A parametric approach to virtual miking for sources of arbitrary directivity

Mirco Pezzoli, *Student, IEEE,* Federico Borra, *Student, IEEE,* Fabio Antonacci, *Member, IEEE,* Stefano Tubaro, *Senior Member, IEEE* and Augusto Sarti, *Senior Member, IEEE,*

***Abstract*—In this manuscript we propose a methodology for the reconstruction of sound fields in arbitrary locations based on the signals acquired by a spatial distribution of compact microphone arrays (virtual miking). The proposed method is suitable for operating in reverberant environments, thanks to a two-stage analysis process, the former of which aims at separating the direct and the diffuse components of the sound field. The method that we propose is inherently parametric, as the sources of the acoustic scene are characterized by parameters describing location and directivity (spherical harmonics expansion), which are extracted from the exterior model of the direct component of the sound field. Once the parameters of the sources are extracted, the direct sound field at an arbitrary location is reconstructed. The diffuse component is reconstructed from the joint knowledge of the diffuse component at the locations of the distributed microphone arrays, under the assumption of isotropic behavior. Results show that the proposed technique is able to analyze the sound field and reconstruct the parameters of the sources that are active in the scene. In addition, the synthesis of the signals at the virtual microphone locations turns out to accurately match (in terms of spatial cues) the actual sound field, as measured by a microphone places in the desired location.**

***Index Terms*—distributed microphone arrays, virtual microphone, source localization, sound field reconstruction.**

I. Introduction

THE reconstruction of the sound field is a well known and challenging problem in the acoustic signal processing community. Also known as sound field recording or virtual miking problem, it consists of the reconstruction of the signal that would be acquired by a *virtual microphone* (VM) arbitrarily placed in the space, starting from the signals acquired by different configurations of microphones. With proper analysis and synthesis techniques, this enables a listener to navigate a recorded acoustic scene and perceive the spatial sound characteristics as they were at the recording location.

The solutions that address this problem in the literature are generally classified into non-parametric and parametric methods. Non-parametric methods [1]–[8] are based on the decomposition of the acquired sound field into spatial Fourier basis functions that are directly derived from the solution of the wave equation. Although promising, particularly for analysis purposes, these methods are usually rather demanding in terms of hardware and computational requirements. Parametric methods [9]–[18], on the other hand, rely on a parametric model to describe a sound field in a compact and general fashion. They usually involve two successive steps. The first step concerns the analysis of the sound scene and the estimation of the model parameters. The second step concerns the synthesis of the desired signal and its implementation depends on the application. For example, if the goal is binaural rendering the synthesis step takes into account the influence of the listener's head and body, through a directional transfer function [19]. For the estimation of a VM signal, instead, the synthesis takes into account other descriptors that are more suitable for describing a VM, such as their pick-up pattern [14], [18].

Early parametric models, such as [9] and [12], are based on a decomposition of the sound field into a direct and a diffuse sound component together with additional information such as the direction of arrival or the position of the acoustic sources. The direct signal component is attributed to multiple plane waves at each time-frequency, while the diffuse sound field component is typically attributed to spatially extended acoustic sources and room reverberation that occurs in enclosed environments. The models that account for both direct and diffuse components are known in the literature as *geometric-based* parametric models [15]. The main drawback of methods such as [9] and [12] is that reconstruction is restricted to the representation of the spatial sound only at the acquisition location. They are, therefore, unable to estimate the sound field in locations that differ from those where the measurement is performed. This limitation, however, was overcome by recently proposed approaches [14], [16]–[18] that exploit a parametric model of the sound field in order to extend the reconstruction at an arbitrary location. In particular, [16] uses the parametric representation of [9]. The parameters are estimated from the signals recorded in a single spatial position by a first-order Ambisonics microphone. However, in order to reconstruct the sound field at arbitrary positions in the space, information about the distances of the sound sources from the recording position are assumed as known a-priori and the physical sources are considered separated in angle with respect to the recording position. In [14], instead, multiple distributed microphone arrays are employed. The authors adopt a geometric-based parametric model where the parameters consist in the positions of an isotropic point-like source together with the direct and diffuse components at such position for each time-frequency bin.

Manuscript received November XX, 2019; revised xx xx, 20xx.

This enables to place a virtual microphone in any point in space, independently of where microphones were placed. The main drawbacks of [14] are that the model does not consider directional sources; and that the locations of the sources must be estimated for each time-frequency bin. In [17], [18] the authors adopt a more complete model of the acoustic sources, which also incorporate their directivity patterns in free-field conditions. This way the advantages of a parametric model of the sound field are preserved, while attaining a more accurate description of acoustic sources. More precisely, the parameters to be estimated consist of the position of the sources, their directivity patterns and the signals emitted. Starting from the signals recorded by a set of distributed microphone arrays, these parameters are estimated by sequential spatial filtering operations with ad hoc designed spatial filters.

In this manuscript we extend the approach presented in [17], [18] to the case of reverberant environments, while retaining the parametric sound field representation inspired by [14], but also refining the model in terms of the acoustic source characteristics. More specifically, the sound field is assumed to be the mix of a direct and a diffuse component. The direct component results from the exterior field emitted by the sources and it is represented using a spherical harmonics expansion. In the analysis phase, the expansion coefficients are obtained from the direct component of the microphone signals, estimated through spectral enhancement, as the solution of a sparse regularized optimization problem that avoids the need to know which source is dominant for each time-frequency bin. This requires the knowledge of the positions of the sources that, unlike in [14], can be estimated using wideband localization algorithms. In the synthesis phase the direct component at the VM is obtained using the estimated spherical harmonics coefficients. The diffuse component, on the other hand, is assumed to be isotropic and homogeneous. However, the estimate at different pairs of microphones can yield different results, due to residuals of the direct component in the estimation of the diffuse one. Similarly to [14], the diffuse component at the VM is estimated as a weighted sum of the estimations at all the microphone pairs. Furthermore, with the proposed approach the pick-up pattern and the sensitivity of the VM can be modeled arbitrarily.

The performance of the proposed method is assessed through an extensive simulation campaign. The employed metrics are aimed at estimating the accuracy of the reconstructed signals in terms of both spatial features, such as the directivity patterns of the sources and the overall reconstruction quality. Some relevant parameters of the sound field at the virtual microphone position are compared with the ground truth, along with the signal to distortion ratio resulting from the reconstruction. We finally investigate the case of a stereo recording application. From all the results, we can conclude that the proposed technique can accurately retrieve the main characteristics of the sound field.

The rest of the manuscript is structured as follows: in Sec. II we present the adopted sound field model; we offer a complete formulation of the problem; and we discuss the block diagram that describes the whole system. In Sec. III we offer a detailed description of each step of the model parameters estimation.

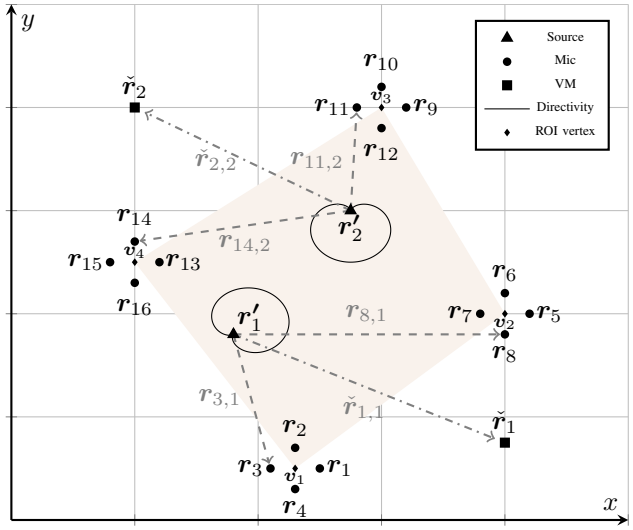


Fig. 1. 2D graphical representation of the model. The setup is presented with $A = 4$ circular microphone arrays of $M = 4$ microphones each. Two sources ($N = 2$) and two VMs ($V = 2$) are present in the scene. The directivity function of the source is superimposed on the plot of the scene.

We then present the synthesis of the VM signal in Sec. IV. We show the results of the simulation campaign in Sec. V. Finally, in Sec. VI we offer some conclusive remarks.

II. Data Model and Problem Formulation

In this section we introduce the data model and the virtual miking problem. We describe the data model in Sec. II-A starting from the definition of the adopted VM model, here we also introduce the parameters that need to be estimated. We then define the model of the signals acquired by the microphones recording the acoustic scene. Finally, in Sec. II-B we describe the virtual miking problem with the help of a block diagram underlying the needs and requirements of the proposed approach.

A. Data Model

Given a Cartesian coordinate system, let us consider N acoustic sources, placed in arbitrary locations $\boldsymbol{r}'_n = [x'_n, y'_n, z'_n]^T$, $n = 1, \ldots, N$; a network of $A \geq 2$ distributed compact microphone arrays with M microphones each, located at $\boldsymbol{r}_i = [x_i, y_i, z_i]^T$, $i = 1, \ldots, M \times A$; and a set of V VMs positioned in $\check{\boldsymbol{r}}_v = [\check{x}_v, \check{y}_v, \check{z}_v]^T$, $v = 1, \ldots, V$, as shown in Fig. 1.

We assume that sources, microphones and VMs lie on the same plane. Moreover, we define the Region Of Interest (ROI) where the sources lie as the polygonal region $\mathcal{R} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_A)$, where $\boldsymbol{v}_a$ is the ath vertex of the polygon. The vertices are defined as the centroids of each array, i.e.,

$$\boldsymbol{v}_a = \frac{1}{M} \sum_{i=(a-1)M+1}^{aM} \boldsymbol{r}_i \,, \qquad a = 1, \ldots, A. \tag{1}$$

When only two arrays are present in the acoustic scene, the definition of the ROI degenerates since we have a polygonal

region $\mathcal{R}$ with only two vertices. In this case, we consider the ROI as the whole plane where sources, microphones and VMs lie.

We model the signal of the vth directional VM in $\check{\boldsymbol{r}}_v$ in the time-frequency domain as the linear combination of a direct sound component and a diffuse sound component [14], i.e.,

$$\begin{aligned} S(t,\omega,\check{\boldsymbol{r}}_v) &= C_v(\omega)S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) \\ &+ Q_v(\omega)S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v),\ n\in\{1,\dots,N\}, \end{aligned} \tag{2}$$

where t is the time-frame index, $\omega = 2\pi f$ the radial frequency with $f > 0$ the temporal frequency, $C_v(\omega) \in \mathbb{R}$ models the VM microphone pick-up pattern and $Q_v(\omega) \in \mathbb{R}$ its sensitivity to the diffuse field. The model in (2) is valid under the assumption that the N source signals are sufficiently sparse in the time-frequency domain [10], [14]. More precisely, when multiple sources are simultaneously active, their signal content in the frequency domain must not overlap significantly, i.e. one source is dominant at each time-frequency bin. Let us denote with $\check{\boldsymbol{r}}_{v,n} = \check{\boldsymbol{r}}_v - \boldsymbol{r}'_n = [\check{x}_{v,n}, \check{y}_{v,n}, \check{z}_{v,n}]$ the vector pointing from the source position to the VM position (see Fig. 1) and with $\check{\rho}_{v,n}$, $\check{\theta}_{v,n}$ and $\check{\phi}_{v,n}$ the coordinates of $\check{\boldsymbol{r}}_{v,n}$ in a spherical coordinate system, i.e.,

$$\begin{aligned} \check{\rho}_{v,n} &= \sqrt{\check{x}^2_{v,n} + \check{y}^2_{v,n} + \check{z}^2_{v,n}}, \\ \check{\theta}_{v,n} &= \arccos\frac{\check{z}_{v,n}}{\check{\rho}_{v,n}}, \\ \check{\phi}_{v,n} &= \arctan\frac{\check{y}_{v,n}}{\check{x}_{v,n}}. \end{aligned} \tag{3}$$

The term $S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ represents the direct sound emitted by the nth source and received by the vth VM and it is modelled as the exterior field [1]

$$\begin{aligned} &S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v) \\ &= \sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell} \beta^n_{\ell\mu}(t,\omega) h_\ell\left(k\check{\rho}_{v,n}\right) Y_{\ell\mu}\left(\check{\theta}_{v,n},\check{\phi}_{v,n}\right), \end{aligned} \tag{4}$$

where $k = 2\pi f/c$, c is the speed of sound, $\beta^n_{\ell\mu}(\omega)$ are the exterior sound field coefficients of the nth source, $h_\ell(\cdot)$ is the ℓth order spherical Hankel function and $Y_{\ell\mu}\left(\check{\theta}_{v,n},\check{\phi}_{v,n}\right)$ is the spherical harmonic of order ℓ and degree μ, defined as

$$Y_{\ell\mu}\left(\check{\theta}_{v,n},\check{\phi}_{v,n}\right) = K_{\ell\mu}P_{\ell\mu}\left(\cos(\check{\theta}_{v,n})\right)e^{j\mu\check{\phi}_{v,n}}, \tag{5}$$

with

$$K_{\ell\mu} = (-1)^\mu\sqrt{\frac{(2\ell+1)}{4\pi}\frac{(\ell-\mu)!}{(\ell+\mu)!}} \tag{6}$$

and $P_{\ell\mu}(\cdot)$ the normalized associated Legendre polynomial. It is worth noting that despite sources, arrays, and VMs are assumed to be lying on the same plane, we consider them as placed in a 3D environment. Hence, in (4) we adopt a 3D propagation model. The term $S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ represents the diffuse sound field component and it is assumed as spatially isotropic and homogeneous, i.e., it arrives with equal strength from all the directions and its mean power does not vary with the position [14], [20]. It is worth noticing that in (2) we implicitly assume that the VM is noiseless.

The signal acquired by the ith omnidirectional microphone placed in $\boldsymbol{r}_i$ is modelled as

$$X\left(t,\omega,\boldsymbol{r}_i\right) = X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) + X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i) + N(t,\omega,\boldsymbol{r}_i). \tag{7}$$

The term $X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)$ is the direct sound emitted by the nth source and received by the ith microphone and, similarly to (4), is modelled as

$$\begin{aligned} &X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) \\ &= \sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell} \beta^n_{\ell\mu}(t,\omega) h_\ell\left(k\rho_{i,n}\right) Y_{\ell\mu}\left(\theta_{i,n},\phi_{i,n}\right), \end{aligned} \tag{8}$$

where $\rho_{i,n}$, $\theta_{i,n}$ and $\phi_{i,n}$ are the spherical coordinates of the vector $\boldsymbol{r}_{i,n} = \boldsymbol{r}_i - \boldsymbol{r}'_n$ (see (3)). The terms $X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)$ and $N(t,\omega,\boldsymbol{r}_i)$ are the spatially isotropic and homogeneous diffuse sound component and the ith sensor self-noise, respectively. The microphone self-noise $N(t,\omega,\boldsymbol{r}_i)$ is modelled as an uncorrelated zero-mean complex Gaussian noise with mean power

$$\Phi_{N,ii}(t,\omega) = E\{N(t,\omega,\boldsymbol{r}_i)N^\star(t,\omega,\boldsymbol{r}_i)\}, \tag{9}$$

where $E\{\cdot\}$ denotes the mathematical expectation and $(\cdot)^\star$ refers to the conjugate of a complex number.

B. Problem Formulation

The virtual miking problem consists of estimating the signal of a generally directional and arbitrarily placed virtual microphone, starting from the signals acquired by a set of distributed microphone arrays recording the acoustic scene. In this manuscript we approach this problem in a parametric fashion. In particular, in Sec. II-A, we developed in (2) and (7) a parametric model for both the VMs and the microphones recording the scene, respectively. The proposed solution can be seen as a system characterized by a set of unknown parameters that need to be estimated. The inputs to the estimation problem are the signals $X(t,\omega,\boldsymbol{r}_i)$, $i = 1,\dots,I$ of the microphones, their positions $\boldsymbol{r}_i$, the characteristics of each VM, namely the position $\check{\boldsymbol{r}}_v$, the pick-up pattern $C_v(\omega)$ and the sensitivity to diffuse noise $Q_v(\omega)$ and the number of sources N. In particular, as regards the latter parameter, it can be estimated using other sensors in the room (e.g., video-camera) or directly from the signals at the microphones as proposed, for example, in [21]–[28]. The output of the algorithm is an estimate $\hat{S}(\check{\boldsymbol{r}}_v)$ of the VM signal $S(\check{\boldsymbol{r}}_v)$.

In Fig. 2 a graphical representation of the proposed solution is depicted. In the block diagram we can identify the two main phases of the procedure namely the *parameters estimation* and the *synthesis* phase. In the former all the parameters needed for the synthesis of the VM signal are estimated. In particular, as it is clear from (2), in order to synthesize the signal at each VM we need to estimate both the direct $S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ and the diffuse $S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ components.

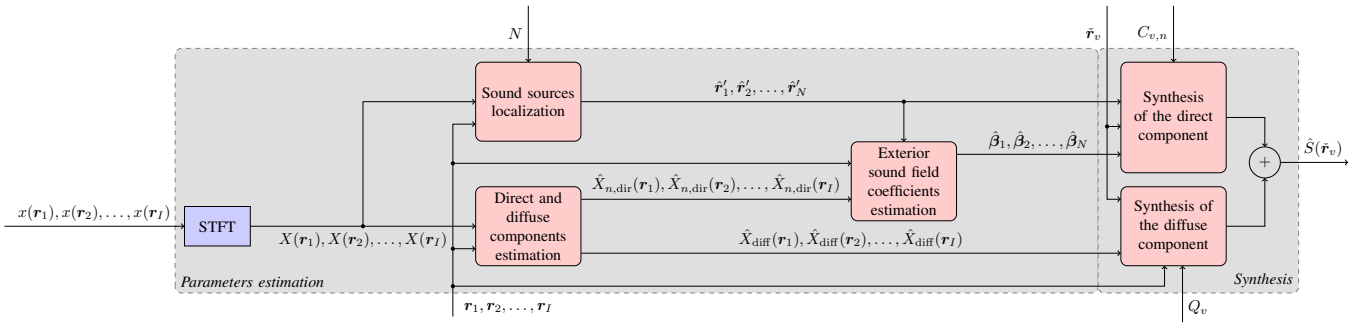


Fig. 2. The virtual miking technique block diagram. The microphone signals $x(\boldsymbol{r}_i)\, i = 1, \ldots, I$ are first transformed using the STFT into $X(\boldsymbol{r}_i)\, i = 1, \ldots, I$ which are used as input for the *Sound sources localization* (Sec. III-A) and *Direct and diffuse components estimation* (Sec. III-B) blocks along with the location of the microphones $\boldsymbol{r}_i$, $i = 1, \ldots, I$. The number of sources N is provided as input to the *Sound sources localization* block. The estimated location of the sources $\hat{\boldsymbol{r}}'_n$, $n = 1, \ldots, N$ and the direct component estimates $\hat{X}_{n,\mathrm{dir}}$ are used as input for the *Exterior sound field coefficients estimation* block (Sec. III-C). As regards the *Synthesis* phase (Sec. IV), the position of the vth VM $\check{\boldsymbol{r}}_v$ is shared by both the *Synthesis of the direct component* (Sec. IV-A) and *Synthesis of the diffuse component* (Sec. IV-B) blocks. In addition, the *Synthesis of the direct component* block requires the vth VM pick-up pattern $C_{v,n}$, $n = 1, \ldots N$, the estimated location of the sources $\hat{\boldsymbol{r}}'_n$, $n = 1, \ldots, N$ and the estimates of the exterior sound field coefficients $\hat{\boldsymbol{\beta}}_n$, $n = 1, \ldots, N$, while the sensitivity of the vth VM Q_v and the estimated diffuse components $\hat{X}_{\mathrm{diff}}$ are given as input to the *Synthesis of the diffuse component* block. Finally, the signal of the vth VM $\hat{S}(\check{\boldsymbol{r}}_v)$ is obtained as the sum of the synthesized direct and diffuse components.

1) VM direct component estimation: The model of the direct component $S_{n,\mathrm{dir}}(t, \omega, \check{\boldsymbol{r}}_v)$ is described in (4). The parameters characterizing the direct sound component of a VM are the source location $\boldsymbol{r}'_n$, $n = 1, \ldots, N$ and the exterior sound field coefficients $\beta^n_{\ell,\mu}(t, \omega)$, $n = 1, \ldots, N$. The positions $\boldsymbol{r}'_n$ of the sources are estimated using the acoustic source localization algorithm described in Sec. III-A.

The estimation of the exterior sound field coefficients $\beta^n_{\ell,\mu}(t, \omega)$ from the microphone signals requires the knowledge of the direct sound component $X_{n,\mathrm{dir}}(t, \omega, \boldsymbol{r}_i)$ at each microphone (see (8)). However, only the microphone signals $X(t, \omega, \boldsymbol{r}_i)$ are directly available. It follows that a procedure for estimating the direct and the diffuse components from $X(t, \omega, \boldsymbol{r}_i)$, $i = 1, \ldots, I$ is required. This procedure must be *blind* with respect to the room transfer function between sources and microphones. This, in fact, is a desirable feature, as measuring the transfer functions is not always feasible for all possible source locations and, in addition, transfer functions can be time-varying.

The algorithm for the estimation of the direct components is described in Sec. III-B. It is worth noticing that the algorithm used for estimating the direct component requires the knowledge of the exterior field spherical harmonic coefficients, which is detailed in the same section. Finally, given the acoustic scene parameters described above, it is possible to synthesize the VM signals. This is detailed in Sec. IV-A.

2) VM diffuse component estimation: The diffuse sound component $S_{\mathrm{diff}}(t, \omega, \check{\boldsymbol{r}}_v)$, can be estimated from the microphone diffuse sound components $X_{\mathrm{diff}}(t, \omega, \boldsymbol{r}_i)$, as detailed in Sec. IV-B. Inputs for the estimation of the diffuse components, as shown in Fig. 2, are the VM position $\check{\boldsymbol{r}}_v$, the microphone positions $\boldsymbol{r}_i$ and the VM sensitivity to the diffuse field $Q_v(\omega)$.

As defined in (2), an estimate of the vth VM signal is obtained as the linear combination of the estimates of the direct component $\hat{S}_{n,\mathrm{dir}}(t, \omega, \check{\boldsymbol{r}}_v)$ and the diffuse component $\hat{S}_{\mathrm{diff}}(t, \omega, \check{\boldsymbol{r}}_v)$.

III. Parameter Estimation

In this section we describe the first phase of the proposed virtual miking approach, which consists of estimating the model parameters described in Sec. II-A. In particular, as underlined in Sec. II-B, the parameters that have to be estimated are the position of the sources (Sec. III-A), the direct and the diffuse components (Sec. III-B) at each microphone, and the coefficients of the exterior sound field model (Sec.III-C).

A. Source Localization

The accurate estimation of the source location is a crucial step, as the estimation of all the other parameters depend on that. Furthermore, it is well-known in the literature [29], [30] that accurate source localization in the presence of strong reverberation is a challenging problem. In this manuscript, we approach the source localization problem as a two-step procedure: in the first step (Sec. III-A1) we estimate a set of source Directions of Arrival (DOAs) for each array, while in the second step (Sec. III-A2) the locations of the sources are found solving the DOAs association and triangulation problem [31].

1) DOA estimation: In the literature different DOA estimation algorithms can be found and they can be mainly divided into two classes: parametric [32]–[34] and spatial methods [35]–[37]. The former class of techniques leverages on assumptions about the covariance structure of the signals, while the latter concerns the computation of a spatial filter, customarily through beamforming.

Here, we adopt a localization based on spatial filtering, where the energy of the directional components of the sound field is evaluated with respect to the reference point $\boldsymbol{v}_a = [x_a, y_a]^T$, i.e., the centroid (1) of the ath array. The pseudospectrum $\Lambda^{(a)}(\alpha, t, \omega)$ of the ath array can be computed as the absolute value of the beamformer output for all the possible directions $\alpha \in (0, 2\pi]$.

As we are interested only in localizing the acoustic sources, similarly to [38]–[40] we compute a wideband extension of the pseudospectra by averaging $\Lambda^{(a)}(\alpha, t, \omega)$ as follows

$$\bar{\Lambda}^{(a)}(t, \alpha) = \left\{ \prod_{w=1}^{W/2} \Lambda^{(a)}(\alpha, t, \omega_w) \right\}^{\frac{2}{W}}, \tag{10}$$

where W is the number of points in the discretized frequency axis, gaining robustness against spatial aliasing and frequency bands with low SNR. It is worth noticing that (10) is not the only possible choice to obtain a wideband pseudospectrum. However, as investigated in [40], this choice provides a wideband pseudospectrum with greater resolution, narrower main-lobe and side-lobes attenuation with respect to the one obtained with the arithmetic mean.

The DOAs of the sources can be measured as the directions $\bar{\alpha}_n^{(a)}(t)$, $n = 1, \ldots, N$ corresponding to the N highest peaks of $\bar{\Lambda}^{(a)}(t, \alpha)$, i.e.,

$$\bar{\boldsymbol{\alpha}}^{(a)}(t) = \mathcal{D}(\bar{\Lambda}^{(a)}(t, \alpha), N), \tag{11}$$

where $\mathcal{D}(\cdot, N)$ is the operator that returns the N highest peaks and $\bar{\boldsymbol{\alpha}}^{(a)}(t) = \left[\bar{\alpha}_1^{(a)}(t), \bar{\alpha}_2^{(a)}(t), \ldots, \bar{\alpha}_N^{(a)}(t)\right]^T$ is the $N \times 1$ vector of the estimated DOAs for the ath array. Obviously, the presence of reflections due to the reverberation is likely to introduce errors in the estimate of the DOAs $\bar{\alpha}_n^{(a)}(t)$ in (11). It is also worth noticing that, in general, correlation between the source signals can negatively affect the estimation of the DOAs. The use of sufficiently short time windows in the STFT and the assumption of uncorrelation among the time-frequency bins, however, attenuates this problem.

With the aim of reducing the impact of reverberation on the location accuracy, we select the DOAs in $\bar{\boldsymbol{\alpha}}^{(a)}(t)$ compatible with source locations inside the ROI described in Sec. II. This is done by intersecting the half-lines with origin $\boldsymbol{v}_a$ and direction $\boldsymbol{D}^{(a)}(t) = \left[\boldsymbol{d}_1^{(a)}(t), \boldsymbol{d}_2^{(a)}(t), \ldots, \boldsymbol{d}_N^{(a)}(t)\right]$, $\boldsymbol{d}_n^{(a)}(t) = [\cos \bar{\alpha}_n^{(a)}(t), \sin \bar{\alpha}_n^{(a)}(t)]^T$ with the polygon $\mathcal{R}$ that defines the ROI

$$\tilde{\boldsymbol{\alpha}}^{(a)}(t) = \mathcal{I}(\boldsymbol{v}_a, \boldsymbol{D}^{(a)}(t), \mathcal{R}), \tag{12}$$

where $\mathcal{I}$ is the operator that returns all the DOAs for which at least one intersection with the polygon edges exists and $\tilde{\boldsymbol{\alpha}}^{(a)}(t) = \left[\tilde{\alpha}_1^{(a)}(t), \tilde{\alpha}_2^{(a)}(t), \ldots, \tilde{\alpha}_{\tilde{N}}^{(a)}(t)\right]^T$ is the resulting DOA vector with dimensions $\tilde{N} \times 1$, $\tilde{N} \leq N$.

2) Association of DOAs and triangulation: Once the DOAs for each array have been estimated, the source locations in Cartesian coordinates are estimated through triangulation. In a multi-source scenario the problem of DOA disambiguation arises, i.e. the matching of the DOAs measured from different arrays corresponding to the same source. In this manuscript, we tackle the disambiguation and triangulation problems employing a localization method based on the Distributed Ray Space Transform (DRST) [17], [18], [41].

The DRST, introduced in [18], is a tool devoted to the mapping of the signals of distributed arrays onto a domain, called *Projective Ray Space* (PRS) [42]. The PRS, derived as a generalization of the *ray space* [38], is the domain of representation of the sound field in the scenarios where the same acoustic scene is observed by multiple viewpoints. This parameterization is based on the description of the plenacoustic function [43] in terms of the acoustic rays. The PRS is defined by the parameters of the implicit equation $l_1 x + l_2 y + l_3 = 0$ that identifies an acoustic ray in 2D. The distinctive characteristics of such parameterization, is that the acoustic primitives, such as sources and reflectors, are mapped in the PRS onto linear subspaces or combinations thereof.

Let us consider a generic point-like acoustic source at a given time instant t placed in $\boldsymbol{r}'(t) = [x'(t), y'(t)]^T$ as in Fig. 3(a). This can be seen as the point originating acoustic rays and, given the source location in homogeneous coordinates $\bar{\boldsymbol{r}}'(t) = [x'(t), y'(t), 1]^T$, a ray emitted by the source satisfies

$$\mathbf{l}^T \bar{\boldsymbol{r}}'(t) = 0, \tag{13}$$

where $\mathbf{l} = \varepsilon[l_1, l_2, l_3]^T, \varepsilon \neq 0$ are the parameters of the projective line describing the ray. As described in [42], the representation of $\boldsymbol{r}'(t)$ is given by the set of rays passing through it, and in the PRS corresponds to a plane (see Fig. 3(b)). DOAs in $\tilde{\alpha}_{\tilde{n}}^{(a)}(t), \tilde{n} = 1, \ldots, \tilde{N}$ in (12) are converted in acoustic rays in the PRS through

$$\begin{aligned}
l_{1,\tilde{n}}^{(a)}(t) &= \varepsilon \sin\left(\tilde{\alpha}_{\tilde{n}}^{(a)}(t)\right); \\
l_{2,\tilde{n}}^{(a)}(t) &= \varepsilon \cos\left(\tilde{\alpha}_{\tilde{n}}^{(a)}(t)\right); \\
l_{3,\tilde{n}}^{(a)}(t) &= \varepsilon[y_a \cos\left(\tilde{\alpha}_{\tilde{n}}^{(a)}(t)\right) - x_a \sin\left(\tilde{\alpha}_{\tilde{n}}^{(a)}(t)\right)], \varepsilon > 0.
\end{aligned} \tag{14}$$

These points will form clusters in the PRS on the planes representing the acoustic sources. In order to associate DOAs to sources and then proceed to the localization task, we adopt techniques of pattern analysis. More precisely, we use a RANSAC algorithm [44] (random maximum consensus) over the set of $\tilde{N} \times A$ points in the PRS. Let us indicate with the subscript $\hat{n}$ the points identified by RANSAC as pertaining to the nth source

$$\hat{\mathbf{l}}_{\hat{n}}(t) = [l_{1,\hat{n}}(t), l_{2,\hat{n}}(t), l_{3,\hat{n}}(t)]^T. \tag{15}$$

Note that $\hat{\mathbf{l}}_{\hat{n}}$ no longer depends on the array index a. The points in (15) are then re-arranged in matrix form such that the condition of (13) becomes

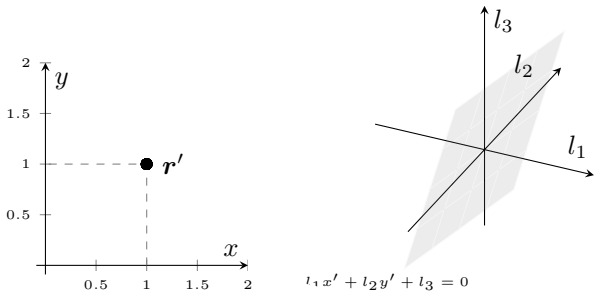
$$\mathbf{L}_n(t) \bar{\boldsymbol{r}}'_n(t) = 0, \tag{16}$$

where $\mathbf{L}_n(t) = [\hat{\mathbf{l}}_1(t), \ldots, \hat{\mathbf{l}}_{\hat{N}}(t)]^T$ and $\hat{N}$ is the number of rays related to the nth source. Finally, the estimate of the source location $\hat{\boldsymbol{r}}'_n(t)$ is obtained as [42]

$$\hat{\boldsymbol{r}}'_n(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{v}_n(t), \tag{17}$$

where $\mathbf{v}_n$ is the singular vector associated to the smallest singular value of the singular value decomposition of the matrix $\mathbf{L}_n^T(t)\mathbf{L}_n(t)$.

It is worth noticing that, in the present section, we made explicit the dependence of the nth source position from the time instant t in order to show that the presented method can account for moving sources. However, for the sake of readability, we will omit such an explicit dependence for the rest of the manuscript.



(a) (b)

Fig. 3. A point-like source in the geometric space (a) is mapped, in the projective ray space (b), onto a plane with normal direction $\bar{\boldsymbol{r}}' = [x', y', 1]^T$

B. Estimation of Direct and Diffuse Components

Once the source locations are obtained, we address the problem of estimating the direct and diffuse components of a microphone signal namely $X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)$ and $X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)$ from the recorded microphone signal $X(t,\omega,\boldsymbol{r}_i)$. This is a crucial step of the process, as the knowledge of $X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)$ is required to estimate the exterior sound field coefficients of the sources (see (8)) and $X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)$ is needed for the estimation of the VM diffuse component $S_{\mathrm{diff}}(t,\omega,\breve{\boldsymbol{r}}_v)$.

The estimation of the direct and the diffuse component is also known as the dereverberation problem. The dereverberation algorithms proposed in the literature can be divided in two categories: inverse filtering algorithms [45]–[47]; and algorithms that estimate and suppress reverberation with spectral subtraction or Wiener filtering [20], [48]. From an operative standpoint, the two categories differ in the fact that the former requires the knowledge of the room transfer function, while the latter does not need it. In this work, as stated in Sec. II-B, the second class meets our requirements.

Following [20] and [48] we can obtain an estimate of the direct sound component $X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)$ at the position $\boldsymbol{r}_i$ as the output of a squared root Wiener filter whose coefficients are computed as [20], [49]

$$G_{\mathrm{dir}}(t,\omega,\boldsymbol{r}_i) = \sqrt{1 - \frac{1}{\mathrm{CDR}(t,\omega,\boldsymbol{r}_i)+1}}, \tag{18}$$

where $\mathrm{CDR}(t,\omega,\boldsymbol{r}_i)$ is the time-frequency dependent signal to diffuse ratio at the ith microphone, defined as

$$\mathrm{CDR}(t,\omega,\boldsymbol{r}_i) = \frac{\Phi_{\mathrm{dir},ii}(t,\omega)}{\Phi_{\mathrm{diff},ii}(t,\omega)}. \tag{19}$$

Here $\Phi_{\mathrm{dir},ii}$ and $\Phi_{\mathrm{diff},ii}$ are the auto-power spectra of the direct and diffuse component, respectively, and are defined as

$$\begin{aligned}\Phi_{\mathrm{dir},ii}(t,\omega) &= E\{X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)\}\\ \Phi_{\mathrm{diff},ii}(t,\omega) &= E\{X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)\}.\end{aligned} \tag{20}$$

As shown in [20], an estimate of $\mathrm{CDR}(t,\omega,\boldsymbol{r}_i)$ can be obtained from the knowledge of the microphone signal coherence function and the diffuse noise coherence function. This can be accomplished using the CDR estimator defined in [20] as (21), where $\mathrm{Re}\{\cdot\}$ is the operator that retrieves the real part of a complex number. The dependencies on time and frequency have been omitted for the sake of readability.

The term $\Gamma_{\mathrm{diff},ii'}(\omega)$ in (21) is the diffuse noise coherence function between the ith and i'th microphones. Assuming a spherically isotropic sound field as in (7), $\Gamma_{\mathrm{diff},ii'}(\omega)$ can be modelled as [20]

$$\Gamma_{\mathrm{diff},ii'}(\omega) = \frac{\Phi_{\mathrm{diff},ii'}(t,\omega)}{\sqrt{\Phi_{\mathrm{diff},ii}(t,\omega)\Phi_{\mathrm{diff},i'i'}(t,\omega)}} = \frac{\sin(kd_{ii'})}{kd_{ii'}}, \tag{22}$$

where

$$\begin{aligned}\Phi_{\mathrm{diff},ii'}(t,\omega) &= E\{X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)X^{\star}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_{i'})\},\\ d_{ii'} &= \|\boldsymbol{r}_i - \boldsymbol{r}_{i'}\|_2\end{aligned} \tag{23}$$

with $\|\cdot\|_2$ the ℓ-2 norm of a vector.

The term $\hat{\Gamma}_{ii'}(t,\omega)$ in (21) is the estimate of the microphone signal coherence function between the ith and the i'th microphone. If we assume that the sensor noise between microphones i and i' is uncorrelated, the microphone signal coherence function can be estimated as [48]

$$\begin{aligned}&\hat{\Gamma}_{ii'}(t,\omega)\\ &= \frac{\hat{\Phi}_{ii'}(t,\omega)}{\sqrt{\hat{\Phi}_{ii}(t,\omega) - \hat{\Phi}_{N,ii}(t,\omega)}\sqrt{\hat{\Phi}_{i'i'}(t,\omega) - \hat{\Phi}_{N,i'i'}(t,\omega)}},\end{aligned} \tag{24}$$

where $\Phi_{\mathrm{N},ii}(t,\omega)$ is the noise auto-power spectrum defined in (9) and

$$\Phi_{ii'}(t,\omega) = E\{X(t,\omega,\boldsymbol{r}_i)\,X^{\star}(t,\omega,\boldsymbol{r}_{i'})\}. \tag{25}$$

The auto and cross spectra can be obtained from the microphone signals by recursive averaging [20]

$$\hat{\Phi}_{ii'}(t,\omega) = \lambda\hat{\Phi}_{ii'}(t-1,\omega) + (1-\lambda)X(t,\omega,\boldsymbol{r}_i)\,X^{\star}(t,\omega,\boldsymbol{r}_{i'}) \tag{26}$$

where λ is a constant in the range $[0,1)$. In our scenario, the microphone pairs are chosen as belonging to the same array. The sensor noise auto-spectra $\hat{\Phi}_{N,ii}(t,\omega)$ and $\hat{\Phi}_{N,i'i'}(t,\omega)$ can be obtained applying recursive averaging on the microphone signals as in (26) when neither acoustic sources nor diffuse noise are present (i.e., only the sensor noise component is active). In order to determine the activity or inactivity of the sources we use the voice activity detector [50]. It is worth noting that [50] assumes that all sources emit speech signals that are sufficiently sparse in the time-frequency domain thus agreeing with the assumption stated in Sec. II-A.

Once an estimate of the CDR at the ith microphone is obtained using (21) we can use (18) to obtain the Wiener filter

$$\mathrm{CDR}(\boldsymbol{r}_i) = \frac{\Gamma_{\mathrm{diff},ii'}\mathrm{Re}\{\hat{\Gamma}_{ii'}\} - |\hat{\Gamma}_{ii'}|^2}{|\hat{\Gamma}_{ii'}|^2 - 1} - \frac{\sqrt{\left(\Gamma_{\mathrm{diff},ii'}\mathrm{Re}\{\hat{\Gamma}_{ii'}\}\right)^2 - \left(\Gamma_{\mathrm{diff},ii'}|\hat{\Gamma}_{ii'}|\right)^2 + (\Gamma_{\mathrm{diff},ii'})^2 - 2\Gamma_{\mathrm{diff},ii'}\mathrm{Re}\{\hat{\Gamma}_{ii'}\} + |\hat{\Gamma}_{ii'}|^2}}{|\hat{\Gamma}_{ii'}|^2 - 1} \tag{21}$$

coefficients that allows to extract the direct component of a the ith microphone signal. However, as highlighted in [20], a more practical implementation of (18) is given by [49]

$$G_{\text{dir}}(t,\omega,\boldsymbol{r}_i) = \max\left\{G_{\min}, \sqrt{1-\frac{\nu}{\text{CDR}(t,\omega,\boldsymbol{r}_i)+1}}\right\}, \tag{27}$$

where ν is the oversubtraction factor and $G_{\min}$ the gain floor. The term ν controls the amount of noise subtracted from the noisy signal. For full noise subtraction, $\nu = 1$ and for over-subtraction $\nu > 1$. The term $G_{\min}$ acts as a lower bound for the filter coefficients weights. This is useful in order to reduce artefacts in the output signal. Inspecting (27), it is clear that high values of CDR leads to low filter gain and vice versa.

Finally, the filter in (27) is used to compute the direct signal component at the ith microphone through [20]

$$\hat{X}_{\text{dir}}(t,\omega,\boldsymbol{r}_i) = G_{\text{dir}}(t,\omega,\boldsymbol{r}_i)U(t,\omega,\boldsymbol{r}_i), \tag{28}$$

where

$$U(t,\omega,\boldsymbol{r}_i) = \sqrt{\frac{Z\left(t,\omega,\boldsymbol{r}_i\right)+Z\left(t,\omega,\boldsymbol{r}_{i'}\right)}{2}}e^{j\arg\{X(t,\omega,\boldsymbol{r}_i)\}}, \tag{29}$$

with $Z\left(t,\omega,\boldsymbol{r}_i\right) = \left|X\left(t,\omega,\boldsymbol{r}_i\right)\right|^2 - \hat{\Phi}_{N,ii}(t,\omega)$, $Z\left(t,\omega,\boldsymbol{r}_{i'}\right) = \left|X\left(t,\omega,\boldsymbol{r}_{i'}\right)\right|^2 - \hat{\Phi}_{N,i'i'}(t,\omega)$ and $\arg\{\cdot\}$ the operator that takes the argument of a complex number. The spatial magnitude averaging performed in (29) is typically used in order to reduce the variance of the estimates for microphone array post-filters [51], [52].

The diffuse component of the microphone signal can be obtained using the filter [48]

$$G_{\text{diff}}(t,\omega,\boldsymbol{r}_i) = \sqrt{1-\left[G_{\text{dir}}(t,\omega,\boldsymbol{r}_i)\right]^2}, \tag{30}$$

where $G_{\text{dir}}(t,\omega,\boldsymbol{r}_i)$ is defined in (27). It follows that an estimate of $X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)$ can be obtained as

$$\hat{X}_{\text{diff}}(t,\omega,\boldsymbol{r}_i) = G_{\text{diff}}(t,\omega,\boldsymbol{r}_i)U(t,\omega,\boldsymbol{r}_i), \tag{31}$$

where $U(t,\omega,\boldsymbol{r}_i)$ is defined in (29). As demonstrated in Appendix A, using the filters in (27) and (30) and assuming that $\nu = 1, G_{\min} = 0$ and that the auto-spectra of the direct, diffuse and noise components at the i microphone and at the i' microphone are the same, the power of the estimated sound field components corresponds to the actual sound power (i.e., $E\{|\hat{X}_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} = E\{|X_{n,\text{dir}}(t,\omega,\boldsymbol{r}_i)|^2\}$ and $E\{|\hat{X}_{\text{diff}}(t,\omega,\boldsymbol{r}_i)|^2\} = E\{|X_{\text{diff}}(t,\omega,\boldsymbol{r}_i)|^2\}$) [49].

C. Exterior Sound Field Coefficients Estimation

Once the direct signal components of the microphone signals have been estimated using (28), and the sources have been localized, we can exploit the model of the direct sound component in (8) in order to estimate the set of spherical harmonics coefficients related to each source in the acoustic scene. Let us define the vector $\hat{\mathbf{x}}_{\text{dir}}(t,\omega)$ containing the estimates of the direct component for all the microphones, i.e.,

$$\left[\hat{\mathbf{x}}_{\text{dir}}(t,\omega)\right]_i = \hat{X}_{\text{dir}}(t,\omega,\boldsymbol{r}_i) \qquad i = 1,\dots,I, \tag{32}$$

where $[\cdot]_i$ is the ith element of the vector. We denote the vector of the coefficients of the spherical harmonic for the nth source as

$$\boldsymbol{\beta}_n(t,\omega) = \left[\beta_{00}^n(t,\omega), \beta_{0-1}^n(t,\omega),\dots,\beta_{LL}^n(t,\omega)\right]^T, \tag{34}$$

where $(\cdot)^T$ is the transpose operator.

Let us define the matrix $\widehat{\mathbf{Y}}_n(k)$ containing the spherical harmonics as in (33), where $\hat{\rho}_{i,n}$, $\hat{\theta}_{i,n}$ and $\hat{\phi}_{i,n}$. are the estimates of $\rho_{i,n}$, $\theta_{i,n}$ and $\phi_{i,n}$ defined in (8) obtained using the estimate of the n source position $\hat{\boldsymbol{r}}'_n$. In the light of the definitions in (32) and (33), the direct sound components acquired by the microphones are given by

$$\begin{aligned}\hat{\mathbf{x}}_{\text{dir}}(t,\omega) &= \left[\widehat{\mathbf{Y}}_1(k)\widehat{\mathbf{Y}}_2(k)\cdots\widehat{\mathbf{Y}}_N(k)\right]\begin{bmatrix}\boldsymbol{\beta}_1(t,\omega)\\ \vdots \\ \boldsymbol{\beta}_N(t,\omega)\end{bmatrix}, \\ &= \widehat{\mathbf{Y}}(k)\boldsymbol{\beta}(t,\omega).\end{aligned} \tag{35}$$

An estimate $\hat{\boldsymbol{\beta}}(t,\omega)$ of $\boldsymbol{\beta}(t,\omega)$ can be obtained as

$$\hat{\boldsymbol{\beta}}(t,\omega) = \widehat{\mathbf{Y}}^{\dagger}(k)\hat{\mathbf{x}}_{\text{dir}}(t,\omega), \tag{36}$$

where † denotes the matrix pseudo-inverse. However, under the assumption that only one source is dominant in each time-frequency bin, we can solve (35) by enforcing the sparsity of the resulting coefficients vector. In particular, we obtain $\hat{\boldsymbol{\beta}}(t,\omega)$ as the result of a group lasso optimization problem [53], i.e.,

$$\begin{aligned}\hat{\boldsymbol{\beta}}(t,\omega) = \underset{\boldsymbol{\beta}(t,\omega)}{\text{argmin}}\ &\frac{1}{2}\|\widehat{\mathbf{Y}}(k)\boldsymbol{\beta}(t,\omega) - \hat{\mathbf{x}}_{\text{dir}}(t,\omega)\|_2^2 \\ &+ \kappa\sum_{n=1}^{N}\|\boldsymbol{\beta}_n(t,\omega)\|_2.\end{aligned} \tag{37}$$

As shown in [54], this problem can be solved using the alternating direction method of multipliers (ADMM).

Discussion: It is worth noticing that, since sources and microphones are assumed to lie on the same plane (i.e., $\theta = \pi/2$), the columns of $\widehat{\mathbf{Y}}_n(k)$ for which $\ell + |\mu|$ is even have been removed. As shown in [55], [56], in fact, when $\theta = \pi/2$ the summation in (4) goes to zero since $Y_{\ell\mu}\left(\pi/2, \phi_{v,n}\right) = 0$.

The truncation order L is usually set as $L = \lceil keR_s/2 \rceil$ where e is the Euler's number, $\lceil\cdot\rceil$ is the ceiling operator and R_s is the radius of the region surrounding a source [1]. Hence, the truncation order should be a function of the source but, in order to simplify the notation, in this manuscript we assumed that the truncation order is the same for all the sources. As stated in [57], the radius R_s and, as a consequence, the value of L can be reduced with a suitable choice of the origin of the reference frame. In this manuscript, the origin coincides with the coordinates of the sources location estimates. Moreover, as it is clear from (36), in order for the system to be over-determined, the following condition should be satisfied: $\left[(L+1)^2 - T_L\right]N < I$, where $T_L = L(L+1)/2$. However, considering the assumption that only one source is dominant in each time-frequency bin, the above-mentioned condition can be relaxed as $\left[(L+1)^2 - T_L\right] < I$. It follows that $L = \min\left(\lceil keR_s/2\rceil, (\sqrt{8I+1}-3)/2\right)$.

IV. SYNTHESIS

In this section we show how the signal of the vth VM modelled as in (2) can be estimated from the set of parameters obtained in Sec. III. In particular, in Sec. IV-A we show how the direct component of the vth VM signal can be estimated while in Sec. IV-B we deal with the diffuse sound component.

A. Synthesis of the Direct Component

An estimate $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ of the direct sound component at the vth VM due to the nth source can be obtained by exploiting the model in (4). More precisely, given the estimate $\hat{\boldsymbol{r}}'_n$ of the nth source position obtained in (17), and the set of exterior field coefficients $\hat{\boldsymbol{\beta}}_n(t,\omega)$ obtained in (37), $\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)$ is obtained through

$$\begin{aligned}&\hat{S}_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)\\&\quad=\sum_{\ell=0}^{L}\sum_{\mu=-\ell}^{\ell}\hat{\beta}^n_{\ell\mu}(t,\omega)h_\ell\left(k\hat{\check{\rho}}_{v,n}\right)Y_{\ell\mu}\left(\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n}\right),\end{aligned}\tag{38}$$

where $\hat{\check{\rho}}_{v,n}$, $\hat{\check{\theta}}_{v,n}$ and $\hat{\check{\phi}}_{v,n}$ are the estimates of $\check{\rho}_{v,n}$, $\check{\theta}_{v,n}$ and $\check{\phi}_{v,n}$ in (4) and can be computed by inserting in (3) the estimate $\hat{\boldsymbol{r}}'_n$ of the nth source location.

The term $C_v(\omega)$ in (2) models the VMs pick-up pattern. Usually this term can be expressed as a function $f(\cdot)$ that depends on the frequency ω, the angle between the vth VM and the nth source $[\iota_{n,v},\zeta_{n,v}]^T=\angle\boldsymbol{r}'_n-\check{\boldsymbol{r}}_v$, with ι the azimuth and ζ the elevation, the orientation $\check{\boldsymbol{o}}_v$ of the vth VM,

$$C_v(\iota_{n,v},\zeta_{n,v},\check{\boldsymbol{o}}_v,\omega)=f\left(\iota_{n,v},\zeta_{n,v},\check{\boldsymbol{o}}_v,\omega\right).\tag{39}$$

The proposed framework enables also the estimation of the direct signal component acquired by a higher order VM located at $\check{\boldsymbol{r}}_v$. In particular, given the estimate $\hat{\boldsymbol{\beta}}_n(t,\omega)$ of the exterior sound field produced by the nth source, the spherical harmonics coefficients acquired by a Z-order VM can be directly obtained through the spherical harmonics addition theorem [58] through

$$\begin{aligned}\gamma^{v,n}_{zb}(t,\omega)&=\sum_{l=0}^{L}\sum_{\mu=-l}^{l}\hat{\beta}^n_{l\mu}(t,\omega)T^{\mu b}_{lz}(k,\hat{\check{\rho}}_{v,n},\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n}),\\&\qquad z=0,\ldots,Z,\quad b=-z,\ldots,z\end{aligned}\tag{40}$$

where

$$\begin{aligned}&T^{\mu b}_{lz}(k,\hat{\check{\rho}}_{v,n},\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n})\\&\quad=4\pi j^{z-l}\sum_{g=0}^{l+z+1}j^g(-1)^{2\mu-b}h_g(k\hat{\check{\rho}}_{v,n})Y^*_{g(b-\mu)}(\hat{\check{\theta}}_{v,n},\hat{\check{\phi}}_{v,n})\\&\qquad\times\sqrt{(2l+1)(2z+1)(2g+1)/(4\pi)}W_1W_2,\end{aligned}\tag{41}$$

with

$$W_1=\begin{pmatrix}l&z&g\\0&0&0\end{pmatrix},\quad W_2=\begin{pmatrix}l&z&g\\\mu&-b&(b-\mu)\end{pmatrix}\tag{42}$$

the Wigner $3-j$ symbols [59]. The knowledge of the spherical harmonics coefficients acquired by the Z-order microphone enables applications like, for example, modal beamforming [60], binaural synthesis [61], [62] etc. In all these applications the spherical harmonics coefficients of the higher order microphone are filtered through

$$C_v(\omega)S_{n,\mathrm{dir}}(t,\omega,\check{\boldsymbol{r}}_v)=\sum_{z=0}^{Z}\sum_{b=-z}^{z}\left[\psi^{v,n}_{zb}(\omega)\right]^*\gamma^{v,n}_{zb}(t,\omega),\tag{43}$$

where $\psi^{v,n}_{zb}(\omega)$ are the spherical harmonics filter coefficients [63].

B. Synthesis of the Diffuse Component

In order to synthesize the diffuse component $S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v)$ of the VM, we use the estimates of the diffuse signal at each microphone obtained in (31). More precisely, given the estimates $\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)$, we compute the power of the diffuse signal component in $\check{\boldsymbol{r}}_v$ as [14]

$$E\{|S_{\mathrm{diff}}(t,\omega,\check{\boldsymbol{r}}_v)|^2\}=\sum_{i=1}^{I}\varpi_i(\omega)E\{|\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\},\tag{44}$$

where $\sum_{i=1}^{I}\varpi_i(\omega)=1$. As stated in [14], $\varpi_i(\omega)$ are real valued weights that depend on the estimation variance of the power of the signal at the ith microphone (i.e., $E\{|X(t,\omega,\boldsymbol{r}_i)|^2\}$), $i=1,\ldots,I$. Since these estimates are usually unavailable, we choose the weights to be inversely proportional with respect to the distance between the vth VM and the ith microphone, i.e.,

$$\varpi_i(\omega)=\frac{1}{\|\boldsymbol{r}_i-\check{\boldsymbol{r}}_v\|_2}\left(\sum_{i'=1}^{I}\frac{1}{\|\boldsymbol{r}_{i'}-\check{\boldsymbol{r}}_v\|_2}\right)^{-1}.\tag{45}$$

For what concerns the estimation of the phase of the diffuse signal component, we experimentally verify that plausible results can be achieved by using as an estimate the phase of the nearest microphone. It is worth noting that such a solution does not provide the desired spatial coherence among closely spaced VMs. It follows that, if one want to further process the signal coming from different VMs, this fact must be taken into account. Nevertheless, as we will see in the following section, this solution is applicable to coincident VMs for simulating, for example, a stereo recording scenario.

Finally, as defined in (2), the function $Q_v(\omega)$ controls the sensitivity of the VM to the diffuse field. In the most general

$$\widehat{\mathbf{Y}}_n(k)=\begin{bmatrix}h_0\left(k\hat{\rho}_{1,n}\right)Y_{00}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right)&h_1\left(k\hat{\rho}_{1,n}\right)Y_{1-1}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right)&\cdots&h_L\left(k\hat{\rho}_{1,n}\right)Y_{LL}\left(\hat{\theta}_{1,n},\hat{\phi}_{1,n}\right)\\h_0\left(k\hat{\rho}_{2,n}\right)Y_{00}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right)&h_1\left(k\hat{\rho}_{2,n}\right)Y_{1-1}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right)&\cdots&h_L\left(k\hat{\rho}_{2,n}\right)Y_{LL}\left(\hat{\theta}_{2,n},\hat{\phi}_{2,n}\right)\\\vdots&\vdots&\ddots&\vdots\\h_0\left(k\hat{\rho}_{I,n}\right)Y_{00}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right)&h_1\left(k\hat{\rho}_{I,n}\right)Y_{1-1}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right)&\cdots&h_L\left(k\hat{\rho}_{I,n}\right)Y_{LL}\left(\hat{\theta}_{I,n},\hat{\phi}_{I,n}\right)\end{bmatrix}\tag{33}$$

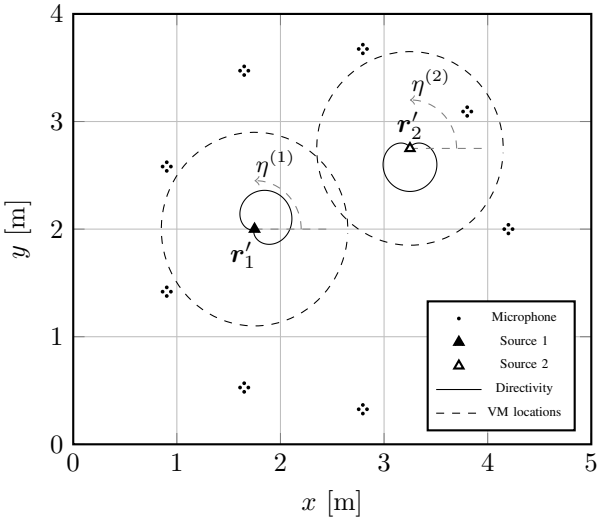


Fig. 4. 2D graphical representation of the first simulation setup. The two sources have a first-order cardioid directivity and they are located in $\boldsymbol{r}'_1 = [1.75, 2]^T$m and $\boldsymbol{r}'_2 = [3.25, 2.75]^T$m.

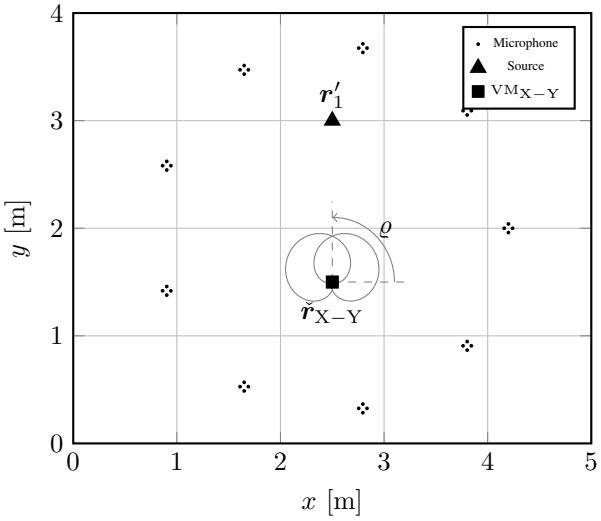


Fig. 5. 2D graphical representation of the second simulation setup. The source is located at $\boldsymbol{r}'_1 = [2.5, 3]^T$m while the VMs in X-Y configuration are placed at $\check{\boldsymbol{r}}_{\mathrm{X-Y}} = [2.5, 1.5]^T$m.

case, this function can be arbitrarily designed. However, in most cases the relationship between $Q_v(\omega)$ and the pick-up pattern of the VM directivity is well approximated by

$$Q_v(\omega) = \sqrt{\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} C_v(\iota, \zeta, \check{\boldsymbol{o}}_v, \omega)^2 \, \sin\zeta \, d\iota \, d\zeta}. \quad (46)$$

The expression in (46) is valid under the assumption of a spherically isotropic sound field and accounts for the fact that the diffuse component of the VM can be attenuated depending on the pick-up pattern $C_v(\iota, \zeta, \check{\boldsymbol{o}}_v, \omega)$.

V. Validation and Results

The proposed virtual miking technique is suitable for many applications thanks to its flexibility in terms of setup configuration. In addition, the possibility of completely characterizing the VM and the intrinsic source model paves the way to the use in advanced spatial audio applications. In order to validate the virtual miking performance, we tested the system through an extensive software simulation campaign. This allows us to control both the characteristics of the sources, such as their directivity pattern and the number and the location of the virtual microphones. Moreover, this eases the development of test cases that require a significant amount of reference VMs, barely deployable in practice. The simulation setup is introduced in Sec. V-A, while in Sec. V-B the different metrics used for the assessment of the VM performance are defined. The simulation results are discussed in Sec V-C.

A. Simulation Setup and Parameters

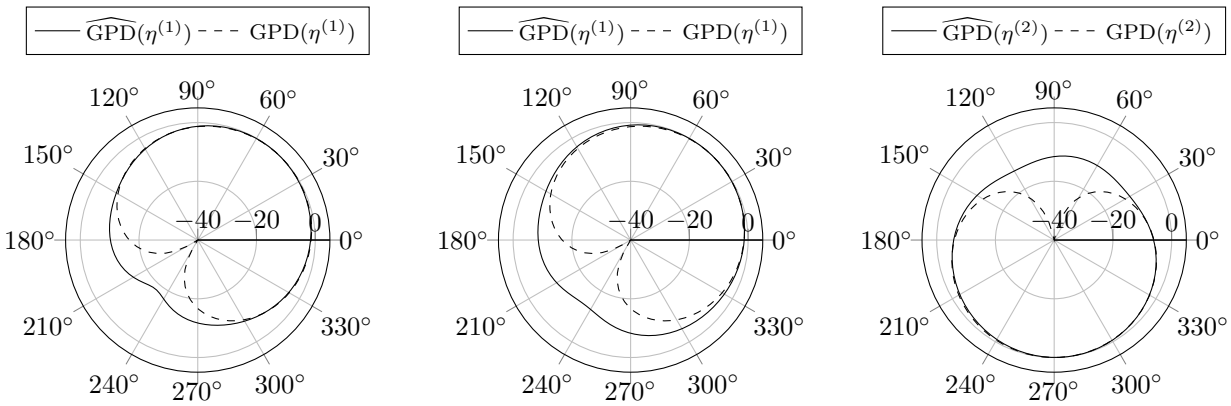
The simulation setup is illustrated in Fig. 4. It consists of $A = 9$ circular microphone arrays with radius 0.04m, accommodating $M = 4$ omnidirectional microphones each. Therefore, the total number of microphones is $I = A \times M = 36$. The sources emit two speech signals simultaneously (female and male taken from [64]), at $\boldsymbol{r}'_1 = [1.75, 2]^T$m and $\boldsymbol{r}'_2 = [3.25, 2.75]^T$m, respectively. When a single source setup is considered, only the source in $\boldsymbol{r}'_1$ is active in the scene at any time. The two sources present a first-order cardioid directivity with looking direction (i.e., direction of maximum energy emission) equal to 45 deg and 270 deg, respectively. A set of $V = V^{(1)} + V^{(2)}$ omnidirectional VMs (i.e., $C(\cdot) = 1$) are placed on two circumferences of radius R centered around the two sources (see Fig. 4). In detail, the positions of the VMs are defined as

$$\begin{aligned} \check{\boldsymbol{r}}_v^{(1)} &= R[\cos\eta_v^{(1)}, \sin\eta_v^{(1)}]^T + \boldsymbol{r}'_1, \quad v = 1, \ldots, V^{(1)} \\ \check{\boldsymbol{r}}_v^{(2)} &= R[\cos\eta_v^{(2)}, \sin\eta_v^{(2)}]^T + \boldsymbol{r}'_2, \quad v = 1, \ldots, V^{(2)}, \end{aligned} \quad (47)$$

where $\eta_v^{(1)} = 2\pi/V^{(1)}v$, $\eta_v^{(2)} = 2\pi/V^{(2)}v$ and $\check{\boldsymbol{r}}_v^{(1)}$, $\check{\boldsymbol{r}}_v^{(2)}$ are the positions of VMs surrounding the first and the second source, respectively. The actual number of VMs, V, depends on the simulation setup. This arrangement of the VMs allows us to capture the directional properties of the sources, testing the capability of the proposed virtual miking technique in rendering a spatial sound perception coherent with the VM position in the scene.

The signal at the microphones (see (7)) is simulated as the convolution between the signal of the sources and the room impulse response (RIR) computed through the image source method [65] implemented in [66]. The room is $5\,\mathrm{m} \times 4\,\mathrm{m} \times 3\,\mathrm{m}$ with a reverberation time of $T_{60} = 0.4\,\mathrm{s}$. As done in [20] and [67], the diffuse component in (7) is computed from the RIR late reverberation part. This can be accomplished suppressing the direct path and the early reflections using a cutoff time T_e set to a typical value of $0.05\,\mathrm{s}$ [68]. The additive noise component in (7) is simulated using a random white Gaussian noise, whose variance is set so that the desired signal to noise ratio at each sensor is $60\,\mathrm{dB}$. The signals are processed at a sampling rate of $16\,\mathrm{kHz}$ and their time-frequency representation is obtained through a 4096 points Short Time Fourier Transform (STFT) with a $0.256\,\mathrm{s}$ Hamming window adopted both in the analysis and synthesis phase and 87.5% overlap.

The localization is performed as described in Sec. III-A, where, for the computation of the pseudospectrum



(a) (b) (c)

Fig. 6. (a) GPD of the VMs and their references in a single source scenario. (b) GPD of the first source when both are active. (c) GPD of the second source when both are active.

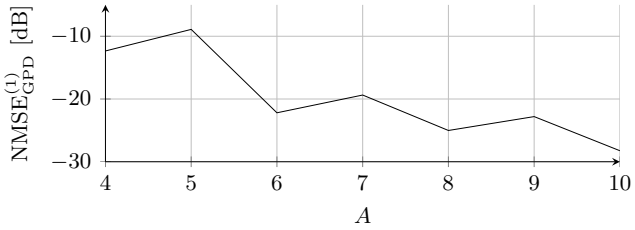
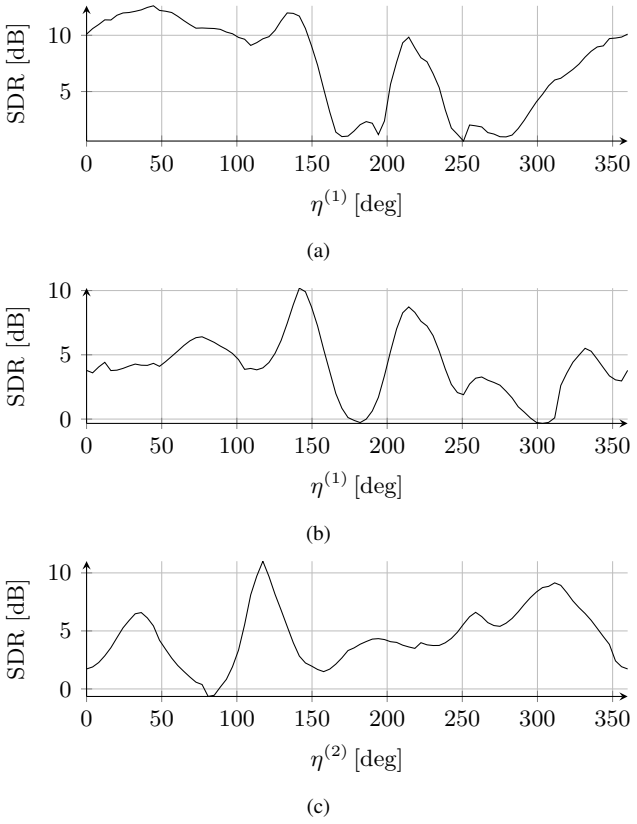


Fig. 7. The $\mathrm{NMSE}_{\mathrm{GPD}}$ as a function of the number of arrays A.

$\Lambda^{(a)}(\alpha,\omega), \forall a = 1,\dots,A$, we adopted the super directive beamformer [35]. The averaged pseudospectum in (10) is compute with $W = 4096$. The actual estimate of the source location $\hat{\boldsymbol{r}}'_n$ is obtained as the median value of 1000 RANSAC executions with different algorithm initializations at each time frame t. An estimate of the direct and diffuse components are obtained as presented in Sec. III-B with $\lambda = 0.68$, $\nu = 1.3$ and $G_{\min} = -30\,\mathrm{dB}$. Given the described setup, the pairs of microphones for the estimation of the cross spectra $\hat{\Phi}_{ii'}(t,\omega)$ are chosen as belonging to the same array by following their order in terms of azimuth with respect to each array center. For what concerns the source parameter estimation, we set the spherical harmonics expansion order in (4) according to the discussion in Sec. III-C.

Moreover, an applicative scenario regarding a spatial acquisition has been simulated.

We test the virtual miking technique in the context of a stereo recording, simulating a X-Y stereo miking setup. In Fig. 5 one omnidirectional source, aligned with the X-Y VM along the y axis, is present. The source emits a female speech signal similarly to the previous scenario. X-Y stereo recording requires the employment of two directional microphone ($\mathrm{VM_L}$ and $\mathrm{VM_R}$) with first-order cardioid pick-up pattern. The microphones are characterized by coincident location $\check{\boldsymbol{r}}_{\mathrm{L}} = \check{\boldsymbol{r}}_{\mathrm{R}} = \check{\boldsymbol{r}}_{\mathrm{X-Y}}$ and different pick-up pattern orientation (39), such that $|\check{\boldsymbol{o}}_{\mathrm{L}} - \check{\boldsymbol{o}}_{\mathrm{R}}| = [\pi/2, 0]^T$. Hence, given the X-Y looking direction $\check{\boldsymbol{o}}_{\mathrm{X-Y}} = [\varrho, 0]^T$, namely, the direction corresponding to the center of the stereo plane, the orientation of the VMs are defined as $\check{\boldsymbol{o}}_{\mathrm{L}}(\varrho) = [\varrho - \pi/4, 0]^T$



(a) (b) (c)

Fig. 8. (a) SDR value referred to a single source scenario. (b) SDR value of the first source when both are active. (c) SDR value of the second source when both are active.

and $\check{\boldsymbol{o}}_{\mathrm{R}}(\varrho) = [\varrho + \pi/4, 0]^T$. Notice that when it is not explicitly stated, we assume the same setup parameters of the previous scenario.

B. Metrics

Accordingly to the different simulation setups, we evaluated the virtual miking performance in terms of generalized power directivity (GPD), signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), direct-to-reverberant ratio (DRR)

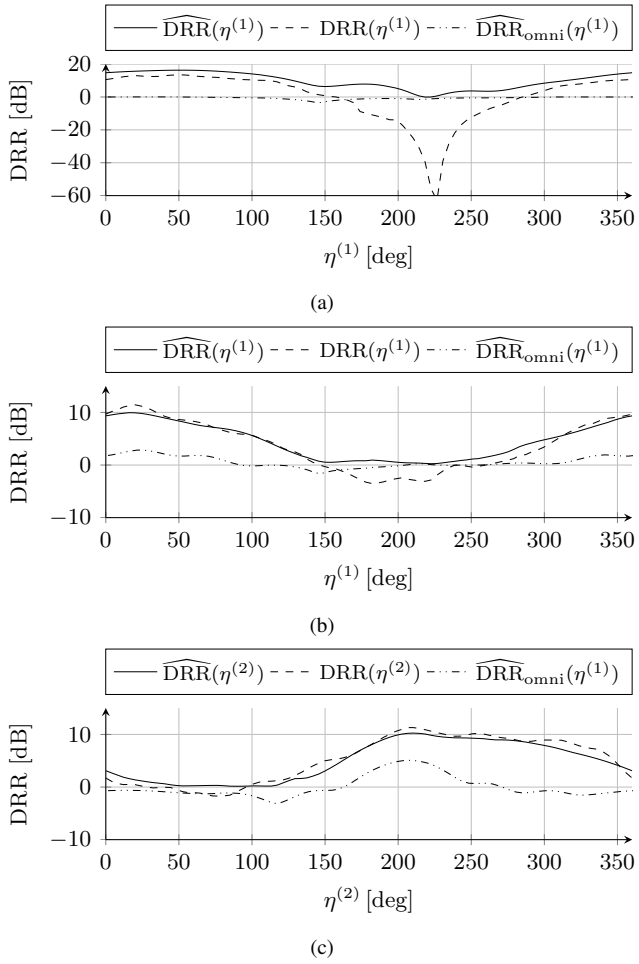


Fig. 9. (a) Estimated DRR and its reference in a single source scenario. (b) Estimated DRR and its reference of the first source when both sources are active. (c) Estimated DRR and its reference of the second source when both sources are active. The subscript omni refers to an estimate obtained assuming an omnidirectional source directivity (i.e., $L = 0$)

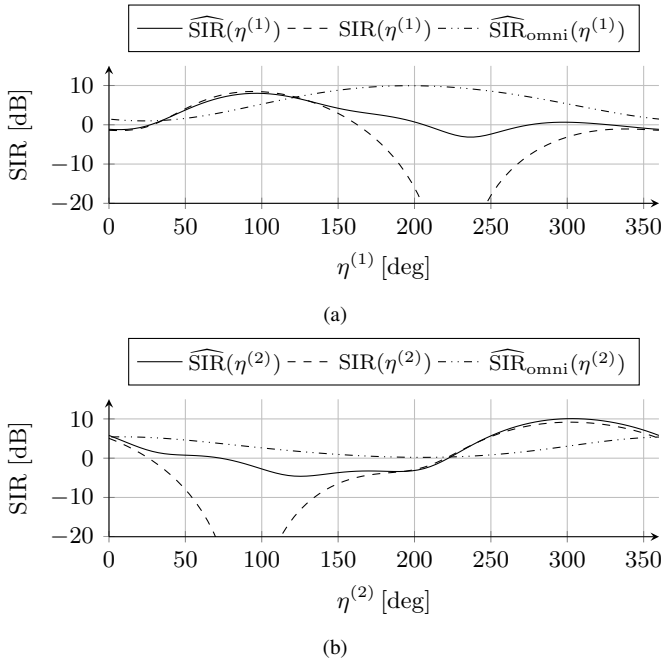


Fig. 10. (a) SIR value referred to the first source compared to its reference. (b) SIR value referred to the second source compared to its reference. The subscript omni refers to an estimate obtained assuming an omnidirectional source directivity (i.e., $L = 0$)

and interchannel level difference (ILD). The mathematical expression of the metrics is given in the following.

a) Generalized Power Directivity (GPD): The GPD of a source is defined as

$$\widehat{\mathrm{GPD}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}{\max\limits_v \sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}, \quad (48)$$

and it measures, for each source, the normalized power of the estimated VM signals surrounding the given source as a function of the VMs angle $\eta_v^{(n)}$ in (47). In the following we indicate with $\mathrm{GPD}(\eta_v^{(n)})$ the same metric computed with reference signals.

b) Signal-to-Distortion Ratio (SDR): The SDR is defined as the ratio between the desired reference signal and the distortion that affects the estimation (i.e. interference, noise and artifacts). We adopt the SDR estimator of [69] for the evaluation of the estimated VM signals with respect to the reference VM signals.

c) Signal-to-Interference Ratio (SIR): The SIR is defined as the ratio of the power of direct component of a desired signal and the sum of the interfering signals. More precisely,

$$\widehat{\mathrm{SIR}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}{\sum_{\bar{n}\neq n} \sum_w \sum_t |\hat{S}_{\bar{n},\mathrm{dir}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}. \quad (49)$$

In the following we indicate with $\mathrm{SIR}(\eta_v^{(n)})$ the same metric computed with reference signals.

d) Direct-to-Reverberant Ratio (DRR): The DRR represents the ratio between the power of the direct and reverberant components

$$\widehat{\mathrm{DRR}}(\eta_v^{(n)}) = \frac{\sum_w \sum_t |\hat{S}_{n,\mathrm{dir}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}{\sum_w \sum_t |\hat{S}_{\mathrm{diff}}(t,\omega_w,\check{\boldsymbol{r}}_v^{(n)})|^2}. \quad (50)$$

This metrics allows us to evaluate the VM in terms of spatial sound characteristics, since the DRR reflects the spatial properties of the signal. In the following we indicate with $\mathrm{DRR}(\eta_v^{(n)})$ the same metric computed with reference signals.

e) Interchannel Level Difference (ILD): The ILD is defined as the ratio between two VMs in a stereo configuration

$$\widehat{\mathrm{ILD}}(\varrho) = \frac{\sum_w \sum_t |\hat{S}(t,\omega_w,\check{\boldsymbol{r}}_{\mathrm{L}},\check{\boldsymbol{o}}_{\mathrm{L}}(\varrho))|^2}{\sum_w \sum_t |\hat{S}(t,\omega_w,\check{\boldsymbol{r}}_{\mathrm{R}},\check{\boldsymbol{o}}_{\mathrm{R}}(\varrho))|^2}. \quad (51)$$

where ϱ is the $\mathrm{VM}_{\mathrm{X-Y}}$ azimuth orientation. In (51) the dependence on the VM orientation is made explicit. In the following we indicate with $\mathrm{ILD}(\varrho)$ the same metric computed with reference signals.

As regards the GPD, SIR and DRR, we also evaluate the estimation in terms of the normalized mean squared error (NMSE). For instance concerning the GPD it is defined as

$$\mathrm{NMSE}^{(n)}_{\mathrm{GPD}} = 10\log_{10}\frac{\sum_v|\widehat{\mathrm{GPD}}(\eta_v^{(n)}) - \mathrm{GPD}(\eta_v^{(n)})|^2}{\sum_v|\mathrm{GPD}(\eta_v^{(n)})|^2}. \tag{52}$$

Moreover, for all the defined metrics we show the results in a decibel scale defined as $10\log_{10}(\cdot)$ of the relative metrics.

C. Results

1) Single-source scenario: The single source scenario is simulated using the setup of Fig. 4, when only the first source in $\boldsymbol{r}'_1$ is active. The estimated source position is $\hat{\boldsymbol{r}}'_1 = [1.7372, 2.0152]^T$ giving a localization error of $\|\hat{\boldsymbol{r}}'_1 - \boldsymbol{r}'_1\| = 0.0198$m. The number of VMs employed is $V = V^{(1)} = 90$, equally spaced around the source at distance from the source of $R = 0.9$m.

The $\widehat{\mathrm{GPD}}$, computed using the synthesized VM signals is reported in Fig 6, compared to the GPD computed with the reference signals. More specifically, Fig. 6 plots $\widehat{\mathrm{GPD}}$ and GPD in three different cases: a) the GDP of the source in $\boldsymbol{r}'_1$ when it is the only active source; b) the GDP of the source in $\boldsymbol{r}'_1$ when both sources are active; c) the GDP of the source in $\boldsymbol{r}'_2$ when both sources are active. In the context of the single source scenario we focus on Fig. 6(a). Notice that the $\widehat{\mathrm{GPD}}\left(\eta^{(1)}\right)$ fits the general trend of GPD with a $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ with respect to the reference of $-22.8\,\mathrm{dB}$. In order to evaluate the influence of the number of arrays A on the accuracy of the estimated source directivity, we performed a set of simulations varying $A \in \{4, 10\}$ with respect to the single-source scenario setup. In particular, the arrays are equally distributed on a circumference of radius 1.7 m centered in the room so that the setup with $A = 9$ corresponds to the one reported in Fig. 4. In Fig. 7 the $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ is reported as a function of the number of the arrays A. As expected, the $\mathrm{NMSE}^{(1)}_{\mathrm{GPD}}$ decreases as A increases since a greater number of arrays guarantees a wider angle coverage. Additionally, from the inspection of Fig. 7 we can notice that between $A = 6$ and $A = 10$ the difference is around $6\,\mathrm{dB}$ in response to an increase of 16 microphones.

The SDR of the VMs is reported in Fig. 8 for the same three cases adopted for GDP. The single source scenario is shown in Fig. 8(a). We can notice here that for the locations where the VM mostly picks up the diffuse component, the SDR estimation is less consistent. Hence, the performance is affected by the VM location and by the averaging of the diffuse estimates of the arrays (Sec. IV-B).

Fig. 9(a) reports the comparison between $\widehat{\mathrm{DRR}}$ and DRR. Note that $\widehat{\mathrm{DRR}}$ is comparable with the DRR profile of an ideal source with first-order cardioid directivity pattern. The ideal DRR reaches $-\infty$ in correspondence of the zero in the source cardioid directivity pattern since no direct component is propagated in this direction. The $\widehat{\mathrm{DRR}}$ does not follow this behavior due to the fact that the GPD is not exactly zero in this direction (see Fig. 6(a)). Moreover, we provide as a comparison the estimated $\widehat{\mathrm{DRR}}_{\mathrm{omni}}$ obtained assuming an omnidirectional sound source. More specifically, we estimated the exterior sound field coefficients (Sec. III-C) by setting the spherical harmonic order $L = 0$ in (33) and consequently, the synthesized direct signal in (38) presents an omnidirectional characteristic. Inspecting Fig. 9(a), we can notice that the addition of the source directivity greatly enhances the estimate of $\widehat{\mathrm{DRR}}$ giving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}}$ of $-14.3\,\mathrm{dB}$, while the $\widehat{\mathrm{DRR}}_{\mathrm{omni}}$ does not follow the actual trend of the reference giving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}_{\mathrm{omni}}}$ of $-1.14\,\mathrm{dB}$.

2) Two-sources scenario: We evaluate the virtual miking technique in a double talk scenario, where both acoustic sources of Fig. 4 are simultaneously active. The estimated position of the sources are $\hat{\boldsymbol{r}}'_1 = [1.8054, 2.0518]^T$ m and $\hat{\boldsymbol{r}}'_2 = [3.3556, 2.7087]^T$ m giving a localization error of 0.0759m and 0.1134m, respectively. We simulate $V = 180$ omnidirectional VMs, equally distributed around the sources with $V^{(1)} = V^{(2)} = 90$ VMs for each source placed accordingly to Fig. 4.

Both $\widehat{\mathrm{GDP}}$ and GDP are reported in Fig. 6(b) and Fig. 6(c). For both sources we can notice that $\widehat{\mathrm{GDP}}$ agrees with GDP suggesting that the spatial radiation characteristics of the sources are correctly reconstructed at the VMs.

Another important measure of the sound field spatial characteristics is the SIR. As shown in Fig. 10, the $\widehat{\mathrm{SIR}}$ follows the behaviour of the actual SIR for a wide range of directions. However, it is worth noting that the SIR goes to $-\infty$ for the directions in correspondence of the zeros in the directivity pattern of the sources. Since the behavior of $\widehat{\mathrm{SIR}}$ is related to the $\widehat{\mathrm{GPD}}$s of the two sources shown in Fig. 6(b) and Fig. 6(c), respectively it cannot reach $-\infty$. In fact, the $\widehat{\mathrm{GPD}}$ is not exactly zero for such directions. Similarly to what is done for the *single-source scenario* in Fig. 9(a), we include in Fig. 10 the $\widehat{\mathrm{SIR}}_{\mathrm{omni}}$ estimated when the two sources are incorrectly assumed as omnidirectional, i.e. setting $L = 0$ in (33). As expected, the estimation of the SIR effectively improves by modelling the source directivity explicitly. In particular, the $\mathrm{NMSE}^{(1)}_{\mathrm{SIR}_{\mathrm{omni}}}$ is equal to $5.3\,\mathrm{dB}$ for the first source (Fig. 10(a)) and the $\mathrm{NMSE}^{(2)}_{\mathrm{SIR}_{\mathrm{omni}}}$ is equal to $-2.4\,\mathrm{dB}$ for the second source (Fig. 10(b)). The $\mathrm{NMSE}^{(1)}_{\mathrm{SIR}}$ and $\mathrm{NMSE}^{(2)}_{\mathrm{SIR}}$ drops to $-12.4\,\mathrm{dB}$ and $-12.6\,\mathrm{dB}$, respectively when the source directivity is taken into account.

As far as the SDR is concerned, results are reported in Fig. 8(b) and Fig. 8(c). Similarly to the single source scenario, the VM performance is influenced by the approximation of the diffuse sound field, with lower SDR values when the diffuse component is mainly present and higher values when the VM is close to an array used for analyzing the sound scene.

Finally, the DRR of the VMs is provided in Fig. 9(b) and Fig.9(c). Analogously to the single source scenario, the $\widehat{\mathrm{DDR}}$ follows the reference value achieving a $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}}$ and a $\mathrm{NMSE}^{(2)}_{\mathrm{DRR}}$ of $-14.3\,\mathrm{dB}$ and $-14.7\,\mathrm{dB}$, respectively. As a comparison, also in Fig. 9(b)and Fig. 9(c) the $\widehat{\mathrm{DDR}}_{\mathrm{omni}}$, estimated assuming two omnidirectional sources, are reported. Both the $\mathrm{NMSE}^{(1)}_{\mathrm{DRR}_{\mathrm{omni}}}$ and the $\mathrm{NMSE}^{(2)}_{\mathrm{DRR}_{\mathrm{omni}}}$ are equal to $-1.7\,\mathrm{dB}$. Such values are approximately $13\,\mathrm{dB}$ higher than the the ones obtained by considering the directivity of the sources. Generally, we can notice that the behavior of the DRR

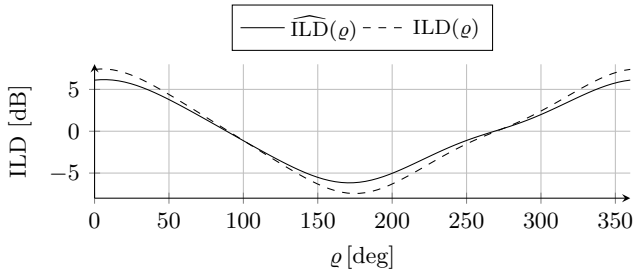


Fig. 11. The ILD of the stereo setup

is mainly determined by the directivity of the sound source. Low DRR values in Fig. 9(b) and Fig. 9(c) are associated to locations where the direct component energy is lower than the diffuse component due to the directivity pattern of each source.

3) Stereo recording scenario: The X-Y stereo microphone is commonly adopted for spatial sound acquisition. The spatial characteristics of the sound field are rendered in the stereo recording through the sound pressure level difference between the two directional microphones. Therefore, we adopted the ILD (51) as a metric to evaluate the ability of the VM in reproducing the spatial features of the stereo setup. The ILD is computed in the scenario of Fig. 5, as a function of the $\mathrm{VM}_{\mathrm{X-Y}}$ azimuth orientation ϱ varying from $0\,$deg to $360\,$deg. The omnidirectional source is localized in $\hat{\boldsymbol{r}}' = [2.5316, 2.9707]^T\,$m with a localization error of $0.0431\,$m. Inspecting the curves of Fig. 11, we can notice an high agreement between the $\widehat{\mathrm{ILD}}$ computed with the estimated signals and the ILD computed with the reference signals. The $0\,$dB value is crossed around $90\,$deg and $270\,$deg, when the sound pressure level at VM_{L} and VM_{R} is equivalent. This coincides with the source located at the center of the stereo plane. In an anechoic scenario, the maximum of the ILD function would occur around $315\,$deg, in correspondence of the zero in the VM_{R} cardioid pattern and the minimum around $225\,$deg, when the VM_{L} amplitude is attenuated. However in Fig. 11, we cannot observe this behavior due to the presence of the diffuse field. Indeed this does not let the signal power of the two microphones in the X-Y configuration go exactly to zero.

The proposed virtual miking procedure aims therefore to provide a promising tool for a wide variety of spatial sound applications. In particular, we showed that the spatial sound characteristics e.g. the DRR, the SIR or the ILD can be effectively approximated by the synthesized VM signal. Moreover, the explicit modeling of the sound source directional characteristics improves the VM estimation especially in terms of its spatial cues with respect to the omnidirectional source model commonly adopted in the literature. Thanks to the possibility of synthesizing the VMs signals in arbitrary locations, the proposed techniques potentially enables a listener to *virtually navigate* a recorded sound field with *six-degree of freedom*. Therefore, we envision the application of the procedure in Virtual or Augmented Reality framework, where capturing the sound field spatial features is a relevant aspect in order to provide an immersive user experience.

VI. Conclusions and Future Works

In this manuscript we proposed a parametric technique for virtual miking in reverberant environments from the analysis of signals captured by a set of distributed microphone arrays. The first step concerns the separation of the direct and diffuse components from the measured signals and the localization of the sources. The direct component is analyzed assuming a spherical harmonic representation of the emitted sound field that inherently describes the directional behaviour of the sources. The diffuse component is assumed to be isotropic and homogeneous. The synthesis of the VM signal is accomplished by properly mixing the estimated direct and diffuse components at the desired location. Hence, the overall system, is a combination of individual sub-systems. This structure brings us two main advantages; on the one hand we can think of distributing part of the computational load locally to each array (e.g., the estimation of the DOAs, the estimation of the direct and diffuse components); on the other hand such a structure allows us to eventually substitute sub-systems depending on the application scenario with the only constraint of maintaining the same input/output relationship. Furthermore, this structure gives us a great insight into the system behaviour and promotes the model interpretability.

Results show that the proposed technique is able to reconstruct the main cues of the VM signal, for instance, the ones related to reverberation (e.g. Direct to Reverberant Ratio) and spatial recording (Interchannel Level Difference). Moreover, by analyzing the metrics at different locations in the space, we showed that the proposed approach is able to capture the spatial characteristic of the recorded acoustic scene.

Future works will be devoted to the investigation on the possibility to jointly optimize the model parameters and to the generalization of the approach to a 3D scenario where microphone arrays and sources are not constrained to lie on the same plane. As far as this last point is concerned, such a generalization will require an extension of the localization approach to 3D geometries, and to keep into account the fact that, in a 3D scenario, all spherical harmonics should be taken into account (i.e., it is no longer possible to remove spherical harmonics for which $\ell + |\mu|$ is even). It follows that, more arrays will be needed in order to guarantee the wider angle coverage required for the estimation of the exterior sound field coefficients.

Appendix A
Demonstration of Equality between Estimated and Actual Power of Direct and Diffuse Components

In this appendix we will demonstrate that

$$\begin{aligned} E\{|\hat{X}_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} &= E\{|X_{n,\mathrm{dir}}(t,\omega,\boldsymbol{r}_i)|^2\} \\ E\{|\hat{X}_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\} &= E\{|X_{\mathrm{diff}}(t,\omega,\boldsymbol{r}_i)|^2\}, \end{aligned} \tag{53}$$

under the assumption that the oversubtraction factor $\nu = 1$, the gain floor $G_{\min} = 0$ and that the direct, diffuse and noise power at the ith and at the i'th microphone are equal.

In order to demonstrate the first equality, we make use of the filter definition given in (27) and the definition of the CDR in (19). It follows that

$$\begin{aligned} E\{|\hat{X}_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\} &= |G_{\mathrm{dir}}(\boldsymbol{r}_i)|^2 E\{|U(\boldsymbol{r}_i)|^2\} \\ &= \frac{\mathrm{CDR}(\boldsymbol{r}_i)}{\mathrm{CDR}(\boldsymbol{r}_i)+1} E\left\{\frac{Z(\boldsymbol{r}_i)+Z(\boldsymbol{r}_{i'})}{2}\right\} \\ &= \frac{\Phi_{\mathrm{dir},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \times \frac{1}{2}[2E\left\{|X_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\right\} + \\ &2E\left\{|X_{\mathrm{diff}}(\boldsymbol{r}_i)|^2\right\} + 2E\left\{|N(\boldsymbol{r}_i)|^2\right\} - 2E\left\{|N(\boldsymbol{r}_i)|^2\right\}] \\ &= \frac{\Phi_{\mathrm{dir},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \frac{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}}{1} = E\{|X_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\}, \end{aligned} \tag{54}$$

where the dependences on the time frame index t and the radial frequency ω have been omitted for the sake of readability.

In order to demonstrate the second equality instead, we make use of the filter definitions given in (30) and (27) and the definition of the CDR in (19). It follows that

$$\begin{aligned} E\{|\hat{X}_{\mathrm{diff}}(\boldsymbol{r}_i)|^2\} &= |G_{\mathrm{diff}}(\boldsymbol{r}_i)|^2 E\{|U(\boldsymbol{r}_i)|^2\} \\ &= \left(1 - \frac{\mathrm{CDR}(\boldsymbol{r}_i)}{\mathrm{CDR}(\boldsymbol{r}_i)+1}\right) E\left\{\frac{Z(\boldsymbol{r}_i)+Z(\boldsymbol{r}_{i'})}{2}\right\} \\ &= \frac{\Phi_{\mathrm{diff},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \times \frac{1}{2}[2E\left\{|X_{n,\mathrm{dir}}(\boldsymbol{r}_i)|^2\right\} + \\ &2E\left\{|X_{\mathrm{diff}}(\boldsymbol{r}_i)|^2\right\} + 2E\left\{|N(\boldsymbol{r}_i)|^2\right\} - 2E\left\{|N(\boldsymbol{r}_i)|^2\right\}] \\ &= \frac{\Phi_{\mathrm{diff},ii}}{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}} \frac{\Phi_{\mathrm{diff},ii}+\Phi_{\mathrm{dir},ii}}{1} = E\{|X_{\mathrm{diff}}(\boldsymbol{r}_i)|^2\}. \end{aligned} \tag{55}$$

REFERENCES

[1] P. Samarasinghe, T. D. Abhayapala, and M. A. Poletti, "3D spatial soundfield recording over large regions," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[2] P. Samarasinghe, T. D. Abhayapala, and M. A. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 647–658, Mar. 2014.

[3] J. G. Tylka and E. Y. Choueiri, "Soundfield navigation using an array of higher-order ambisonics microphones," in *Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016, p. 10.

[4] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 135–139, Jan. 2018.

[5] Y. Takida, S. Koyama, and H. Saruwataril, "Exterior and interior sound field separation using convex optimization: Comparison of signal models," in *26th European Signal Processing Conference,(EUSIPCO)*. IEEE, 2018, pp. 2549–2553.

[6] S. Koyama and L. Daudet, "Sparse representation of a spatial sound field in a reverberant environment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 172–184, Mar. 2019.

[7] F. Borra, I. D. Gebru, and D. Marković, "Soundfield reconstruction in reverberant environments using higher-order microphones and impulse response measurements," in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. IEEE, 2019, pp. 281–285.

[8] ——, "1st-order microphone array system for large area sound field recording and reconstruction: Discussion and preliminary results," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, p. 5.

[9] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[10] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional audio coding: Virtual microphone-based synthesis and subjective evaluation," *Journal of the Audio Engineering Society*, vol. 57, no. 9, pp. 709–724, 2009.

[11] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[12] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010, pp. 6–7.

[13] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*. Edinburgh, United Kingdom: IEEE, 2011, pp. 185–190.

[14] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

[15] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, Mar. 2015.

[16] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," in *Audio Engineering Society Conference: International Conference on Audio for Virtual and Augmented Reality*, 2018, p. 11.

[17] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro, "Estimation of the sound field at arbitrary positions in distributed microphone networks based on distributed ray space transform," in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*. IEEE, 2018, pp. 186–190.

[18] M. Pezzoli, F. Borra, F. Antonacci, A. Sarti, and S. Tubaro, "Reconstruction of the virtual microphone signal based on the distributed ray space transform," in *26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1537–1541.

[19] J. G. Tylka and E. Y. Choueiri, "Comparison of techniques for binaural navigation of higher-order ambisonic soundfields," in *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015, p. 13.

[20] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[21] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2008.

[22] 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 *Independent Component Analysis and Signal Separation*, T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 742–750.

[23] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.

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

[25] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 459–463.

[26] S. Pasha, J. Donley, and C. Ritz, "Blind speaker counting in highly reverberant environments by clustering coherence features," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1684–1687.

[27] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, and B. Firner, "Crowd++: Unsupervised speaker count with smartphones," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 43–52. [Online]. Available: https://doi.org/10.1145/2493432.2493435

[28] F. Stöter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2019.

[29] A. Nehorai and E. Paldi, "Acoustic vector sensor array processing," in *26th Asilomar Conference on Signals, Systems & Computers, (ACSSC)*. IEEE, 1992, pp. 192–198.

[30] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 53, no. 1, pp. 44–53, 2005.

[31] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

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

[33] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1132–1143, 1990.

[34] B. Ottersten, M. Viberg, and T. Kailath, "Performance analysis of the total least squares ESPRIT algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1122–1135, 1991.

[35] H. L. Van Trees, *Optimum array processing*. Wiley Online Library, 2002, vol. 1.

[36] R. Berkun, I. Cohen, and J. Benesty, "Combined beamformers for robust broadband regularized superdirective beamforming," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 5, pp. 877–886, 2015.

[37] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone arrays*. Springer, 2001, pp. 19–38.

[38] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "Soundfield imaging in the ray space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2493–2505, 2013.

[39] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro, "The ray space transform: A new framework for wave field processing," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5696–5706, Nov 2016.

[40] M. R. Azimi-Sadjadi, A. Pezeshki, L. L. Scharf, and M. E. Hohil, "Wideband doa estimation algorithms for multiple target detection and tracking using unattended acoustic sensors," in *Unattended/Unmanned Ground, Ocean, and Air Sensor Technologies and Applications VI*, vol. 5417. International Society for Optics and Photonics, 2004, pp. 1–11.

[41] F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, and A. Sarti, "A fast ray space transform for wave field processing using acoustic arrays," in *28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2020.

[42] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, "Multiview soundfield imaging in the projective ray space," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1054–1067, 2015.

[43] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.

[44] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[45] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[46] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, 2007.

[47] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, vol. 3. IEEE, 2004, pp. iii–889.

[48] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2337–2346, 2012.

[49] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[50] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[51] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, vol. 5. IEEE, 1988, pp. 2578–258.

[52] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[53] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[54] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Foundations and Trends, 2011.

[55] T. D. Abhayapala and A. Gupta, "Higher order differential-integral microphone arrays," *The Journal of the Acoustical Society of America*, vol. 127, pp. EL227–EL233, May 2010.

[56] P. Samarasinghe, H. Chen, A. Fahim, and T. D. Abhayapala, "Performance analysis of a planar microphone array for three dimensional soundfield analysis," in *Workshop on Applications of Signal Processing to Audio and Acoustics, (WASPAA)*. IEEE, 2017, pp. 249–253.

[57] B. Rafaely, "Spatial alignment of acoustic sources based on spherical harmonics radiation analysis," in *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*. IEEE, 2010, pp. 1–5.

[58] P. A. Martin, *Multiple scattering: interaction of time-harmonic waves with N obstacles*. Cambridge University Press, 2006, no. 107.

[59] H. Appel, *3.1 The 3j-symbol*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1968, pp. 8–12.

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

[61] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 2010.

[62] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[63] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[64] "Sound quality assessment material recording for subjective tests," https://tech.ebu.ch/publications/sqamcd, European Broadcasting Union, Tech. Rep., 2008. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[65] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[66] E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep, Tech. Rep. 2.4, 2006.

[67] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio*. Wiley Online Library, 2018.

[68] H. Kuttruff, *Room acoustics*. CRC Press, 2016.

[69] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1669, July 2006.

Mirco Pezzoli (S'20) received the M.S. degree (cum laude), in 2017, in computer engineering from the Politecnico di Milano, Italy, where he is currently pursuing the Ph.D. degree in information engineering with the Dipartimento di Elettronica, Informazione and Bioingegneria. His main research interests are space-time audio signal processing and musical acoustics.

Federico Borra (S'17) received the B.S. degree, in 2014, and the M.S. degree (cum laude), in 2016, in computer engineering from the Politecnico di Milano, Italy. In 2020 he received his Ph.D. degree in information engineering from the Politecnico di Milano, Italy, where he is currently a postdoctoral researcher. His main research interests concern space-time audio signal processing.

Fabio Antonacci (M'14) was born in Bari, Italy, on July 26, 1979. He received the Laurea degree in 2004 in telecommunication engineering and the Ph.D. degree in information engineering in 2008, both from the Politecnico di Milano, Milan, Italy. He is currently an Assistant Professor at the Politecnico di Milano. His research focuses on space-time processing of audio signals, for both speaker and microphone arrays (source localization, acoustic scene analysis, rendering of spatial sound) and on modeling of acoustic propagation. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee and of the EURASIP SAT on Audio, Speech and Music Signal Processing.

Stefano Tubaro (SM'01) was born in Novara, Italy, in 1957. He completed his studies in Electronic Engineering at the Politecnico di Milano, Milan, Italy, in 1982. He then joined the Dipartimento di Elettronica, Informazione e Bioingegneria of the Politecnico di Milano, first as a Researcher of the National Research Council, and then (in November 1991) as an Associate Professor. Since December 2004, he has been appointed as a Full Professor of telecommunication at the Politecnico di Milano. His current research interests include advanced algorithms for video and sound processing. He is the author of more than 180 scientific publications on international journals and congresses and the coauthor of more than 15 patents. In the past few years, he has focused his interest on the development of innovative techniques for image and video tampering detection and, in general, for the blind recovery of the "processing history" of multimedia objects. He coordinates the research activities of the Image and Sound Processing Group at the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano. He had the role of Project Coordinator of the European Project ORIGAMI (A new paradigm for high-quality mixing of real and virtual) and of the research project ICT-FET-OPEN REWIND (REVerse engineering of audio-VIsual coNtent Data). This last project was aimed at synergistically combining principles of signal processing, machine learning, and information theory to answer relevant questions on the past history of such objects. He is a member the IEEE Multimedia Signal Processing Technical Committee and of the IEEE SPS Image Video and Multidimensional Signal Technical Committee. He was in the organization committee of a number of international conferences including the IEEE MMSP 2004/2013, IEEE ICIP 2005, IEEE AVSS 2005/2009, IEEE ICDSC 2009, IEEE MMSP 2013, IEEE ICME 2015. From May 2012 to April 2015, he was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, and is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

Augusto Sarti (M'04–SM'13) received his Ph.D. Information Engineering from the University of Padova, Italy, in 1993, with a joint graduate program with University of California, Berkeley. In 1993, he joined the Faculty of the Politecnico di Milano, Italy, where he is currently a Full Professor. In 2013, he also joined the University of California, Davis. He coordinates the activities of the Musical Acoustics Laboratory and the Sound and Music Computing Laboratory of the Politecnico di Milano. He promoted/coordinated and/or contributed to numerous European projects in the area of multimedia signal processing. He has coauthored over 300 scientific publications on international journals and congresses and numerous patents in the multimedia signal processing area. His main research interests are in the area of audio and acoustic signal processing, with particular focus on sound analysis, synthesis, and processing; space-time audio processing; geometrical acoustics; music information extraction and music modeling. He served in the IEEE Technical Committee on Audio and Acoustics Signal Processing for two terms. He served as Associate Editor of IEEE/ACM Tr. On Audio Speech and Language Processing, and as Senior Area Editor of IEEE Signal Processing Letters, and in 2017 he received the "Outstanding Editorial Board Member Award" by the IEEE Signal Processing Society. He co-chaired the IEEE Intl. Conf. on Advanced Video and Signal based Surveillance (AVSS-05); he chairman the Digital Audio Effects conference (DAFx-09); and he co-chaired the IEEE Intl. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-19). He was in the organizing committees of numerous International Conferences, including IEEE ICASSP-14; the ACM/IEEE Intl. Conf. on Distributed Smart Cameras (ICDSC-09); the IEEE Intl. Workshop on Haptic Audio-Visual Environment and Game (HAVE-09); and the European Signal Processing Conference (EUSIPCO-2018). He is currently serving in the EURASIP board of directors.