A METHODOLOGY FOR SOUND SCENE MANIPULATION BASED ON THE RAY SPACE TRANSFORM

Francesco Picetti, Federico Borra, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano
via Ponzio 34/5 - 20133 Milano, Italy

ABSTRACT

In this paper we devise a methodology for analysing and subsequently manipulating a sound scene acquired by means of a uniform linear array of microphones. The array signal is transformed in the ray space, i.e. a domain where acoustic rays are points; here we extract the source position and orientation in space and its radiation pattern, while its signal is extracted by a near-field beamformer. These descriptors can be easily manipulated and provided to any parametric rendering system. Through simulations we have proven the capability of the proposed method to perform different manipulations.

***Index Terms*—** Audio space time processing, ray space transform, sound scene manipulation

1. INTRODUCTION

Sound field processing is a research topic that aims at extracting information about the acoustic characteristics of both the environment and the objects in a sound scene. The ability to sense spatial properties of sound enables humans to experience immersivity, i.e. the sense of presence in a scene. In the last decades this topic has raised interest in both research and industry communities, for its great variety of applications.

For these purposes, the parametric spatial sound processing paradigm is gaining the trend in the literature [1], [2]. The main idea is to represent an input audio scene in a way that is independent of any reproduction format, thus enabling optimal reproduction over any given playback system as well as flexible scene modification [3].

In this paper we aim at analysing and subsequently manipulating a sound scene, adopting such a parametric model. Specifically, the parameters we are interested in, are the source signal, its position and orientation in space and its radiation pattern, defined as the angular-frequency dependence of the amplitude of the emitted sound field [4].

The scene is acquired by means of a uniform linear array (ULA) of microphones and transformed in the ray space, which is a domain where acoustic rays are mapped onto points and sources onto lines. In this domain, the source position is estimated by a linear regression, while the radiation pattern and the orientation are extrapolated in the circular harmonics domain through an optimization problem; the source signal is extracted by a specifically tailored near-field beamformer.

The scene parameters can be modified and provided to any parametric rendering system, such as *Ambisonics* [5] and *Wave Field Synthesis* [6]. In the following, the sound scene is characterized by a single sound source lying on the same plane of the ULA in a free field scenario. The manipulation refers to the source translation and the rotation about its axis; the source signal and radiation pattern are the *identity* of any sound object, therefore their modification is not a proper scene manipulation.

In terms of novelty, we highlight the advantages of processing the sound scene in the ray space: in addition to exploiting the point-line duality for source localization with very fast and robust methods, in this paper the radiation pattern and the source orientation are estimated directly from the ray space coefficients of the array signal. Therefore, the near-field beamformer for signal extraction is designed upon the information processed in the ray space.

2. SIGNAL MODEL AND BACKGROUND

In this section we define the data model adopted in this manuscript and review the Ray Space Transform (RST) introduced in [7] that will be applied to the array signal.

Being $\mathbf{u}(t_\tau, \omega_k) \in \mathbb{C}^{L\times 1}$ the Short Time Fourier Transform (STFT) of the microphone array at the τ-th time frame and k-th frequency bin, we can state that [8]

$$\mathbf{u}(t_\tau, \omega_k) = \mathbf{g}(\mathbf{R}|\mathbf{r}'(t_\tau), \omega_k) \circ \mathbf{p}(t_\tau, \omega_k) S(t_\tau, \omega_k) + \mathbf{e}(t_\tau, \omega_k), \tag{1}$$

where $\mathbf{R} = [\mathbf{r}_0, \dots, \mathbf{r}_{L-1}]$ collects the microphone positions, $\mathbf{r}'(t_\tau)$ is the source position, $\circ$ denotes the Hadamard product, $\mathbf{p}(t_\tau, \omega_k)$ collects the radiation pattern weights, $S(t_\tau, \omega_k) \in \mathbb{C}$ is the STFT of the source signal, $\mathbf{e}(t_\tau, \omega_k)$ is a spatially white noise and finally $\mathbf{g}(\mathbf{R}|\mathbf{r}'(t_\tau), \omega_k) \in \mathbb{C}^{L\times 1}$ collects the free field acoustic transfer function values from the source to the array in the form of the Green propagator

$$[\mathbf{g}(\mathbf{R}|\mathbf{r}'(t_\tau), \omega_k)]_l = \frac{e^{-j\omega_k \|\mathbf{r}_l - \mathbf{r}'(t_\tau)\|/c}}{4\pi\|\mathbf{r}_l - \mathbf{r}'(t_\tau)\|}, \tag{2}$$

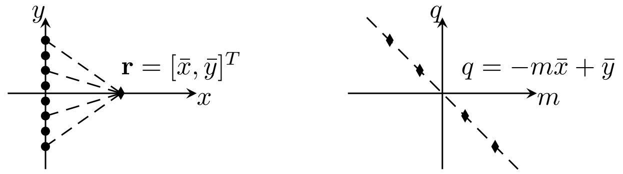
y
$\mathbf{r} = [\bar{x}, \bar{y}]^T$
x
q
$q = -m\bar{x} + \bar{y}$
m

Fig. 1: The point in the geometric space (left) transforms in a linear pattern in the ray space (right).

where $[\cdot]_l$ is the l-th element of the vector. The RST maps the signal acquired by the ULA onto a domain called *ray space*; each (m, q) point on this domain represents a ray lying on a line having slope m and intercept q [9]. Consequently, a generic point $\mathbf{r} = [\bar{x}, \bar{y}]^T$ in the geometric space appears in the ray space as the linear pattern $q = -m\bar{x} + \bar{y}$ [10], as shown in fig. 1. In detail, the RST applies a Gaussian window sliding along the array, modulated by complex exponential basis functions that span a set of prescribed directions. Being the ULA displaced along the y axis, it can be formalized as

$$[\mathbf{Y}]_{i,w}(\omega_k) = d\sum_{l=0}^{L-1} U_l(\omega_k) e^{-j\frac{\omega}{c}y_l\frac{m_w}{\sqrt{1+m_w^2}}} e^{-\pi\frac{(y_l-q_i)^2}{\sigma^2}}, \quad (3)$$

where $[\cdot]_{i,w}$ is the matrix element at i-th row and w-th column, d is the distance between two adjacent array elements, L is the number of array elements, $U_l(\omega)$ is the signal of the l-th microphone (placed at y_l), c is the speed of sound, q_i with $i = 0, \ldots, I-1$ is the centre of the i-th spatial window, m_w with $w = 0, \ldots, W-1$ encodes the direction of the ray passing through the window centre, and finally σ is related to the width of the Gaussian window.

3. SOUND SCENE ANALYSIS

In this section we go through the proposed analysis procedure, depicted in fig. 2, that enables the extraction of the sound scene parameters. Since the analysis is performed at each time frame independently, in the following the time dependency is omitted.

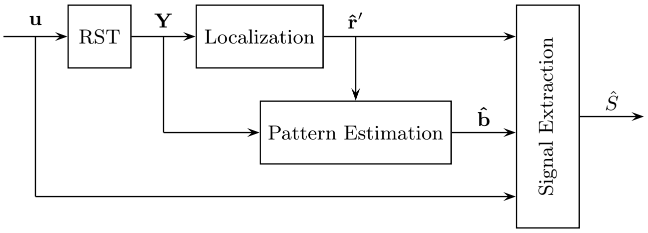

Fig. 2: Block diagram of the analysis stage.

3.1. Source Localization

The localization is performed upon considering the wideband extension $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I\times W}$ of the RST, computed as the geometric mean along the $K/2$ frequency bins of the absolute value of the coefficients matrix $|\mathbf{Y}(\omega_k)|$ [7].

From $\boldsymbol{\mathcal{Y}}$ one peak is identified for each i-th row at values $\hat{m}_i$ and amplitude λ_i. The found peaks are collected in the vector $\hat{\mathbf{m}} = [\hat{m}_0, \ldots, \hat{m}_{I-1}]^T$; then we introduce the matrices $\hat{\mathbf{M}} = [-\hat{\mathbf{m}}, \mathbf{1}]$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_0, \ldots, \lambda_{I-1})$; finally we build the vector $\mathbf{q} = [q_0, \ldots, q_{I-1}]^T$ that collects the y-coordinate of the centres of the spatial windows.

The position of the acoustic source is then estimated through a weighted least-squares regression [11]:

$$\hat{\mathbf{r}}' = [\hat{x}', \hat{y}']^T = \left(\hat{\mathbf{M}}^T\mathbf{\Lambda}\hat{\mathbf{M}}\right)^{-1}\hat{\mathbf{M}}^T\mathbf{\Lambda}\mathbf{q}. \quad (4)$$

3.2. Radiation Pattern Estimation

Once the position of the source has been extracted, we can estimate the radiation pattern from $\mathbf{Y}(\omega_k)$ as:

$$[\hat{\mathbf{p}}(\omega_k)]_i = 4\pi\rho_i\left(d\sum_{l=0}^{L-1} e^{-\pi(y_l-q_i)^2/\sigma^2}\right)^{-1} [|\mathbf{Y}(\omega_k)|]_{i,w'}, \quad (5)$$

where $\rho_i = \sqrt{\hat{x}' + (q_i - \hat{y}')^2}$ accounts for the attenuation due to the acoustic propagation, the term in brackets accounts for the Gaussian windowing introduced by the RST and $[\mathbf{Y}(\omega_k)]_{i,w'}$ are the RST coefficients along the linear pattern emerging from the localization, i.e. (i, w') such that $q_i = -m_{w'}\hat{x}' + \hat{y}'$.

The radiation pattern $\hat{\mathbf{p}}(\omega_k)$ is *sensed* by the array in a limited set of directions $\theta_i = \arctan((\hat{y}' - q_i)/\hat{x}')$ from the source position to the centres of the spatial windows; however we would like to know the radiation pattern value for all the possible directions. Therefore, we assume that the source radiation pattern can be represented with a limited set of Circular Harmonics (CH) coefficients, similarly to [12]. The decomposition assumes a set of cosine basis functions:

$$[\hat{\mathbf{p}}(\omega_k)]_i = \sum_{n=0}^{N-1} b_n(\omega_k)\cos\left(n\left(\theta_i - \phi\right)\right), \quad (6)$$

where N is the order of the CH decomposition and $b_n(\omega_k)$ is the n-th order coefficient. We define a coefficients vector $\mathbf{b}(\omega_k) \in \mathbb{R}^{N\times 1}$ and a matrix $\hat{\mathbf{A}} \in \mathbb{R}^{I\times N}$ collecting the cosine functions:

$$\left[\hat{\mathbf{A}}\right]_{i,n} = \cos(n(\theta_i - \hat{\phi})), \quad (7)$$

where $\hat{\phi}$ is the estimated source orientation, defined as the direction at which the mean along the frequency of the radiation pattern is maximum:

$$\hat{\phi} = \theta_{i'} \quad \text{with} \quad i' = \arg\max_i \frac{2}{K}\sum_k [\hat{\mathbf{p}}(\omega_k)]_i, \quad (8)$$

having assumed the source is facing the array. Thus, equation (6) can be written in matrix form:

$$\hat{\mathbf{p}}(\omega_k) = \hat{\mathbf{A}}\mathbf{b}(\omega_k). \tag{9}$$

Therefore, $\mathbf{b}(\omega_k)$ can be estimated using a linearly constrained minimization problem [13]:

$$\begin{aligned} \hat{\mathbf{b}}(\omega_k) = \underset{\mathbf{b}}{\arg\min} \quad & \|\hat{\mathbf{p}}(\omega_k) - \hat{\mathbf{A}}\mathbf{b}\|^2 \\ \text{s.t.} \quad & \mathbf{Cb} \geq \mathbf{0} \end{aligned} \tag{10}$$

where $\mathbf{0} \in \mathbb{Z}^{N_\theta \times 1}$ is a vector of zeros and $\mathbf{C} \in \mathbb{R}^{N_\theta \times N}$ collects the CH basis functions in the same form of (7) but $\forall\theta \in [0, 2\pi]$, having N_θ uniformly sampled directions. Here, the constraint forces the radiation pattern to be positive.

3.3. Signal Extraction

Once $\hat{\mathbf{b}}(\omega_k)$ has been obtained, we can compute the radiation pattern weight for the array, $\tilde{\mathbf{p}}(\omega_k)$ adopting (9), and the propagation terms $\tilde{\mathbf{g}}(\omega_k) = \mathbf{g}(\mathbf{R}|\hat{\mathbf{r}}', \omega_k)$ from the estimated source position $\hat{\mathbf{r}}'$ to the microphone positions $\mathbf{R}$.

We aim at extracting the source signal $S(\omega_k)$ directly from the array signal (1) by designing a filter $\tilde{\mathbf{h}}(\omega_k)$ such that

$$\begin{aligned} \tilde{\mathbf{h}}(\omega_k) = \underset{\mathbf{h}}{\arg\min} \quad & \mathbf{h}^H\mathbf{h} \\ \text{s.t.} \quad & \mathbf{h}^H\left(\tilde{\mathbf{p}}(\omega_k) \circ \tilde{\mathbf{g}}(\omega_k)\right) = 1, \end{aligned} \tag{11}$$

yielding to the *near-field* beamformer:

$$\tilde{\mathbf{h}}(\omega_k) = \frac{\tilde{\mathbf{p}}(\omega_k) \circ \tilde{\mathbf{g}}(\omega_k)}{\|\tilde{\mathbf{p}}(\omega_k) \circ \tilde{\mathbf{g}}(\omega_k)\|} \tag{12}$$

The source signal is extracted, at each time frame, as $\hat{S}(\omega_k) = \tilde{\mathbf{h}}(\omega_k)^H\mathbf{u}(\omega_k)$ and, once all the time frames and frequency bins have been analysed, the original signal can be reconstructed with the Inverse STFT.

4. SOUND SCENE SYNTHESIS

This section describes the spatial sound scene manipulation and synthesis. In the following we consider a set of L_S points, collected by the matrix $\mathbf{R}_S$, at which we reconstruct the sound field after its manipulation. The time and frequency dependencies are explicitly reported for the sake of clarity; however the synthesis is performed in the same STFT framework of the analysis stage.

Let $\mathbf{r}'_S(t_\tau)$ be the modified source position; the acoustic propagator is $\mathbf{g}_S(t_\tau, \omega_k) = \mathbf{g}(\mathbf{R}_S|\mathbf{r}'_S(t_\tau), \omega_k)$, computed as in (2). The manipulated radiation pattern is computed in the CH domain as in (9):

$$\begin{aligned} [\mathbf{A}_S(t_\tau)]_{l,n} = & \quad \cos(n(\theta_l(t_\tau) - \phi_S(t_\tau)) \\ \mathbf{p}_S(t_\tau, \omega_k) = & \quad \mathbf{A}_S(t_\tau)\hat{\mathbf{b}}(\omega_k), \end{aligned} \tag{13}$$

where θ_l is the direction from the desired source position $\mathbf{r}'_S(t_\tau)$ to the l-th reconstruction point, and $\phi_S(t_\tau)$ is the desired source orientation. Finally we can compute the manipulated sound field at the evaluation points as in (1):

$$\mathbf{u}_S(t_\tau, \omega_k) = \left(\mathbf{g}_S(t_\tau, \omega_k) \circ \mathbf{p}_S(t_\tau, \omega_k)\right) \hat{S}(t_\tau, \omega_k). \tag{14}$$

We remark that the CH coefficients $\hat{\mathbf{b}}(\omega_k)$ and the source signal $\hat{S}(t_\tau, \omega_k)$ are inherited from the analysis, while the manipulation concerns only the source position and orientation.

5. RESULTS

In this section we evaluate through simulations the proposed methodology. The scene consists of a single sound source emitting a white noise signal with a 4-th order directive pattern facing the array. The radiation pattern extraction is performed by a first order CHD. Without loss of generality, the synthesized sound field is evaluated at the same microphone positions of the analysis stage, i.e. $\mathbf{R}_S = \mathbf{R}$. Table 1 reports the scene characteristics and the RST and STFT parameters.

In order to validate the synthesized sound field against the sound field as if it would be generated by a virtual source, we compare the l-th microphone signal $u_l(t)$ of the desired sound field with its synthesized version $\hat{u}_l(t)$ obtained from (14). In particular, we compute the Normalized Mean Square Error between the actual and synthesized signals averaged over all the sensors in the array:

$$\text{NMSE} = 10\log_{10}\left(\frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_t |u_l(t) - \hat{u}_l(t)|^2}{\sum_t |u_l(t)|^2}\right). \tag{15}$$

We perform the simulations by varying the variance of the spatial white noise $\mathbf{e}(t_\tau, \omega_k)$ (1) in order to obtain three different values of SNR with respect to the 8-th microphone, e.g.

Sound Scene	
speed of sound	$c = 343\ \mathrm{m\,s^{-1}}$
source orientation	$\phi = \pi$
radiation pattern (9)	$\mathbf{b} = [0.62, 0.3, 0.02, 0.05, 0.01]^T$
array elements and distance	$L = 16$, $d = 10$ cm
microphone radiation pattern	omnidirectional
RST	
q axis sampling interval	$\bar{q} = d/2 = 5$ cm
number of spatial windows	$I = 32$
m axis sampling interval	$\bar{m} \approx 0.0028$
number of visible directions	$W = 350$
angular aperture	$m = \pm 4 \quad (\theta \approx \pm 76°)$
Gaussian windows width	$\sigma = 30$ cm
STFT	
frequency range	$250 \div 1700$ Hz
length, overlap and shape	50 ms - 75% - Hamming
FFT sampling	8 kHz, 512 bins

Table 1: Simulation parameters.

20, 30 and 40 dB. The reported results are obtained by averaging over 100 realizations.

5.1. Scene Manipulation

For the sake of clarity we do manipulate only one parameter at a time: first we modify the source orientation, then we translate the source from its analysed position. Fig. 3 helps visualizing the manipulations. We adopt the polar coordinates in order to clarify the impact on the synthesis of the distance ρ and the angle η of the source with respect to the centre of the array, i.e. the axes origin. During the analysis stage the source is placed at $\mathbf{r}' = [1,0]^T\mathrm{m}$ for the first three cases, while for the fourth one is placed in $[3,0]^T\mathrm{m}$. In the following we report the list of the performed manipulations.

Orientation: the source is rotated about its axis while its position is left unchanged in $[1,0]^T\mathrm{m}$. Fig. 4a shows the results against $\Delta\phi = \phi_S - \hat{\phi}$. It is possible to notice a smooth and graceful increase of the NMSE as the difference of the orientation increases from the original one, due to the fact that during the manipulation, for large values of $\Delta\phi$, part of the radiation pattern is synthesized from the model (9) for directions different from the analysis ones.

Angular coordinate: the source moves on a circumference centred on the axes origin with radius $1\mathrm{m}$ starting at the analysis position $[1,0]^T\mathrm{m}$, while keeping its original orientation fixed at $\phi_S = \pi$. The results are reported in fig. 4b; as the source departs from the analysis position, the synthesized directions deviate from the analysis set and therefore the NMSE slightly increases.

Increasing distance: the source moves on a straight line starting at position $[1,0]^T\mathrm{m}$ up to $[3,0]^T\mathrm{m}$ i.e. increasing the distance ρ (in the figure, reported as $\overrightarrow{\rho}$). The results in fig. 4c shows a decrease of the NMSE due to the fact that, as the source gets away without changing its orientation, the set of directions *sensed* by the array shrinks and therefore the radiation pattern weights are not extrapolated but instead interpolated from the analysis data.

Decreasing distance: the source moves on a straight line starting at position $[3,0]^T\mathrm{m}$ down to $[1,0]^T\mathrm{m}$ i.e. decreasing ρ (in the figure, reported as $\overleftarrow{\rho}$). The results in fig. 4d shows an increase of the NMSE due to the fact that as the source moves closer to the array without changing its orientation, the set of directions *sensed* by the array widens.

Although the original pattern has an higher order than the synthesized one, the system achieves good performances in terms of NMSE, which remains strictly negative even for low SNR values. As expected, the higher the SNR is, the lower NMSE can be obtained.

6. CONCLUSIONS

In this paper we have proposed a methodology for the manipulation of sound scenes adopting a parametric approach in the ray space. The source localization and the extraction of its radiation pattern are performed through a least squares regression and a constrained quadratic programming problem, respectively. The source signal is extracted by a near-field beamformer. The estimate parameters are intuitively modified in order to manipulate the acoustic scene and then provided to any parametric rendering system. We have provided numerical evaluations to prove the effectiveness of the proposed solution. The system exhibits robustness to the additive white noise at different SNR values with low performance degradation. The proposed methodology can be easily extended by deploying multiple microphone arrays [9], [14].

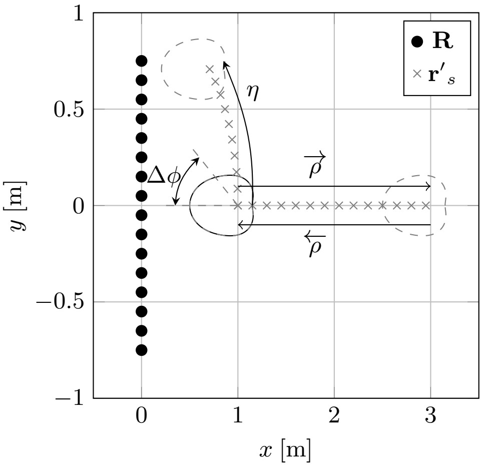

Fig. 3: The simulated manipulations.

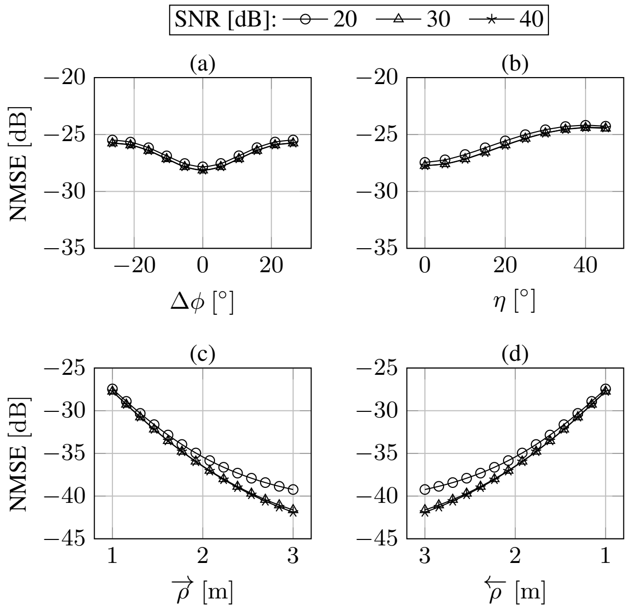

Fig. 4: Simulation results for three different SNR levels when manipulating: (a) the orientation; (b) the angular coordinate; (c) increasing distance; (d) decreasing distance.

7. REFERENCES

[1] Konrad Kowalczyk, Oliver Thiergart, M Taseska, G Del Galdo, Ville Pulkki, L. Cristoferetti, and Emanuël A. P. Habets, "Parametric Spatial Sound Processing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, 2015.

[2] James D. Johnston, Jean-Marc Jot, Zoran Fejzo, and Steve R. Hastings, "Beyond Coding: Reproduction of Direct and Diffuse Sounds in Multiple Environments," in *AES 129th Convention*, San Francisco, CA, USA, 2010, Audio Engineering Society.

[3] Michael M. Goodwin and Jean-Marc Jot, "Spatial Audio Scene Coding," in *AES 123th Convention*, San Francisco, CA, USA, 2008, Audio Engineering Society.

[4] Earl G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, Academic Press, London, UK, 1999.

[5] Peter Fellgett, "Ambisonics. part one: General system description," *Studio Sound*, vol. 17, no. 8, pp. 20–22, 1975.

[6] A. J. Berkhout, D. De Vries, and P. Vogel, "Acoustic Control by Wave Field Synthesis," *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[7] Lucio Bianchi, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "The ray space transform: A new framework for wave field processing," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5696–5706, 2016.

[8] Antonio Canclini, Luca Mucci, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Estimation of the radiation pattern of a violin during the performance using plenacoustic methods," in *AES 138th Convention*, Warsaw, PL, 2015, pp. 1–10, Audio Engineering Society.

[9] Dejan Marković, Fabio Antonacci, Augusto Sarti, and Stefano 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.

[10] Dejan Marković, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Soundfield imaging in the ray space," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 12, pp. 2493–2505, 2013.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer, 2nd edition, 2017.

[12] Filippo Maria Fazi, Vincent Brunel, Philip A. Nelson, Lars Hörchens, and Jeongil Seo, "Measurement and Fourier-Bessel Analysis of Loudspeakers Radiation Patterns Using a Spherical Array of Microphones.," in *AES 124th Convention*, Amsterdam, NL, 2008, Audio Engineering Society.

[13] Stephen P. Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[14] Federico Borra, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Extraction of acoustic sources for multiple arrays based on the ray space transform," in *2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, 2017, pp. 146–150, IEEE.