta = 0.02$.

It should be noted that different estimators in PCSF involve different frequency limits with respect to spatial aliasing. In particular, observe that $l_{\mathbf{x}} = l_{\mathbf{y}} = \sqrt{2}R$ while $l_{\mathbf{u}} = l_{\mathbf{v}} = 2R$, corresponding to maximum frequencies of approximately $f_A = 4.3$ kHz and $f_A = 3.0$ kHz respectively, according to the analysis made in Section II-C. We have observed that although the alias-free limit for estimators $E_{\mathbf{u},\mathbf{v}}$ and $E_{-\mathbf{u},\mathbf{v}}$ is well below the maximum frequency of analysis used in the evaluation (4.5 kHz), this does not seem to cause serious problems. We believe that this is a consequence of the overall fusion process, which causes erroneous candidate DOAs to be eliminated during the candidate and local DOA fusion steps.

A. Simulated Environment

Simulation results are presented for the case of a rectangular room with dimensions of $L_x \times L_y \times L_z = 5 \times 6 \times 3$ m and with the center of the circular array placed at [2.6 3.2 1.4] m. For these simulations, the image source method of Allen and Berkley [39] was implemented in Matlab using the toolbox

provided in [40]. For the source signals we used recordings of continuous speech from different subjects of 6 s duration each. In each simulation the sound sources have approximately equal power and the signal-to-noise ratio (SNR) is estimated as the ratio of the power of the first speech signal to the power of the noise signal at the first microphone.

Results reported in terms of DOA estimation performance in the simulated environment are derived with the following settings; in all cases, the sources are placed at a distance of 1.4 m from the center of the array and the number of sources Q is assumed known, although Q can be estimated jointly with the sound source locations using MP . This is done in order to focus the comparison on source localization performance solely (source counting results are presented at a later point). For the case of $Q = 1$ speakers, we span all azimuth angles in $[-180, 180)$ with 50 equidistant steps. For $Q > 1$, 50 different speaker placements are considered for each Q, where speakers are randomly distributed along the azimuth plane with the restriction that no speaker is closer than 20 degrees to another speaker. As a metric of the DOA estimation performance, we use the Mean Absolute Estimation Error (MAEE) which measures the absolute difference between the true DOA and the estimated DOA, in degrees, averaged over all sources, orientations and time-frames of the source signals [13]. Results are reported in terms of MAEE by excluding the transition period, i.e., time-frames with index lower than $L = 43$.

The first results presented for the scope of this evaluation intend to highlight the dependency of localization performance on the atom width, but also the large differences characterizing the performance of the DOA estimation techniques under different acoustic conditions. Specifically, in Fig. 6, we demonstrate the MAEE as a function of a single atom-width ($\mathbf{b}_i^n = \mathbf{b}_i^w, \forall i$), for four different acoustic conditions; highly reverberant environment with almost no noise in Fig. 6(a), and lightly reverberant environment with substantial amount of different types of noise; spherically isotropic noise [41] in Fig. 6(b), common mode noise in Fig. 6(c) and spatially white noise in Fig. 6(d). In principle, spherical isotropic noise is a more realistic model of acoustic noise in the case of a coincident microphone array operating inside a reverberant environment [30]. On the other hand, common mode noise type considers the case of additive white noise which is identical to all microphone channels. Apart from radio frequency interference which can be the cause of this type of noise for audio circuits [42], the particular noise model can be linked to the case of a microphone array with a sound reproduction unit (loudspeaker) embedded in it (e.g., Amazon Echo). As the loudspeaker and the microphones are embedded inside the same device, the acoustic paths from the loudspeaker to the microphones are almost identical, which forces the interference component to vary trivially from one microphone to the other.

In all cases, results in Fig. 6 are shown for a scenario with 3 static speech sound sources. The simulations were performed for spherical isotropic white noise of 30 dB SNR and $\text{RT}_{60} = 400$ ms in (a), spherical isotropic white noise of 5 dB SNR and $\text{RT}_{60} = 200$ ms in (b), common mode noise of 5 dB SNR and $\text{RT}_{60} = 200$ ms in (c) and spatially white noise of 5 dB SNR and $\text{RT}_{60} = 200$ ms in (d).

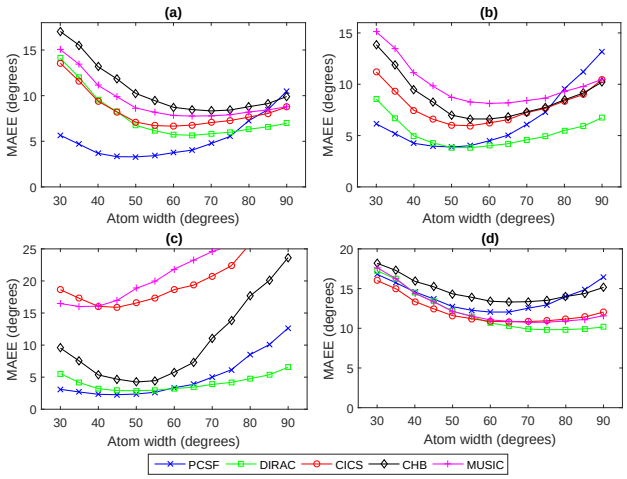

Fig. 6. MAEE for each technique as a function of the length of the Blackman sequence used for constructing the atoms for MP. Results are shown for three speakers in a simulated environment of $\text{RT}_{60} = 400$ ms and SNR=30 dB of isotropic noise in (a), $\text{RT}_{60} = 200$ ms and SNR=5 dB of isotropic noise in (b), $\text{RT}_{60} = 200$ ms and SNR=5 dB of common mode noise in (c) and $\text{RT}_{60} = 200$ ms and SNR=5 dB of spatially white noise in (d).

Figure 6(a) reveals that for an atom width of 55 degrees, PCSF may significantly improve sound source localization performance in conditions of high reverberation and relatively low noise, exhibiting twice as good score, in terms of MAEE, in comparison to DIRAC and CICS. While not very different in terms of performance, PCSF and DIRAC exhibit very good tolerance to isotropic noise and common mode noise in Figs. 6(b) and (c). The robustness of DIRAC to common mode noise can be understood from the fact that the method relies on the sound pressure differences across microphone pairs, so that a common component among the microphones is cancelled out. Similarly, the common noise components are expected to vanish in the same way that the isotropic noise component vanishes for PCSD in Eq. (5). On the contrary, it can be observed that common mode noise has a disastrous effect on CICS as well as on MUSIC. Finally, from Fig. 6(d) it can be observed that at 5 dB SNR of spatially white noise DIRAC, CICS and MUSIC may potentially perform better than PCSF.

An additional conclusion derived from Fig. 6(a) is that CHB and MUSIC provide poor performance in reverberant conditions. Indeed, we have observed that with the given array, both CHB and MUSIC give rise to very noisy histograms. To a large degree, we believe this is a consequence of the strict WDO assumption and the fact that the methods do not make any TF point selection to extract directional information. An additional reason for CHB is that, with the given sensor array configuration, the method can only implement beamforming based on first-order harmonics, which limits significantly the method's spatial resolution.

Further conclusions which may be drawn by Fig. 6 is the fact that the atom width indeed severely affects the localization error but the values that guarantee the best performance for each technique are more or less consistent for different amounts and types of noise. Finally, the graphs illustrate that the atoms widths which maximize performance for PCSF are

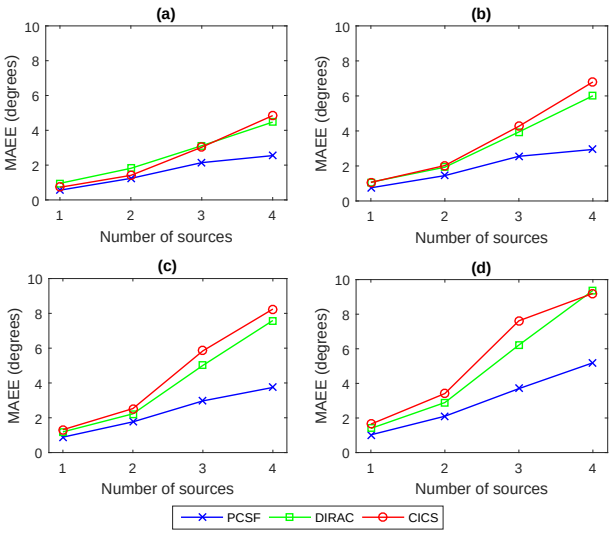

Fig. 7. MAEE for each technique at 30 dB SNR of spherical isotropic noise as a function of the number of speakers for $RT_{60} = 200$ ms in (a), $RT_{60} = 300$ ms in (b), $RT_{60} = 400$ ms in (c) and $RT_{60} = 500$ ms in (d)

smaller than those for DIRAC and CICS. For the results that follow, two atom widths are used for implementing MP, as suggested in Section IV. The narrow and wide atom widths are set to 40 and 60 degrees, respectively, for PCSF and to 45 and 65 degrees, for DIRAC and CICS. These values were empirically chosen, by taking into account both the localization and counting performance, and were found to be representative of the best that each technique can offer across a wide range of RT_{60} values and number of sources.

An additional series of simulations was performed with the goal to provide a more detailed comparison for the different techniques in terms of localization performance across different reverberation times and number of continuously active static speakers. In order to achieve a better clarity of presentation, CHB and MUSIC, which show the least competitive performance among all techniques, are ignored in this evaluation and results are shown for three techniques only, namely PCSF, DIRAC and CICS. The MAEEs are calculated at 30 dB SNR of spherical isotropic noise and plotted for reverberation times ranging from $RT_{60} = 200$ to $RT_{60} = 500$ ms with a step of 100ms in Fig. 7. The graphs illustrate that DIRAC may achieve slightly better performance than CICS at high reverberation times. On the other hand, PCSF achieves evidently better localization performance in comparison to both DIRAC and CICS, a fact which becomes more prominent as reverberation and number of acoustic sources increases. A general cause for this improvement is that PCSF produces less noisy histograms, making it easier to detect the sound source locations. In Fig. 5(b), (c) and (d) we illustrate an example of this qualitative difference in the histograms obtained with PCSF, DIRAC and CICS respectively, for a representative scenario of 3 simultaneous speakers at $RT_{60} = 300$ ms (DOAs are accumulated across the entire duration of the sound signals). As it will be shown later, this advantage is also reflected in the source counting performance.

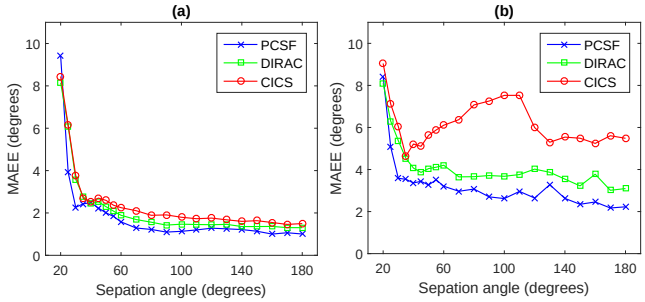

Fig. 8. MAEE for two speakers inside a simulated environment as a function of source separation angle. Results are shown at 10 dB SNR for $RT_{60} = 200$ ms in (a), $RT_{60} = 400$ ms in (b).

TABLE II
SOURCE COUNTING SUCCESS RATES FOR PCSF, DIRAC AND CICS AT RT_{60}=300 AND 500 MS.

Number of Sources (Q)	RT_{60} 300 ms PCSF $\gamma = 0.018$	DIRAC $\gamma = 0.056$	CICS $\gamma = 0.076$	500 ms PCSF $\gamma = 0.032$	DIRAC $\gamma = 0.065$	CICS $\gamma = 0.082$
1	100.0%	100%	99.0%	99.6%	98.0%	84.9%
2	97.4%	93.0%	88.7%	85.7%	83.5%	77.9%
3	90.6%	77.9%	78.3%	75.6%	68.0%	66.0%
4	80.0%	55.5%	46.8%	62.4%	45.9%	24.8%

In an additional set of simulations, we investigate the spatial resolution of each method, i.e., how close two sources can be in terms of angular distance while accurately estimating their DOA. Fig. 8 shows the MAEE against the angular distance for pairs of static, continuously active speakers at 10 dB SNR of isotropic noise for $RT_{60} = 200$ ms in (a) and $RT_{60} = 400$ ms in (b). While keeping the angular distance of the sources fixed, the resulting MAEE values are averaged across 19 different rotations, where each static source pair is rotated from its initial setting from 0 to 180 degrees with a step of 10 degrees. It can be seen that the proposed method achieves better spatial resolution than DIRAC and CICS and this advantage is more prominent in Fig. 8(b) where reverberation time is higher. Furthermore, the localization performance drops rapidly for separation angles less than 25 degrees, which can be seen as a lower limit for localizing two closely spaced sound sources.

Up to this point, we have presented DOA estimation results assuming that we know the exact number of sound sources that contribute to the sound scene at each moment. In practice, this information is not available and it should be inferred from the data, as explained earlier in Section IV.

In Table II we present source counting results for the simulated environments of RT_{60} =300 and 500 ms at 30 dB SNR of spherical isotropic noise. Results are presented in terms of success rates, corresponding to the percentage of time-frames correctly counting the number of sources, under the assumption that the minimum possible number of active sound sources is 1 and the maximum is 5. The acoustic conditions for these simulations were exactly the same as the ones used for DOA estimation in the previous subsection, and the success rates were averaged across all 50 source orientations and time-frames excluding the transition periods. In order to make a fair comparison, we tuned the value of γ for

each technique and each RT_{60} to the value that maximizes the average success rate score for 1 up to 4 number of sources, for a value of the parameter T_ϵ common to all techniques and equal to 0.15. As it can be seen in Table II, the source counting performance varies from one technique to the other in proportion to the localization performance, with PCSF achieving a significant advantage compared to DIRAC and CICS, especially for $Q = 3$ and 4 simultaneous sources.

B. Real Environment

Apart from simulations, we present results obtained from a real environment, a typical office room with approximately the same dimensions and placement of the microphone array as in the simulations and with an RT_{60} approximately equal to 300 ms. For the recordings we used four Shure SM93 microphones (omnidirectional) with a TASCAM US2000 8-channel USB soundcard. The microphone locations were stabilized using a red plastic circular case and the microphone array was placed on top of a wooden circular table which was close to the center of the room. In this experiment, we have derived and used a wideband estimation of the signal energy at each microphone in order to compensate for different microphone gains, possibly introduced by the imperfect tuning of the gain knobs on the microphone pre-amplifier. Although this was found to improve performance for all three methods, it is certainly far from saying that an accurate calibration of the acquisition system was performed.

Data was produced by recording three male and three female continuously active static speakers at a distance of approximately 1.3 m from the center of the microphone array. Each speaker was recorded at one or two different locations, covering 13 different locations uniformly distributed along the azimuth plane and particularly at $\pm 166, \pm 139, \pm 111, \pm 83, \pm 56, \pm 28$ and 0 degrees. As each speaker was recorded separately, we were able to superimpose different recordings in order to investigate cases with multiple simultaneous speakers. We note that while the speakers were at approximately the same distance, the signals resulting from the different speakers exhibited a ± 3 dB variation in their powers, as expected for such a real life experiment. Finally, the SNR at each recording was estimated at approximately 25 dB, mostly caused by ventilation noise. Results with respect to the DOA estimation performance are presented assuming a known number of sources for $Q = 1$ up to 4 speakers as follows; for $Q = 1$ the MAEE is averaged across all 13 different cases while for $Q > 1$ we consider all the Q out of 13 combinations that rise given the available locations, resulting to 78, 286 and 715 combinations for $Q = 2$, 3 and 4 sources respectively. In all cases, the parameters involved for DOA estimation for each technique were exactly the same as in the simulations.

The outcome of this experiment is shown in terms of MAEE in Fig. 9 where it can be seen that PCSF has by far better localization performance compared to DIRAC and CICS. Furthermore, the relevant advantage in the performance gained from PCSF, which in the simulations appears for higher number of sources, now appears even for the case of a single speaker.

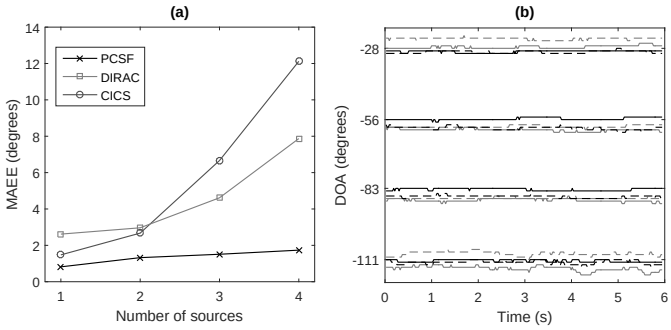

Fig. 9. Sound source localization in the real environment. MAEE as a function of the number of speakers in (a) and estimated DOA as a function of time for four different single active speaker locations for PCSF (blue), DIRAC (green), CICS (red) and CHB (black) in (b). The true DOAs are at -28, -56, -83 and -111 degrees.

As an attempt to explain the additional advantage that PCSF seems to achieve when going from simulated to real conditions, we have plotted in Fig. 9(b) the estimated DOA as a function of time for each technique for four indicative single speaker locations at -28, -56, -83 and -111 degrees. The graph indicates strong systematic errors between the estimated and the actual DOA, a fact that is not observed in the simulated environment. More particularly, DOAs estimated with PCSF deviate ± 2 degrees with respect to the ground truth, while for the other methods this deviation reaches ± 5 degrees at several time instants. We can think of two main reasons for explaining these phenomena; 1) insufficient calibration and/or minor differences in the magnitude and phase responses of the input channels and 2) acoustic diffraction effects, caused by the body of the plastic case where the microphones are mounted on. Interestingly, PCSF seems to be tolerant to these issues, producing DOA estimates which are much closer to the ground truth. To our opinion, this is additional evidence for the robustness of the proposed technique on certain types of model mismatch, which can be attributed to the way that different direction estimates are fused.

With respect to counting performance in the real environment, results are presented for PCSF with $\gamma = 0.018$ in terms of a confusion matrix in Table III. Following the DOA estimation performance, the counting results are also very satisfactory, achieving a success rate equal to 94.3% and 87.8% for three and four number of sources respectively. As an indication about the counting success rates of the other techniques, the corresponding values where 71.6% and 43.8% for DIRAC ($\gamma = 0.050$) and 64.6% and 43.7% for CICS ($\gamma = 0.068$).

In Fig. 10 we present an example of joint DOA estimation and counting, derived with PCSF, DIRAC and CICS inside the real environment, for a scenario where the number of speakers increases from one to four and then gradually decreases again to one. For this experiment, the four speakers were located at -85, -28, 0, and 140 degrees. The graph illustrates the estimated DOA as a function of time (coloured lines), while the actual locations and times of activation of each speaker are represented by the gray lines. As it can be seen, PCSF estimates correctly the DOAs and number of speakers at most

TABLE III
CONFUSION MATRIX OF COUNTING SUCCESS RATES FOR THE PROPOSED METHOD IN THE REAL ENVIRONMENT

		$\hat{Q}$				
		1	2	3	4	5
Q	1	100%	0.0%	0.0%	0.0%	0.0%
	2	4.3%	95.6%	0.1%	0.0%	0.0%
	3	0.3%	4.8%	94.3%	0.5%	0.0%
	4	0.1%	0.4%	10.7%	87.8%	1.0%

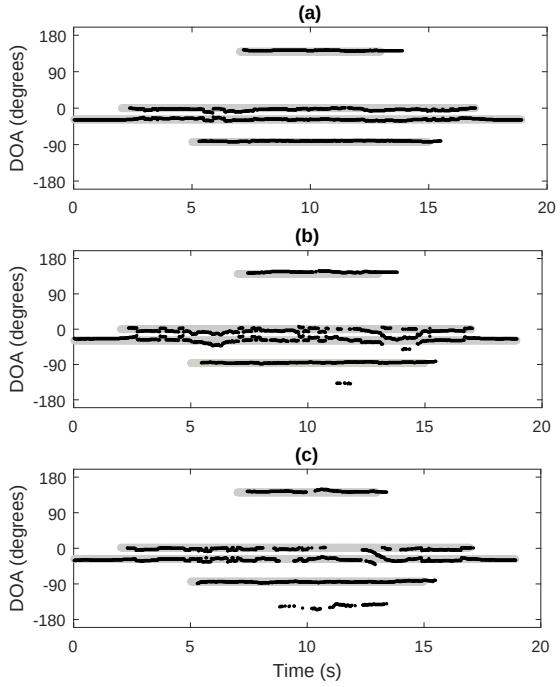

Fig. 10. Estimated DOA of four static speakers in a real environment using PCSF in (a), DIRAC in (b) and CICS in (c).

of the time-frames. On the other hand, the other two methods fail to detect the correct number of sources at several time instants or make erroneous DOA estimations. Particularly, at certain time-frames, both CICS and DIRAC detect a sound source at approximately -140^o, although there was no sound source there. This is caused by erroneous DOA estimations appearing in the histograms of the two methods, a phenomenon which is much less intense with PCSF. A final thing to note from Fig. 10 is that the estimation of each source is prolonged for some period of time after he/she stops talking or respectively is delayed when he/she starts talking. This is due to the block based decision used for processing all the estimated directions, which requires 1 second of history in order to construct the histogram upon which DOA estimation and counting relies. For the same reason, short discontinuities in the different speakers' activities are smoothed out.

C. Real-time Implementation

Real-time implementation is very important for many applications involving sound source localization. An important advantage of PCSF in comparison to other DOA estimation techniques, such as CICS, MUSIC and CHB that rely on a grid-search, is that the method establishes a closed-form solution between the DOA and the observed microphone signals. While this has a very positive impact on the accuracy and the complexity associated to direction estimation, the additional steps required for candidate and local DOA fusion further increase the complexity of the algorithm. With the square microphone array as an example, this involves the fact that, not one, but four different angle estimation systems operate in parallel, and the fact that all pairwise angular distances need to be calculated in order take a decision upon the candidate DOA set.

We performed several tests using the data acquired in the real environment and we observed that, implemented in Matlab with an Intel Core i7 @3.4 GHz CPU, joint DOA estimation and counting with the proposed method is characterized by a real time factor (RTF) of 80 %, where RTF is defined as the ratio of the computation time to the input duration. Potentially, by optimizing the numerics and by using a lower level programming language, such as C++, the computational resources required for a real-time implementation may be further reduced. Interestingly, the implemented versions of DIRAC and CICS are evidently lighter processes, operating at a RTF of approximately 40% and 60% respectively on the same computer. Nevertheless, the increased computational demand associated to PCSF is largely beshadowed by a significant improvement in DOA estimation and counting performance.

VI. CONCLUSION

PCSF operates on a smoothed – in time and in frequency – observation of the local signal covariance matrix to establish an analytic relation between two different cross-spectra terms and the incident acoustic direction. Inherent to this technique is the presence of multiple DOA estimation subsystems which operate in parallel, producing a multiplicity of candidate DOAs at each TF point. We have defined a metric of coherence, based on the property of divergence of the different DOA estimators, for assessing the reliability of different signal portions, so that only TF bins with a high quality of directional information are exploited for local DOA estimation. The final set of local DOAs is provided as input to a recently proposed histogram-based method which allows for joint estimation of the number of sources and their locations. Tests performed with a small sized square array revealed that the proposed technique may achieve a significant improvement in terms of source localization and counting performance, compared to other state of the art techniques, with the relevant advantage in performance increasing proportional to reverberation and to the number of sound sources. Furthermore, the technique shows good tolerance to spherical isotropic noise as well as to common mode noise.

In our future plans, we intend to evaluate PCSF for the case of an eight-sensor uniform circular array, using eight parallel DOA estimation subsystems which come with this particular array topology. Furthermore, it would be interesting to extend the method for 3D DOA estimation and counting, in conjunction with a properly designed microphone array.

ACKNOWLEDGMENT

This research has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 644283, Project LISTEN.

REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830–1847, 2004.

[2] M. Cobos, J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian Mixture Model," *Digital Signal Processing*, vol. 21, no. 1, pp. 66–76, 2011.

[3] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with Dirichlet prior," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009, pp. 33–36.

[4] A. Torres, M. Cobos, B. Pueo, and J. Lopez, "Robust acoustic source localization based on modal beamforming and time-frequency processing using circular microphone arrays," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1511–1520, 2012.

[5] N. Stefanakis and A. Mouchtaris, "Direction of arrival estimation in front of a reflective plane using a circular microphone array," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 622–626.

[6] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Process.*, vol. 91, pp. 1781–1788, 2011.

[7] S. Araki, H. Sawada, R. Mukay, and S. Makino, "DOA estimation for multiple sparse sources with arbitrary arranged multiple sensors," *J. Sign. Process. Syst.*, vol. 63, pp. 265–275, 2011.

[8] S. Mohan, M. Lockwood, M. Kramer, and D. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, p. 21362147, 2008.

[9] N. Tho, S. Zhao, and J. D., "Robust DOA estimation of multiple speech sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 2287–2291.

[10] A. Moore, C. Evers, P. Naylor, D. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct path dominance test," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2014, pp. 2296–2300.

[11] M. Puigt and Y. Deville, "A new time-frequency correlation-based source separation method for attenuated and time shifted mixtures," in *Proc. of the 8th International Workshop (ECMS and Doctoral School) on Electronics, Modelling, Measurement and Signals*, 2007, pp. 34–39.

[12] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 2625–2628.

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

[14] M. Aktas, T. Akgun, and H. Ozkan, "Acoustic direction finding in highly reverberant environment with single acoustic vector sensor," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2014, pp. 2346–2350.

[15] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, 2011.

[16] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[17] F. Belloni and V. Koivunen, "Unitary root-MUSIC technique for uniform circular array," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, (ISSPIT)*, 2003, pp. 451–454.

[18] S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 2009–2014.

[19] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2007, pp. 18–21.

[20] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, *Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays*. Springer Berlin Heidelberg, 2010, pp. 81–88.

[21] E. Fishler, M. Grosmann, and H. Messer, "Detection of signals by information theoretic criteria: general asymptotic performance analysis," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1027–1036, 2002.

[22] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Neural Information Processing Systems*. MIT Press, 2003, pp. 281–288.

[23] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*, ser. Academic Press. Elsevier, 2010.

[24] N. Ohwada and K. Suyama, "Multiple sound sources tracking method based on subspace tracking," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 217–220.

[25] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proceedings of the International Workshop for Acoustics Echo and Noise Control, (IWAENC)*, 2008.

[26] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proceedings of the 8th International Conference on Independent Component Analysis and Blind Source Separation*, 2009, pp. 742–750.

[27] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Source counting in real-time sound source localization using a circular microphone array," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Conference*, 2012, pp. 521–524.

[28] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2007, pp. 778–782.

[29] M. Kallinger, F. Kuech, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Analysis and adjustment of planar microphone arrays for application in directional audio coding," in *Proc. of 124th Convention of Audio Engineering Society*, 2008, paper 7374.

[30] N. Ito, N. Ono, E. Vincent, and S. Sagayama, "Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 2818–2821.

[31] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[32] L. Kuncheva, *Combining pattern classifiers*. John Wiley and Sons, 2004.

[33] T. Dietterich, "Ensemble methods in machine learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*. Springer-Verlag, 2000, pp. 1–15.

[34] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007.

[35] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Der Galdo, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Proc. of Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 37–40.

[36] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2002, pp. 529–532.

[37] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, and A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 1473–1477.

[38] M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Signal Process.*, vol. 47(7), pp. 1890–1902, 1999.

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

[40] D. Jarrett, E. Habets, M. Thomas, and P. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1462–1472, audiolabs-erlangen.de/fau/professor/habets/software/smirgenerator 2012.

[41] E. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[42] G. Ballou, *Handbook for Sound Engineers*. CRC Press, 2015.