KU LEUVEN

STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Authors: Tengan E., Dietzen T., Elvander F., van Waterschoot T. Title: Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization Journal/Conference/Book: IEEE/ACM Transactions on Audio, Speech and Language Processing
Archived version	☐ Submitted version ■ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ☐ EU H22020 ■ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other
Abstract	In this paper, a method is proposed for estimating the direction of arrival (DOA) of multiple broadband sound sources. This is achieved by solving a group-sparse optimization problem, which models an observed broadband steered response power (SRP) map as a linear function of power spectral densities (PSDs), associated to a set of candidate DOAs. The estimation of the source DOAs is then accomplished by identifying peaks in the resulting spatial power density, i.e., the estimated direction-specific PSDs integrated over frequency. The proposed method is motivated by its potential to reveal more distinct peaks in the estimated spatial power density than those directly observed in the SRP map, which can be beneficial to the robustness in DOA estimation performance when multiple sources need to be distinguished under varying acoustic conditions. An implementation of the proposed method using the alternating direction method of multipliers (ADMM) is presented, and the DOA estimation performance is evaluated with both simulated and experimental data. Results show that, especially in reverberant scenarios, the proposed method presents an advantage in locating closely spaced sources when compared to SRP-PHAT, the group-sparse iterative covariance-based estimation (GSPICE) method, and the wideband MUSIC method with geometric averaging. Furthermore, it is observed that for a compact microphone array, the proposed method overall maintained its performance even when using SRP maps with lower grid resolutions than the sampling requirements of the broadband SRP function. Finally, results obtained with experimental data showed the applicability of the proposed method in a practical meeting room environment.
Author contact	elisa.tengan@esat.kuleuven.be

Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization

Elisa Tengan, Thomas Dietzen, Filip Elvander, and Toon van Waterschoot

***Abstract*—In this paper, a method is proposed for estimating the direction of arrival (DOA) of multiple broadband sound sources. This is achieved through the solution of a group-sparse optimization problem, which models an observed broadband steered response power (SRP) map as a linear function of power spectral densities (PSDs), corresponding to a set of candidate DOAs, and forming a PSD vector. Given the assumption of spatial sparsity, the estimation of the source DOAs is then accomplished by identifying peaks in the resulting spatial power density, i.e., the estimated direction-specific PSDs integrated over frequency. The motivation behind the proposed method lies in its potential to reveal more distinct peaks in the estimated spatial power density than those directly observed in the broadband SRP map, which can be beneficial to the robustness in DOA estimation performance when multiple sources need to be distinguished under varying acoustic conditions. An implementation of the proposed method using the alternating direction method of multipliers (ADMM) is presented, and the DOA estimation performance is evaluated with both simulated and experimental data. Results show that, especially in reverberant scenarios, the proposed method presents an advantage in locating closely spaced sources when compared to the conventional SRP-PHAT, the group-sparse iterative covariance-based estimation (GSPICE) method, and the wideband MUSIC method with geometric averaging. Furthermore, it is observed that for a compact microphone array, the proposed method overall maintained its performance even when using SRP maps computed with grid resolutions that are lower than the sampling requirements of the broadband SRP function. Finally, results obtained with experimental data showed the validity and applicability of the proposed method in a practical meeting room environment.**

***Index Terms*—source localization, direction-of-arrival estimation, group sparsity, steered response power**

I. Introduction

The field of sound source localization research has remained prevalent due to its wide-ranging applications, including domains as diverse as hearing aid technology [1], [2], multi-party videoconferencing [3], acoustic surveillance [4] and autonomous robots [5]. Over time, significant research efforts have been devoted to addressing the challenges of maintaining a robust source localization performance in realistic scenarios.

Many source localization methods rely on the use of microphone arrays, primarily due to the spatial diversity they offer and that can be effectively exploited [6]. Some well-established categories of array processing methods for source localization include beamforming methods [7], [8], subspace-based approaches [9]–[11], as well as generalized cross-correlation (GCC) methods [12]–[14]. Additionally, machine learning [15], [16] and compressed sensing techniques [17]–[20], as well as methods based on optimal transport [21]–[23], have also emerged as relevant alternatives.

Among the conventional GCC-based methods, the steered response power (SRP) has been thoroughly used as the foundation for new localization methods being further developed, mostly due to its robustness in adverse acoustic conditions such as noisy and reverberant environments when compared to, for instance, subspace-based approaches [13], [14], [24]–[27]. Moreover, the use of the phase transform (PHAT) weighting on the GCC function yields the well-established SRP-PHAT framework, which has been demonstrated to improve performance on the localization of speech sources in reverberant environments [13], [14], [24], [28], [29].

Since the conventional SRP and SRP-PHAT methods rely on finding the candidate direction that yields the maximum value over a spatial acoustic map [24], when multiple sources are present, differentiating true source-related peaks from spurious peaks caused by noise or acoustic reflections, as well as resolving closely located sources, can become a difficult task when employing these methods [29]–[31]. As a consequence, various strategies have been proposed to improve multi-source localization under an SRP-based framework [31]–[44]. Some techniques propose the recursive removal of the influence of previously located sound sources in the observed SRP maps, thereby enabling the localization of the remaining sources [32]–[35]. Spatial clustering [36], [37] and grid refinement methods [38], [39] represent other relevant approaches in this context. Exploiting sparsity in the time-frequency domain [31], [40], [41] and in the acoustic scene [42]–[44] has also been investigated.

In this paper, as an alternative contribution to the task of multi-source localization under an SRP-based framework, we propose a method for estimating the direction of arrival (DOA) of multiple broadband sound sources in far-field scenarios

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of FWO Mandate SB 1S86520N and KU Leuven internal funds C14/21/075 "A holistic approach to the design of integrated and distributed digital signal processing algorithms for audio and speech communication devices". The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

E. Tengan, T. Dietzen and T. van Waterschoot are with the Department of Electrical Engineering (ESAT/STADIUS), KU Leuven, Leuven, Belgium (email: elisa.tengan@esat.kuleuven.be; thomas.dietzen@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be). F. Elvander is with the Department of Information and Communications Engineering, Aalto University, Espoo, Finland (email: filip.elvander@aalto.fi).

based on group-sparse modeling. Group-sparse regularization, previously applied in the context of broadband source localization with co-prime arrays [45] and single directional microphones [46], [47], is here employed to model a broadband SRP map. In particular, the cross-correlations used for computing the observed SRP map are modeled as a linear function of (unknown) source power spectral densities (PSDs), supported on an overcomplete search grid of candidate DOAs. Given the assumption of spatial sparsity, only a small number of candidate DOAs will have non-zero PSDs, yielding a group-sparse structure to the vector constructed by concatenating the DOA-specific PSDs from the entire search grid. We propose to estimate the full set of PSDs by solving a linear inverse problem augmented by a group-sparsity penalty. The source DOAs are then identified by the peaks of the spatial power density, i.e., the estimated direction-specific PSDs integrated over frequency. The motivation behind the proposed method lies in its potential to reveal more distinct peaks in the resulting spatial power density than those directly observed in the broadband SRP map [17], [42], [43], which can be beneficial to the robustness of the DOA estimation performance when multiple sources need to be distinguished under varying acoustic conditions.

The proposed method is based on our previous work [48], in which the search grid of candidate DOAs used to compose the PSD vector to be estimated was assumed to coincide with the search grid of the observed SRP map that is being modeled, and the estimation of the PSD vector was performed by solving a Group Lasso problem [49]. In this paper, the extension of the method from [48] is twofold: Firstly, we reformulate the optimization problem to accommodate scenarios where the search grid used for computing the SRP map can have a lower resolution than the one used in the estimation of the source PSD vector. Within this context, the uniform spatial sampling requirements of the broadband SRP function, for a given microphone array configuration, are also derived in this work. We note that an analytical expression of the sampling requirements of the GCCs used in the computation of the SRP map has been previously presented in [50], in terms of time difference of arrival (TDOA), and therein used to determine the pair-specific maximum signal bandwidth for which the sampling of the GCC function can be performed without aliasing, given a chosen grid resolution. Moreover, different approaches aimed at efficient spatial sampling of the SRP have also been proposed, based on the spatial distribution of TDOAs [51]–[53]. Alternatively, our focus lies in establishing, through a different perspective, a reference for the required uniform spatial grid resolution, and consequently, the number of required broadband SRP samples, to ensure the highest feasible spatial resolution in the observed SRP map. Secondly, the method from our previous work [48] is further extended by using a group square-root Lasso [54] formulation for the optimization problem in the estimation of the PSD vector, as opposed to the more conventional group Lasso formulation [49] previously employed. As shown in [55], the group square-root Lasso is equivalent to a group-sparse covariance fitting problem, and its formulation can facilitate the required regularization tuning based on group-sparse covariance fitting criteria [54], [55]. This adaptation is particularly valuable in practical applications where the acoustic noise conditions are not known a priori. An efficient implementation of the proposed framework using the alternating direction method of multipliers (ADMM) [56] is also presented.

In order to evaluate the proposed method's DOA estimation performance, a series of simulations subject to different acoustic conditions are performed. We first consider a two-speaker scenario and subsequently vary the levels of noise and reverberation, as well as the angular separation and broadband power level difference between sources. We then consider scenarios with up to five sound sources, involving speech and instrumental music. The effects of the grid resolution of the observed SRP map, as well as of limiting the number of iterations within the proposed method's ADMM scheme, are analyzed. Furthermore, the method's performance is evaluated with experimental data [57]. This study greatly extends the aspects considered for evaluating the proposed method when compared with those presented in our previous work [48]. Moreover, the proposed method is compared with the conventional SRP-PHAT method, as well as with the group-sparse iterative covariance-based estimation (GSPICE) method [55], here used for estimating the PSDs directly from the microphone array cross-correlations, and with the wideband MUSIC method with geometric averaging [9], [58]. Results show that, especially in reverberant scenarios, the proposed method presents an advantage in locating closely spaced sources when compared to the SRP-PHAT, GSPICE, and wideband MUSIC. Furthermore, it is shown that for a compact microphone array, the proposed method overall maintains its performance even when using SRP maps computed with grid resolutions that are lower than the sampling requirements of the broadband SRP function. This observation, along with results obtained for limited numbers of iterations within the ADMM-scheme, indicate potential approaches to reduce the proposed method's overall computational complexity while maintaining its DOA estimation performance. Finally, results obtained with experimental data show the validity and applicability of the proposed method in a practical meeting room environment.

This paper is organized as follows. In Sec. II, the signal model is defined. In Sec. III, the broadband SRP formulation, as well as its spatial sampling requirements, are detailed. In Sec. IV, the proposed method is presented. In Sec. V, the simulation setups and their implementation are detailed, with results discussed in Sec. VI. Finally, Sec. VII concludes the paper.

Notation: vectors and matrices are denoted by lower- and upper-case boldface letters, respectively. For matrices $\mathbf{A}$ (and vectors $\mathbf{a}$ where applicable), $\mathbf{A}^\top$, $\mathbf{A}^H$, $\Re\{\mathbf{A}\}$, and $[\mathbf{A}]_{m,n}$ respectively denote the transpose, the Hermitian transpose, the real part, and the element of $\mathbf{A}$ at index (m, n). For vectors $\mathbf{a}$, $\|\mathbf{a}\|_p$ denotes the l_p-norm, the operation $\mathrm{diag}(\mathbf{a})$ creates a diagonal matrix from the elements of $\mathbf{a}$, and we use $\hat{\mathbf{a}}$ to denote an estimate of $\mathbf{a}$. For scalars a, the expectation operator is denoted by $\mathbb{E}\{a\}$, the complex conjugate by a^* and an estimate by $\hat{a}$. The operation $\mathrm{mod}(a, b)$ denotes a modulo b, and $\lceil a \rceil$ denotes the ceiling of a. Euler's number is denoted by e and

the imaginary unit by j.

II. Signal model

We here consider scenarios where sound sources are located in the far field and in the same horizontal plane as a planar microphone array with M elements, indexed by $m = 1, \ldots, M$. The time difference of arrival (TDOA) of a propagating sound wave arriving from an azimuth angle γ that is observed at a microphone pair (m, m') is given by [14]

$$\Delta t_{m,m'}(\gamma) = (\mathbf{p}_m - \mathbf{p}_{m'})^\top \mathbf{r}_\gamma / c. \tag{1}$$

Here, $\mathbf{p}_m \in \mathbb{R}^3$ and $\mathbf{p}_{m'} \in \mathbb{R}^3$ are the Cartesian coordinate vectors of microphones m and m', respectively, i.e., $\mathbf{p}_m = [x_m, y_m, z_m]^\top$ and $\mathbf{p}_{m'} = [x_{m'}, y_{m'}, z_{m'}]^\top$. Furthermore, $\mathbf{r}_\gamma$ denotes the incident direction vector of unit length for azimuth angle γ, and c is the speed of sound. The azimuth angle of a sound source will also be referred to as its DOA.

In the discrete Fourier transform (DFT) domain, the m-th microphone signal is denoted by $y_m(\omega_k)$, with ω_k denoting the discrete angular frequency. Assuming that the microphone signals are sampled in the time domain with a sampling rate f_s, and that the signal $y_m(\omega_k)$ is computed from a fast Fourier transform (FFT) of length $2K$, the discrete frequencies correspond to $\omega_k = k\pi f_s / K$, with $k \in \{-K+1, \ldots, K\}$. Moreover, while considering a total of J sound sources located at DOAs $\vartheta_1, \ldots, \vartheta_J$, the signal $y_m(\omega_k)$ is modeled as

$$y_m(\omega_k) = \mathbf{w}_m^H(\omega_k)\mathbf{s}(\omega_k) + n_m(\omega_k), \tag{2}$$

where $n_m(\omega_k)$ denotes the undesired component, which can be due to noise and reverberation. The vector $\mathbf{s}(\omega_k) \in \mathbb{C}^J$ corresponds to the source signal vector at the array's first microphone, here used as reference:

$$\mathbf{s}(\omega_k) = [s_1(\omega_k),\ \ldots\ , s_J(\omega_k)]^\top, \tag{3}$$

with $s_i(\omega_k)$ denoting the signal emitted by the i-th source. The vector $\mathbf{w}_m(\omega_k) \in \mathbb{C}^J$ is thus the transfer function relative to the array's first microphone:

$$\mathbf{w}_m(\omega_k) = \left[e^{j\omega_k \Delta t_{m,1}(\vartheta_1)},\ \ldots\ , e^{j\omega_k \Delta t_{m,1}(\vartheta_J)}\right]^\top, \tag{4}$$

such that, based on (3), $\mathbf{w}_1(\omega_k)$ corresponds to $\mathbf{w}_1(\omega_k) = [1, \ldots, 1]^\top$.

The generalized cross-correlation between recorded signals from a microphone pair (m, m') in the DFT domain is defined as

$$\psi_{m,m'}(\omega_k) = \mathbb{E}\{y_m(\omega_k) y^*_{m'}(\omega_k)\}. \tag{5}$$

If the source signals are assumed to be mutually uncorrelated, then the cross-correlation between $s_i(\omega_k)$ and $s_{i'}(\omega_k)$, for $i, i' \in \{1, \ldots, J\}$, is given by

$$\mathbb{E}\{s_i(\omega_k) s^*_{i'}(\omega_k)\} = \begin{cases} \phi_{S_i}(\omega_k) & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases} \tag{6}$$

where $\phi_{S_i}(\omega_k)$ is the power spectral density (PSD) of the i-th source signal. Furthermore, from (1), one can note that $\Delta t_{m,m'}(\gamma) = \Delta t_{m,1}(\gamma) - \Delta t_{m',1}(\gamma)$. Thus, the cross-correlation $\psi_{m,m'}(\omega_k)$ can be expressed as

$$\psi_{m,m'}(\omega_k) = \sum_{i=1}^{J} \phi_{S_i}(\omega_k) e^{-j\omega_k \Delta t_{m,m'}(\vartheta_i)} + \eta^2_{m,m'}(\omega_k), \tag{7}$$

where $\eta^2_{m,m'}(\omega_k)$ denotes the cross-correlation of the undesired component[1] for microphone pair (m, m').

III. Broadband steered response power (SRP) mapping

A. Broadband SRP formulation

For conciseness, a microphone pair (m, m') will be denoted by index p in the remainder, such that a given TDOA $\Delta t_{m,m'}(\gamma)$ will be denoted by $\Delta t_p(\gamma)$. For an M-element microphone array, the total number of different microphone pairs P is given by

$$P = M(M-1)/2. \tag{8}$$

Since $\psi_p(\omega_k)$ is Hermitian, we consider its single-sided form, with $k \in \{0, \ldots, K\}$. Assuming the DC component to be negligible ($\psi_p(\omega_0) = 0$), the broadband SRP for a given candidate DOA angle γ is defined as [59]

$$z(\gamma) = 2 \sum_{p=1}^{P} \sum_{k=1}^{K} \Re\{\psi_p(\omega_k) e^{j\omega_k \Delta t_p(\gamma)}\}. \tag{9}$$

The exponential term in (9) corresponds to the phase shift applied to the pair-specific cross-correlation $\psi_p(\omega_k)$ in order to obtain a frequency-domain delay-and-sum beamformer steered towards candidate direction γ [13]. Considering a total of N SRP samples obtained at angles from the SRP grid $\Gamma_N = \{\gamma_1, \ldots, \gamma_N\}$, we define the broadband SRP map vector $\mathbf{z} \in \mathbb{R}^N$ as

$$\mathbf{z} = [z(\gamma_1),\ \ldots\ , z(\gamma_N)]^\top, \tag{10}$$

which can be expressed as a matrix-vector product as follows. We define, for a microphone pair p, the matrix $\mathbf{H}_p \in \mathbb{C}^{N \times K}$ and the vector $\boldsymbol{\psi}_p \in \mathbb{R}^K$, respectively, as

$$[\mathbf{H}_p]_{n,k} = e^{j\omega_k \Delta t_p(\gamma_n)}, \tag{11}$$

$$[\boldsymbol{\psi}_p]_k = \psi_p(\omega_k). \tag{12}$$

By stacking multiple matrices $\mathbf{H}_p$ and vectors $\boldsymbol{\psi}_p$, respectively, as

$$\mathbf{H} = [\mathbf{H}_1,\ \ldots\ , \mathbf{H}_P], \tag{13}$$

$$\boldsymbol{\psi} = \left[\boldsymbol{\psi}_1^\top,\ \ldots\ , \boldsymbol{\psi}_P^\top\right]^\top, \tag{14}$$

the broadband SRP map $\mathbf{z}$ can be expressed as

$$\mathbf{z} = 2\Re\{\mathbf{H}\boldsymbol{\psi}\}. \tag{15}$$

Conventionally, in a scenario with a single broadband source at DOA ϑ_1, performing the DOA estimation using an observed

[1] We note that there may be cross-correlation terms between the source signals and the undesired component, stemming from the latter's broad definition in (2). However, such cross-correlation terms do not impact the formulation of the optimization problem within the proposed method.

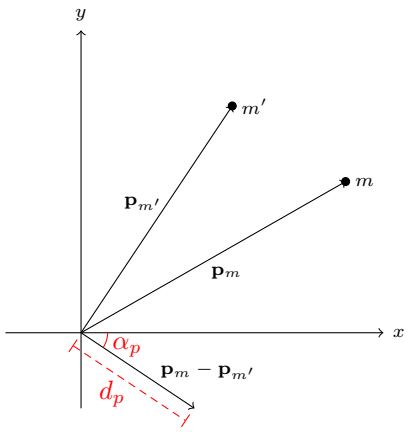


Fig. 1: Example of a microphone pair on the horizontal plane with its corresponding inter-microphone Euclidean distance d_p and rotation α_p.

SRP map involves finding the index n from the set of candidate DOAs that yields the maximum value of $z(\gamma_n)$:

$$\hat{n} = \arg\max_n z(\gamma_n), \tag{16}$$

such that the single source's estimated DOA is given by

$$\hat{\vartheta}_1 = \gamma_{\hat{n}}. \tag{17}$$

B. Uniform spatial sampling requirements

In order to describe the uniform spatial sampling requirements [2] of a broadband SRP map used for performing DOA estimation, we start by analyzing the TDOA expression of a given microphone pair. As previously presented in (1), the TDOA depends on the incident direction vector $\mathbf{r}_\gamma$, which, in the case of far-field propagation and sound sources on the same horizontal plane as a planar microphone array, corresponds to [14]

$$\mathbf{r}_\gamma = -[\cos\gamma, \sin\gamma, 0]^\top. \tag{18}$$

The vector $(\mathbf{p}_m - \mathbf{p}_{m'})$ in (1) is given by

$$(\mathbf{p}_m - \mathbf{p}_{m'}) = \begin{bmatrix} x_m - x_{m'} \\ y_m - y_{m'} \\ 0 \end{bmatrix} = \begin{bmatrix} d_p \cos\alpha_p \\ d_p \sin\alpha_p \\ 0 \end{bmatrix} \tag{19}$$

where d_p denotes the Euclidean distance of the microphone pair as in $d_p = \|\mathbf{p}_m - \mathbf{p}_{m'}\|_2$, and α_p denotes the microphone pair's rotation on the horizontal plane. In Fig. 1, an example of a microphone pair on the horizontal plane and its corresponding distance d_p and rotation α_p are illustrated.

[2]The sampling requirements specified in this section are related to the resolution required by the broadband SRP function such that, when computed for a grid of candidate DOAs, an SRP map with a distinct peak is obtained. Hence, these requirements do not refer to the relation between the spacing between microphones and the signal's wavelength, for which the effect of spatial aliasing is commonly defined [60].

By decomposing the cross-correlation $\psi_p(\omega_k)$ into its magnitude and phase components $|\psi_p(\omega_k)|$ and $\angle\psi_p(\omega_k)$, respectively, the broadband SRP in (9) can be reformulated as

$$z(\gamma) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} \Re\{|\psi_p(\omega_k)| e^{j\angle\psi_p(\omega_k)} e^{j\omega_k \Delta t_p(\gamma)}\} \tag{20a}$$

$$= 2\sum_{p=1}^{P}\sum_{k=1}^{K} |\psi_p(\omega_k)| \cos(\angle\psi_p(\omega_k) + \omega_k \Delta t_p(\gamma)), \tag{20b}$$

which corresponds to, essentially, a weighted sum of cosines with frequencies being dependent on $\Delta t_p(\gamma)$. By substituting (18) and (19) into (1), the TDOA expression becomes $\Delta t_p(\gamma) = \frac{d_p}{c}\cos(\alpha_p - \gamma)$, which allows for the broadband SRP in (20b) to be reformulated as

$$z(\gamma) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} |\psi_p(\omega_k)| \cos\left(\angle\psi_p(\omega_k) - \tfrac{\omega_k d_p}{c}\cos(\alpha_p - \gamma)\right). \tag{21}$$

To obtain the approximate spatial frequency of $z(\gamma)$, we linearize the term $\cos(\alpha_p - \gamma)$ around an arbitrary look direction, denoted γ_0, by means of a first-order Taylor series approximation:

$$\cos(\alpha_p - \gamma) \approx \cos(\alpha_p - \gamma_0) + \left(\frac{\mathrm{d}}{\mathrm{d}\gamma}\cos(\alpha_p - \gamma)\Big|_{\gamma=\gamma_0}\right)(\gamma - \gamma_0)$$
$$= \cos(\alpha_p - \gamma_0) + \sin(\alpha_p - \gamma_0)(\gamma - \gamma_0), \tag{22}$$

with $\frac{\mathrm{d}}{\mathrm{d}\gamma}\cos(\alpha_p - \gamma)\Big|_{\gamma=\gamma_0}$ denoting the derivative of $\cos(\alpha_p - \gamma)$ with respect to γ, evaluated in $\gamma = \gamma_0$. Hence, by inserting the above approximation in (21), we obtain

$$z(\gamma) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} |\psi_p(\omega_k)| \cos\big(\angle\psi_p(\omega_k) - \tfrac{\omega_k d_p}{c}(\cos(\alpha_p - \gamma_0) + \sin(\alpha_p - \gamma_0)(\gamma - \gamma_0))\big), \tag{23}$$

such that the spatial frequency of $z(\gamma)$ at look direction γ_0, for frequency ω_k and microphone pair p, is given by

$$\bar{\omega}_{k,p,\gamma_0} = \frac{\omega_k d_p}{c}\sin(\alpha_p - \gamma_0). \tag{24}$$

Based on the Nyquist-Shannon sampling theorem [61], and by assuming the microphone signals to be bandlimited by ω_b, i.e., $\psi_p(\omega_k) = 0$ for $\omega_k > \omega_b$, we define, for microphone pair p, the corresponding critical spatial sampling period (in radians) as

$$\Delta\gamma_{p,\gamma_0} = \frac{\pi}{\max_{\omega_k \le \omega_b} \bar{\omega}_{k,p,\gamma_0}} = \frac{\pi c}{\omega_b d_p \sin(\alpha_p - \gamma_0)}. \tag{25}$$

Note that the smallest value of $\Delta\gamma_{p,\gamma_0}$ for a given microphone pair p, denoted $\Delta\gamma_{p_{\min}}$, occurs at $\gamma_0 = \alpha_p - \pi/2$ and is equal to $\Delta\gamma_{p_{\min}} = (\pi c)/(\omega_b d_p)$. This observation is consistent with the one presented in [50]. Additionally, note that for a given microphone array configuration, the microphone pair with the largest inter-microphone distance, denoted $d_{\max}$, describes the smallest critical spatial sampling period over the entire array, corresponding to

$$\Delta\gamma = \frac{\pi c}{\omega_b d_{\max}}. \tag{26}$$

Therefore, if the SRP grid Γ_N used to compute the SRP map $\mathbf{z}$ is intended to be spatially uniform, one could consider $\Delta\gamma$ as a lower bound for the grid resolution, which guarantees that $z(\gamma)$ is sufficiently sampled over γ for all microphone pairs and look directions. In that case, the minimum number of required SRP samples N_{min}, uniformly distributed over 2π, is given by

$$N_{\text{min}} = \left\lceil \frac{2\pi}{\Delta\gamma} \right\rceil = \left\lceil \frac{2\omega_b d_{\text{max}}}{c} \right\rceil. \tag{27}$$

The knowledge of N_{min} is particularly relevant for compact microphone arrays. It has the potential to indicate whether the number of SRP samples to be considered, and therefore, the computational complexity of SRP-based DOA estimation methods, can be reduced while maintaining the highest achievable spatial resolution of the SRP map, given the available array configuration.

IV. Proposed Method

The conventional SRP approach detailed in Sec. III-A is evidently straightforward when performing single-source localization. However, when considering the case of multiple sources, especially under adverse acoustic conditions, such approach can be subject to DOA estimation errors caused by the presence of spurious peaks in the observed SRP map, as well as to the inability to resolve closely located sources. Therefore, in the following section, we present a method that aims to further improve the DOA estimation of multiple sources while considering the aforementioned limitations of the conventional SRP approach.

A. Group-sparse fitting of a broadband SRP map

We define a candidate source PSD vector $\boldsymbol{\varphi}_{S,l} \in \mathbb{R}^K$ as

$$\boldsymbol{\varphi}_{S,l} = \left[\phi_S(\omega_1, \theta_l) \;\; \dots \;\; \phi_S(\omega_K, \theta_l)\right]^\top, \tag{28}$$

where $\phi_S(\omega_k, \theta_l)$ is the candidate source PSD with DOA θ_l for frequency ω_k. By considering a set of L candidate DOAs $\Theta_L = \{\theta_1, \dots, \theta_L\}$ uniformly distributed over 2π and stacking different $\boldsymbol{\varphi}_{S,l}$ vectors, for $l = 1, \dots, L$, and $L \gg J$, the overcomplete vector $\boldsymbol{\varphi}_S \in \mathbb{R}^{KL}$ of direction-dependent source PSDs is obtained as

$$\boldsymbol{\varphi}_S = \left[\boldsymbol{\varphi}_{S,1}^\top \;\; \dots \;\; \boldsymbol{\varphi}_{S,L}^\top\right]^\top. \tag{29}$$

In addition, we define a vector $\mathbf{a}_{p,l} \in \mathbb{C}^K$ composed of the frequency-domain phase shifts related to a specific microphone pair p and candidate direction θ_l as

$$\mathbf{a}_{p,l} = \left[e^{-j\omega_1 \Delta t_p(\theta_l)} \;\; \dots \;\; e^{-j\omega_K \Delta t_p(\theta_l)}\right]^\top, \tag{30}$$

such that we can obtain a matrix $\mathbf{A} \in \mathbb{C}^{KP \times KL}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \dots & \mathbf{A}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{P,1} & \dots & \mathbf{A}_{P,L} \end{bmatrix}, \tag{31}$$

where $\mathbf{A}_{p,l} = \text{diag}\left(\mathbf{a}_{p,l}\right)$ Based on (7), we model the cross-correlation vector $\boldsymbol{\psi}$ in (14) as

$$\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\varphi}_S, \tag{32}$$

so that the broadband SRP vector in (15) can be expressed as

$$\mathbf{z} = 2\Re\{\mathbf{H}\mathbf{A}\boldsymbol{\varphi}_S\} = 2\Re\{\mathbf{H}\mathbf{A}\}\boldsymbol{\varphi}_S = \mathbf{B}\boldsymbol{\varphi}_S, \tag{33}$$

where the second equality follows from the PSDs being real-valued, and where we define $\mathbf{B} = 2\Re\{\mathbf{H}\mathbf{A}\}$. With this, we propose to estimate the PSDs corresponding to the candidate DOAs, and thereby the source DOAs, by solving

$$\underset{\boldsymbol{\varphi}_S \geq 0}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S\right\|_2 + \sum_{l=1}^{L} \lambda_l \left\|\boldsymbol{\varphi}_{S,l}\right\|_2 \tag{34}$$

where λ_l is a regularization parameter. The proposed optimization problem's formulation corresponds to a nonnegative group square-root Lasso problem [54]. The regularization term $\sum_{l=1}^{L} \lambda_l \left\|\boldsymbol{\varphi}_{S,l}\right\|_2$ promotes solutions with a group-sparse structure, i.e., only a few of the subvectors $\boldsymbol{\varphi}_{S,l}$ composing $\boldsymbol{\varphi}_S$ should be nonzero, which agrees with the assumption of having a limited number of point sources ($J \ll L$). By estimating $\hat{\boldsymbol{\varphi}}_S$, and consequently, $\hat{\boldsymbol{\varphi}}_{S,1}, \dots, \hat{\boldsymbol{\varphi}}_{S,L}$, the source signals' DOAs are estimated by identifying the peaks on the spatial power density, obtained by integrating the PSD values over frequency for each subvector $\hat{\boldsymbol{\varphi}}_{S,l}$, as previously proposed in [48]. The proposed group-sparse optimization-based method is motivated by the potential to reveal more distinct peaks in the spatial power density than those directly observed in the broadband SRP map [17], [42], [43], which can positively contribute to the robustness of the DOA estimation performance in scenarios where multiple sources need to be distinguished under varying acoustic conditions. For a total of J sources, the estimated DOAs $\hat{\vartheta}_1, \dots, \hat{\vartheta}_J$ are then chosen to be the candidate directions related to the J largest peaks. One may note that the grid of candidate DOAs Θ_L employed to compose $\boldsymbol{\varphi}_S$ is not required to be equal to the SRP grid Γ_N used to obtain the observed SRP map $\mathbf{z}$, as opposed to the approach previously proposed in [48].

The choice of formulating a group square-root Lasso problem as opposed to the more conventional group Lasso [49] relates to the possibility of tuning the regularization parameter λ_l, which affects the level of group sparsity in the solution obtained, independently from the noise level [54]. Under the conditions of homoscedastic noise, it has been shown in [55] that the group square-root Lasso is equivalent to a relaxed version of the covariance fitting problem, which motivates the selection of the regularization parameter based on the corresponding covariance fitting criteria [62]. In that case, considering same-size groups in the definition of $\boldsymbol{\varphi}_S$, the reference value $\tilde{\lambda}_l$ according to which the optimal choice of λ_l is scaled is given by [54], [55]

$$\tilde{\lambda}_l = \sqrt{\frac{\left\|\left[\|\mathbf{b}_{(l-1)K+1}\|_2^2 \;\; \dots \;\; \|\mathbf{b}_{lK}\|_2^2\right]^\top\right\|_2}{N}}, \tag{35}$$

for $l = 1, \dots, L$, with $\mathbf{b}_q$ denoting the q-th column of $\mathbf{B}$.

B. Group square-root Lasso implementation

In order to solve the problem in (34), we present an implementation based on the alternating direction method of multipliers (ADMM) [56].

By using auxiliary variables, denoted $\mathbf{u}_1$ and $\mathbf{u}_2$, the problem in (34) can be reformulated as

$$\underset{\mathbf{u}_1,\mathbf{u}_2\geq 0}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{u}_1\right\|_2 + \sum_{l=1}^{L}\lambda_l \left\|\mathbf{u}_{2,l}\right\|_2 \tag{36a}$$

$$\text{subject to} \quad \mathbf{u}_1 = \mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S \tag{36b}$$

$$\mathbf{u}_2 = \boldsymbol{\varphi}_S \tag{36c}$$

where $\mathbf{u}_{2,l}$ corresponds to the l-th group composing the vector $\mathbf{u}_2$, such that it can be represented as

$$\mathbf{u}_2 = \begin{bmatrix}\mathbf{u}_{2,1}^\top & \dots & \mathbf{u}_{2,L}^\top\end{bmatrix}^\top. \tag{37}$$

The augmented Lagrangian of the optimization problem in (36) is given by

$$\begin{aligned}\mathcal{L}_\mu(\boldsymbol{\varphi}_S,\mathbf{u}_1,\mathbf{u}_2,\mathbf{d}_1,\mathbf{d}_2) = \frac{1}{2}\left\|\mathbf{u}_1\right\|_2 + \sum_{l=1}^{L}\lambda_l\left\|\mathbf{u}_{2,l}\right\|_2 \\ + \frac{\mu}{2}\left\|\mathbf{z}-\mathbf{B}\boldsymbol{\varphi}_S-\mathbf{u}_1-\mathbf{d}_1\right\|_2^2 \\ + \frac{\mu}{2}\left\|\boldsymbol{\varphi}_S-\mathbf{u}_2-\mathbf{d}_2\right\|_2^2\end{aligned} \tag{38}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the (scaled) dual variables, and μ can be interpreted as a dual update step size [56].

For a given ADMM iteration, indexed as q, we define the following shorthands:

$$\boldsymbol{\zeta}_1(q) = \mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S(q+1) - \mathbf{d}_1(q), \tag{39}$$

$$\boldsymbol{\zeta}_2(q) = \boldsymbol{\varphi}_S(q+1) - \mathbf{d}_2(q), \tag{40}$$

such that the updates for each variable in (38) can be written as

$$\boldsymbol{\varphi}_S(q+1) = \underset{\boldsymbol{\varphi}_S}{\arg\min}\,\mathcal{L}_\mu(\boldsymbol{\varphi}_S,\mathbf{u}_1(q),\mathbf{u}_2(q),\mathbf{d}_1(q),\mathbf{d}_2(q)), \tag{41}$$

$$\mathbf{u}_1(q+1) = \underset{\mathbf{u}_1}{\arg\min}\,\frac{1}{2}\left\|\mathbf{u}_1\right\|_2 + \frac{\mu}{2}\left\|\boldsymbol{\zeta}_1(q)-\mathbf{u}_1\right\|_2^2, \tag{42}$$

$$\mathbf{u}_2(q+1) = \underset{\mathbf{u}_2\geq 0}{\arg\min}\sum_{l=1}^{L}\lambda_l\left\|\mathbf{u}_{2,l}\right\|_2 + \frac{\mu}{2}\left\|\boldsymbol{\zeta}_2(q)-\mathbf{u}_2\right\|_2^2, \tag{43}$$

$$\mathbf{d}_1(q+1) = \mathbf{u}_1(q+1) - \boldsymbol{\zeta}_1(q), \tag{44}$$

$$\mathbf{d}_2(q+1) = \mathbf{u}_2(q+1) - \boldsymbol{\zeta}_2(q). \tag{45}$$

It can be readily verified that the updates for the primal variables are given by

$$\begin{aligned}\boldsymbol{\varphi}_S(q+1) = \left(\mathbf{B}^\top\mathbf{B}+\mathbf{I}_{KL}\right)^{-1}\left[\mathbf{B}^\top(\mathbf{z}-\mathbf{u}_1(q)-\mathbf{d}_1(q))\right. \\ \left.+\mathbf{u}_2(q)+\mathbf{d}_2(q)\right],\end{aligned} \tag{46}$$

$$\mathbf{u}_1(q+1) = \mathrm{T}\left(\boldsymbol{\zeta}_1(q), 1/\mu\right), \tag{47}$$

$$\mathbf{u}_{2,l}(q+1) = \mathrm{T}\left(\max\left(\boldsymbol{\zeta}_{2,l}(q),0\right), \frac{\lambda_l}{\mu}\right), \tag{48}$$

where $\mathbf{I}_{KL}$ denotes a $KL\times KL$ identity matrix, and $\mathrm{T}(\mathbf{x},\kappa)$ is defined as the soft-thresholding operator

$$\mathrm{T}(\mathbf{x},\kappa) = \left[\max\left(1-\frac{\kappa}{\|\mathbf{x}\|_2},0\right)\right]\mathbf{x}. \tag{49}$$

In (48), the update of $\mathbf{u}_2$ is computed group-wise, for $l = 1,\dots,L$. After computing the updates of $\mathbf{d}_1$ and $\mathbf{d}_2$ as in (44) and (45), respectively, the whole iterative process is repeated until convergence [56]. The proposed method's entire framework, employing the presented ADMM-based implementation, is summarized in Algorithm 1. It may be noted that the per-iteration complexity is dominated by (46), which, if implemented by using a Cholesky factorization of $(\mathbf{B}^\top\mathbf{B}+\mathbf{I}_{KL})$, has a complexity of $\mathcal{O}(K^2L^2)$. Furthermore, if $N < KL$, then the per-iteration complexity is $\mathcal{O}(NKL)$.

Algorithm 1 The proposed DOA estimation algorithm

initiate $\hat{\boldsymbol{\varphi}}_S = \mathbf{0}^{KL\times 1}$
initiate $j := 0$, $\mathbf{u}_2(0) = \hat{\boldsymbol{\varphi}}_S$,
  $\boldsymbol{\varphi}_S(0) = \mathbf{d}_2(0) = \mathbf{0}^{KL\times 1}$, $\mathbf{u}_1(0) = \mathbf{d}_1(0) = \mathbf{0}^{N\times 1}$
repeat {ADMM scheme}
  update $\boldsymbol{\varphi}_S(j+1)$ as in (46)
  update $\mathbf{u}_1(j+1)$ as in (47)
  for $l \leftarrow 1$ to L **do**
    update $\mathbf{u}_{2,l}$ as in (48)
  end for
  $\mathbf{d}_1(j+1) = \mathbf{u}_1(j+1) - \boldsymbol{\zeta}_1(j)$
  $\mathbf{d}_2(j+1) = \mathbf{u}_2(j+1) - \boldsymbol{\zeta}_2(j)$
  $j \leftarrow j+1$
until convergence
store $\hat{\boldsymbol{\varphi}}_S = \mathbf{u}_2(\text{end})$
estimate $\hat{\vartheta}_i$, for $i = 1,\dots,J$, as the J largest peaks integrated over frequency in $\hat{\boldsymbol{\varphi}}_S$

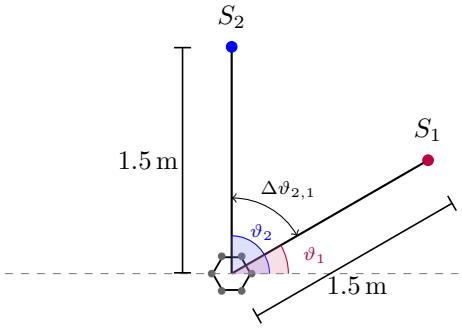


Fig. 2: Illustration of a simulated scenario with two sources at a radial distance of 1.5 m from the center of a 6-element uniform circular array.

V. Evaluation

In order to evaluate the DOA estimation performance of the proposed method, a series of simulations is performed under different acoustic conditions. In Sec. V-A, the parameters describing the different simulated acoustic scenarios are defined. In Sec. V-B, the choice of the search grids of candidate DOAs Γ_N and Θ_L are detailed. In Sec. V-C, the methods with which the proposed method is compared, as well as the the performance metric used to evaluate the DOA estimation, are specified. The particular characteristics of each scenario and the results obtained are further detailed and discussed in Sec. VI.

A. Simulation setup

Different far-field propagation scenarios are simulated in a rectangular room with dimensions $6.3\times 5.1\times 2.9$ m. A

uniform circular array with a radius of $5\,\mathrm{cm}$ and $M = 6$ microphones is considered to be centered at Cartesian coordinates $[3.1, 2.7, 1.5]^\top\,\mathrm{m}$, with a total of $P = 15$ distinct microphone pairs, and the largest inter-microphone distance corresponding to the array's diameter ($d_{\max} = 0.1\,\mathrm{m}$). The implementation of the image source method [63] presented in [64] is used for simulating the room acoustics, characterized by the reverberation time (T_{60}).

The simulated point sources are denoted by S_i, where i is the source index, and are placed in the same horizontal plane as the microphone array, at a radial distance of $1.5\,\mathrm{m}$ from the center of the array. Each source emits a randomly selected segment of a pre-determined signal type. The signal type of each source and the section in which it is present is detailed in Table I. With the exception of Sec. VI-F, all sources are simulated with equal broadband power. Spatially white speech-shaped noise is generated by filtering white Gaussian noise with a 16th-order finite impulse response (FIR) linear prediction filter obtained from a babble noise signal [65]. With $\sigma^2_{S_i}$ denoting the broadband power of the i-th source's reverberant signal at the array's first microphone, the signal-to-noise ratio (SNR) is defined as

$$\mathrm{SNR} = \frac{\sigma^2_{S_1}}{\sigma^2_n}, \tag{50}$$

where σ^2_n denotes the broadband power of the spatially white component. With a sampling rate of $f_s = 16\,\mathrm{kHz}$, the simulated time-domain microphone signals have a duration of $3\,\mathrm{s}$, and their short-time Fourier transform (STFT) is computed using a 512-point square-root Hann window and 50% overlap, resulting in a time-domain window of $32\,\mathrm{ms}$ and a total of 187 frames. We assume that in all scenarios, the microphone signal's bandwidth is limited by $\omega_b = 2\pi f_b$, with $f_b = 8\,\mathrm{kHz}$. The cross-correlations in (5) are obtained by averaging the PHAT-weighted STFT cross-correlations over a total of N_f frames [3]. Considering the single-sided broadband SRP formulation in (9), a total of $K = 256$ frequency bins are used.

For each scenario, a total of $R = 100$ realizations are performed, and for each realization r, the first source's DOA $\left(\vartheta_1^{(r)}\right)$ is randomly generated over the range of $360°$ without constraining it to lie on the candidate DOA search grids Γ_N and Θ_L being used. The remaining sources' DOAs are then set as

$$\vartheta_i^{(r)} = \vartheta_{i-1}^{(r)} + \Delta\vartheta_{i,i-1}^{(r)}, \quad \text{for } i = 2, \ldots, J, \tag{51}$$

where $\Delta\vartheta_{i,i-1}^{(r)}$ denotes the angular separation between two neighboring sources for realization r. This separation is randomly selected within a specified range in each scenario considered, and an example illustrating the spatial configuration of two sources is presented in Fig. 2.

In addition to the aforementioned simulated acoustic setups, we utilize measured room impulse responses (RIRs) from the database detailed in [57] to assess the proposed method's DOA estimation performance within a realistic meeting room environment, and the corresponding scenarios being tested are further specified in Sec. VI-I.

[3]Since the weighting operation is not accounted for in the cross-correlation model presented in Sec. II, the PSD vector estimated with the proposed method will be inherently weighted as well in this simulation setup. However, due to the DOA estimation being based solely on the location of the nonzero elements of the PSD vector, it is expected that the proposed method's overall performance should not be affected.

TABLE I: Source signal types.

Source	Signal type	Section
S_1	speech (male speaker) [66]	VI-A, VI-G
S_2	speech (female speaker) [66]	VI-A, VI-G
S_3	speech (male speaker) [66]	VI-G
S_4	speech (female speaker) [66]	VI-G
S_5	instrumental music (guitar) [67]	VI-G

B. Search grids of candidate DOAs

The set of candidate DOAs Θ_L, used in the proposed method's optimization problem to define the candidate source PSD vector $\boldsymbol{\varphi}_S$, is generated as a search grid with a fixed resolution of $5°$, which yields a total of $L = 72$ candidate DOAs, uniformly distributed over $360°$. The SRP grid Γ_N, used to compute the observed SRP map $\mathbf{z}$, is defined as a subset of Θ_L ($\Gamma_N \subseteq \Theta_L$), such that the N SRP samples composing $\mathbf{z}$ are also uniformly distributed over $360°$. Based on the sampling requirements determined in Sec. III-B, we note that, due to the largest inter-microphone distance of the uniform circular array corresponding to its diameter ($d_{\max} = 0.1\,\mathrm{m}$), the SRP map $\mathbf{z}$ should be composed, in theory, of at least $N_{\min} = 30$ uniformly distributed samples over the range of $360°$, which corresponds to a grid resolution of approximately $\Delta\gamma \approx 12.2°$. Therefore, varying N allows for analyzing the different levels at which the possible spatial aliasing effects over the observed SRP map $\mathbf{z}$ can affect the DOA estimation performance.

C. Compared methods and performance metric

The proposed method is implemented as in Algorithm 1, with the regularization parameter set to $\lambda_l = 0.001\tilde{\lambda}_l$, with $\tilde{\lambda}_l$ defined in (35), for $l = 1, \ldots, L$. We note that although this heuristic approach to the selection of λ_l may be suboptimal, it is motivated by the objective of presenting a performance evaluation of the proposed method that is generalisable across the different acoustic conditions considered in this study. The proposed method's DOA estimation performance is compared to that of the conventional SRP-PHAT method, such that the effect of locating peaks in the estimated direction-specific PSDs rather than directly over the observed broadband SRP map can be assessed in the DOA estimation performance. The proposed method is also compared with another group-sparse regression approach, namely the group-sparse iterative covariance-based estimation (GSPICE) method, proposed in [55] and briefly explained in the Appendix. The GSPICE method is here used to estimate the source PSD vector $\boldsymbol{\varphi}_S$, and consequently, the sources' DOAs, by directly fitting the microphone array cross-correlations vector $\boldsymbol{\psi}$ to the proposed source PSD model, according to (32). The comparison between the proposed method and GSPICE primarily aims to evaluate the effect in DOA estimation performance and computational complexity of combining pairwise GCC information and modeling an observed SRP map as a linear function of

TABLE II: Main parameters for each set of results, divided by subsections.

Section	N	N_f	$\Delta\vartheta^{(r)}_{i,i-1}$	SNR	T_{60}	J
VI-A	12, 18, 24, 36, 72	64	$[15°, 30°]$	0 dB	0s	2
VI-B	24, 72	1, 16, 32, 64, 128	$[15°, 30°]$	0 dB	0s	2
VI-C	24, 72	64	$[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, $[120°, 180°]$	0 dB	0.2 s	2
VI-D	24, 72	64	$[15°, 30°]$	-18 dB to 18 dB	0.2 s	2
VI-E	24, 72	64	$[15°, 30°]$	0 dB	0 s to 0.6 s	2
VI-F	24, 72	64	$[15°, 30°]$	0 dB	0.2 s	2
VI-G	24, 72	64	$[15°, 30°]$	0 dB	0 s	1 to 5
VI-H	24, 72	64	$[15°, 30°]$	0 dB	0 s	2
VI-I	24, 72	64	$[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, $[120°, 180°]$	0 dB	0.659 s	2

PSDs, compared to directly utilizing pairwise GCCs for this purpose. Finally, the proposed method is also compared with the wideband MUSIC method with geometric averaging [9], [58], which will be referred to as WMUSIC in the remainder. We note that the choice of the SRP grid Γ_N does not apply to GSPICE and WMUSIC in this comparison, as both methods are employed considering the set of candidate DOAs Θ_L. For all methods, the DOAs are estimated by finding the J largest local maxima on the corresponding maps obtained, and for the proposed method and GSPICE, the convergence criteria are defined with respect to the relative change in the estimated vectors between iterations, with a maximum number of iterations imposed.

The absolute DOA estimation error for a specific source and realization is computed as

$$\epsilon_i^{(r)} = |\mathrm{mod}(\vartheta_i^{(r)} - \hat{\vartheta}_i^{(r)}, 360°)|, \quad \text{for } i = 1, \ldots, J, \tag{52}$$

and the score indicating whether a source is considered to be successfully located is defined as

$$\rho_i^{(r)} = \begin{cases} 1, & \text{if } \epsilon_i^{(r)} \leq \epsilon_{th} \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

where ϵ_{th} corresponds to an absolute error threshold, in degrees. Here, ϵ_{th} is set to 5° in all cases. With the exception of Sec. VI-F, in which the DOA estimation success is evaluated per source, the success rate used for evaluation is computed as the average number of successes over all sources and realizations of the scenario being considered, as

$$\bar{\rho} = \frac{1}{JR} \sum_{i=1}^{J} \sum_{r=1}^{R} \rho_i^{(r)}. \tag{54}$$

VI. Results

In this section, we analyze the effect of each parameter described in Sec. V-A on the DOA estimation performance of the proposed method, in comparison with SRP-PHAT, GSPICE and WMUSIC. In Table II, a summary of the parameter values, used in each setup configuration being considered in the following subsections, is presented.

A. Number of SRP samples and computational complexity

Firstly, we aim to observe the effect of the choice of the candidate DOA search grid Γ_N used to compute the observed SRP map $\mathbf{z}$ in the DOA estimation performance of the proposed method and the conventional SRP-PHAT method. We also evaluate the computational complexity of the proposed method compared to SRP-PHAT, GSPICE and WMUSIC. In this context, the computational complexity of SRP-PHAT is here determined by the cost of computing $\mathbf{z}$, which corresponds to $\mathcal{O}(NKP)$. For GSPICE, the computational complexity corresponds to $\mathcal{O}(K^2L^2)$ [55]. The complexity of WMUSIC is dominated by the per-frequency eigenvalue decomposition and computation of the pseudospectrum [68], and corresponds to $\mathcal{O}(M^3K + M^2KL)$. For the proposed method, as previously mentioned in Sec. IV-B, the computational complexity corresponds to $\mathcal{O}(NKL)$.

We consider a scenario with two sources (S_1 and S_2), with an angular separation of $\Delta\vartheta^{(r)}_{2,1} \in [15°, 30°]$. The SNR is set to 0 dB and the simulations are performed under anechoic conditions ($T_{60} = 0$ s). A total of $N_f = 64$ STFT cross-correlation frames are averaged in order to obtain the cross-correlation vector $\boldsymbol{\psi}$, and the number of points composing the SRP search grid Γ_N and used to compose the SRP map vector $\mathbf{z}$ is varied as $N = 12, 18, 24, 36, 72$, yielding the grid resolutions of $30°, 20°, 15°, 10°$ and $5°$, respectively.

In Fig. 3a, the success rates obtained with each method, and for the case of the proposed method and SRP-PHAT, with different values of N, are presented. The shaded areas depict the 95% binomial proportion confidence interval computed for each method's success rate, based on the Clopper-Pearson method [69], and this representation will be maintained in the following sections. It can be observed that, in this anechoic scenario, the proposed method overall outperforms SRP-PHAT for varying N. Moreover, for $N \geq 24$, the proposed method attains a performance that is comparable to GSPICE and overall better than WMUSIC. Since the minimum number of required SRP samples for the scenario considered corresponds to $N_{\text{min}} = 30$ (cf. Sec. V-B), this result indicates that while the knowledge of N_{min} can be beneficial for having a reference when designing the SRP grid Γ_N, a sufficiently robust per-

formance can nonetheless still be achieved with the proposed method for lower grid resolutions of the SRP map $\mathbf{z}$ being employed, and therefore, with a reduction in computational complexity. Based on these observations, and for ease of interpretation, the simulation results in the following subsections are presented for $N = 24$ and $N = 72$.

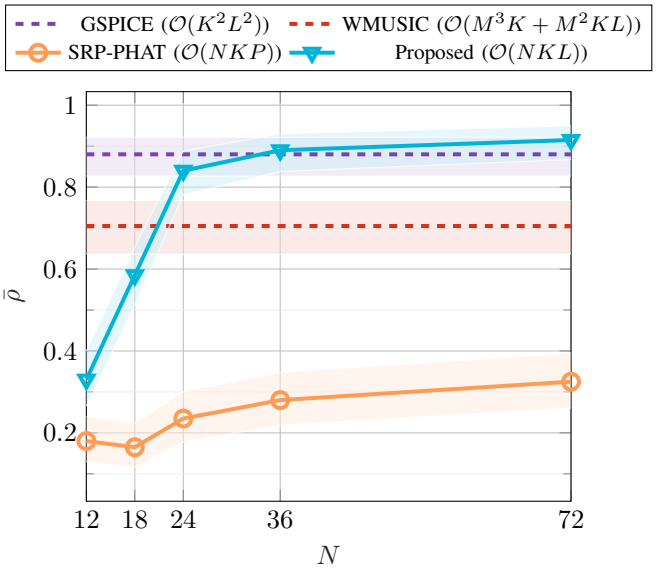
In Fig. 3b, the computational complexity of each method in terms of the number of multiplications per iteration (N_{mult}), and for the case of the proposed method and SRP-PHAT, with different values of N, is presented. We note that, in this set of simulations, the proposed method required on average 90 iterations to reach convergence, whereas GSPICE required on average 40 iterations. Nevertheless, the average runtime for the proposed method was lower than for GSPICE. When the complexity is evaluated alongside the success rates presented in Fig. 3a, we can observe that despite the proposed method being more computationally expensive than the conventional (and non-terative) SRP-PHAT and WMUSIC methods, it still allows for an improvement in DOA estimation performance that is comparable to GSPICE, at a significantly lower computational cost per iteration.

B. Number of averaging STFT cross-correlation frames

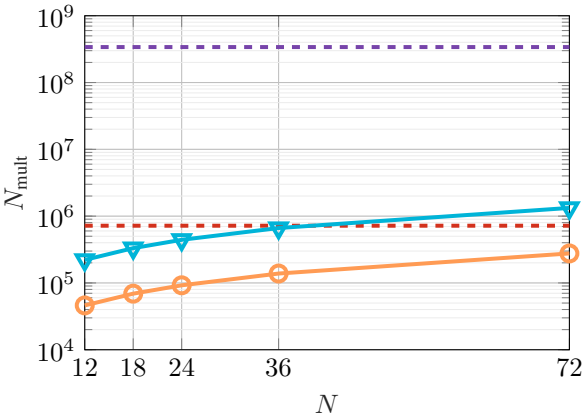
In order to observe the effect of averaging over multiple STFT cross-correlation frames on the DOA estimation performance of the proposed method, we again consider the case of two sources (S_1 and S_2), with an angular separation of $\Delta\vartheta_{2,1}^{(r)} \in [15°, 30°]$. The SNR is set to 0 dB and the simulations are performed under anechoic conditions ($\text{T}_{60} = 0\,\text{s}$). The number of STFT cross-correlation frames being considered (N_f) is varied between 1 and 128. In Fig. 4, the success rate obtained with the proposed method and SRP-PHAT for $N = 24$ and $N = 72$, as well as the success rate obtained with GSPICE and WMUSIC, is presented. It can be observed that in the two-source scenario considered, the DOA estimation performance improves for GSPICE, WMUSIC, and the proposed method, as the number of STFT cross-correlation frames N_f increases. However, in the case of SRP-PHAT, it is suspected that the close spacing between sources in the scenario tested causes the performance to be mainly limited by N, such that an increase in N_f did not substantially affect the method's performance, and the success rate was lower than all other methods in comparison. We can also observe that for $N_f \leq 64$, the proposed method overall performs better than WMUSIC and similar to GSPICE, at a lower computational complexity than the latter method, indicating a potential benefit of its use in cases where the number of available frames is limited.

C. Angular separation between sources

In this set of simulations, we aim to investigate the influence of the angular separation between two sources (S_1 and S_2) of equal broadband power on the DOA estimation performance obtained with the proposed method, the conventional SRP-PHAT, GSPICE, and WMUSIC. We vary the angular separation between the two sources such that, for each scenario considered, $\Delta\theta_{2,1}^{(r)}$ is within one of the following intervals: $[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, and $[120°, 180°]$.



(a) Success rate $\bar{\rho}$ as a function of N.



(b) Number of per-iteration multiplications N_{mult} as a function of N.

Fig. 3: Success rate and number of per-iteration multiplications N_{mult} as a function of N, with $L = 72$, SNR $= 0\,\text{dB}$ and $\text{T}_{60} = 0\,\text{s}$.

The SNR is set to 0 dB and $\text{T}_{60} = 0.2\,\text{s}$. We consider the cases where $N = 24$ and $N = 72$ for the proposed method and SRP-PHAT, and the success rates obtained with each method are presented in Fig. 5. We again observe that the performance of SRP-PHAT is more dependent on the choice of N than the one obtained when employing the proposed method. We also observe that, for all methods, the DOA estimation performance overall improves as the angular separation between sources increases, with the success rates of GSPICE, WMUSIC, SRP-PHAT ($N = 72$) and the proposed method being similar when the angular separation between sources is above 30°. Given this observation, we note that although employing the proposed method in such scenarios may not provide a distinct benefit in DOA estimation or computational complexity, it still outperforms GSPICE, WMUSIC and SRP-PHAT when $\Delta\theta_{2,1} \in [5°, 15°]$ and $\Delta\theta_{2,1} \in [15°, 30°]$, indicating its specific advantage in differentiating closely located sources.

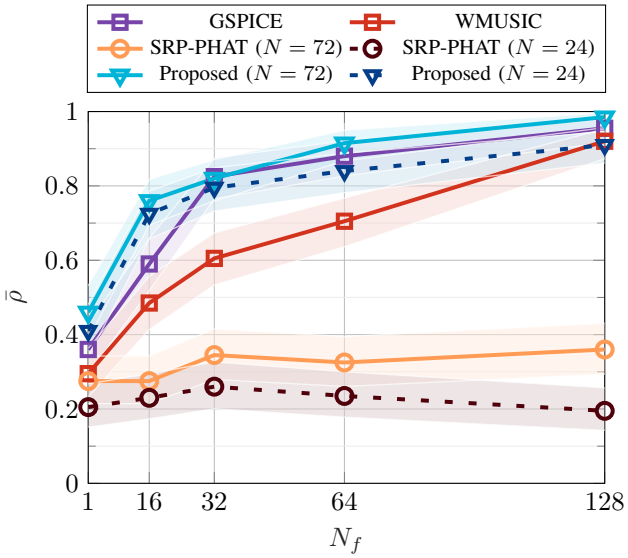


Fig. 4: Success rate obtained for different numbers of averaging STFT cross-correlation frames N_f, with SNR $= 0\,\mathrm{dB}$ and $\mathrm{T}_{60} = 0\,\mathrm{s}$.

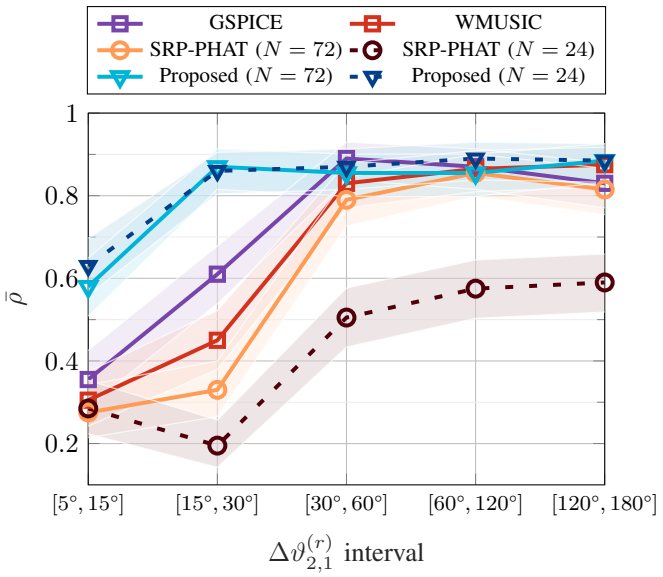


Fig. 5: Success rate obtained for different angular separations between sources, with SNR $= 0\,\mathrm{dB}$ and $\mathrm{T}_{60} = 0.2\,\mathrm{s}$.

D. Noise level

In this set of simulations, the effect of the spatially white noise level on the DOA estimation performance of the proposed method in comparison to the conventional SRP-PHAT, GSPICE, and WMUSIC is analyzed by varying the SNR between $-18\,\mathrm{dB}$ and $18\,\mathrm{dB}$, for the case of two sources (S_1 and S_2) of equal broadband power. The reverberation time is set to $\mathrm{T}_{60} = 0.2\,\mathrm{s}$ and the angular separation is $\Delta\vartheta_{2,1}^{(r)} \in [15°, 30°]$. The success rates obtained with each method for the different SNRs considered, as well as for $N = 24$ and $N = 72$ in the case of SRP-PHAT and the proposed method, are presented in Fig. 6. It can be observed

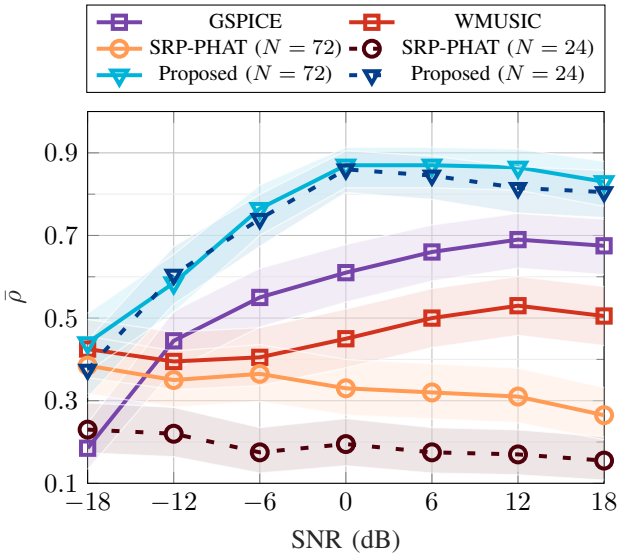


Fig. 6: Success rate obtained for different SNRs, with $\mathrm{T}_{60} = 0.2\,\mathrm{s}$.

that for GSPICE, WMUSIC and the proposed method, the DOA estimation performance overall improves as the SNR increases, with GSPICE and WMUSIC seemingly attaining their highest success rate at SNR$= 12\,\mathrm{dB}$, and the proposed method attaining its highest success rate at SNR$= 0\,\mathrm{dB}$. However, SRP-PHAT demonstrates less sensitivity to noise levels compared to the other methods. These observations suggest that the proximity between the simulated sources and the level of reverberation are likely the main factors limiting the performance of the compared methods. Notably, the proposed method outperforms all others for SNR$\geq -12\,\mathrm{dB}$, maintaining similar success rates for both $N = 24$ and $N = 72$.

E. Reverberation

The effect of reverberation on the DOA estimation performance of the proposed method is evaluated for the case of two sources (S_1 and S_2) of equal broadband power and with $\Delta\theta_{2,1}^{(r)} \in [15°, 30°]$. For SRP-PHAT and the proposed method, the number of SRP samples is again set to $N = 24$ and $N = 72$. The SNR is set to $0\,\mathrm{dB}$ and the reverberation time is varied between $0\,\mathrm{s}$ and $0.6\,\mathrm{s}$, with $\mathrm{T}_{60} = 0\,\mathrm{s}$ corresponding to the anechoic case. The success rates obtained with each method being compared are presented in Fig. 7, where it is observed that the proposed method, GSPICE and WMUSIC present a DOA estimation performance that is seemingly more sensitive to an increase in T_{60} than the conventional SRP-PHAT. Nevertheless, the proposed method still outperforms all compared methods when $\mathrm{T}_{60} \geq 0.2\,\mathrm{s}$, indicating its potential advantage in locating closely spaced sources in reverberant environments. As in the previous subsections, we again observe that the proposed method's average accuracy scores obtained with $N = 24$ and $N = 72$ mostly coincide.

F. Broadband power ratio between two sources

In order to investigate the effect of a possible power level difference between sound sources on the DOA estimation

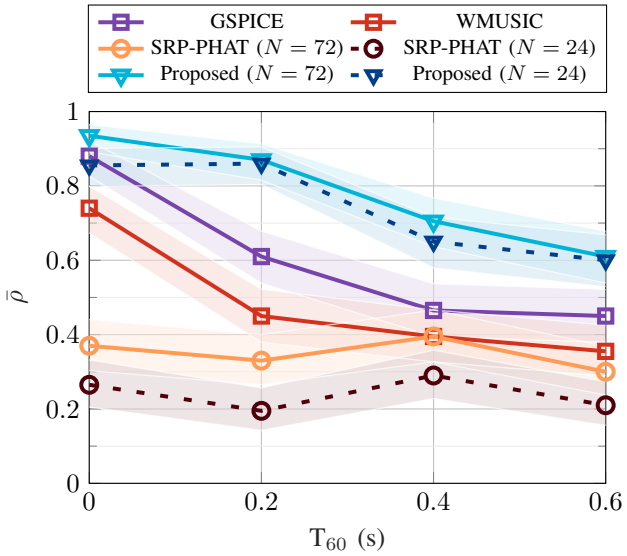
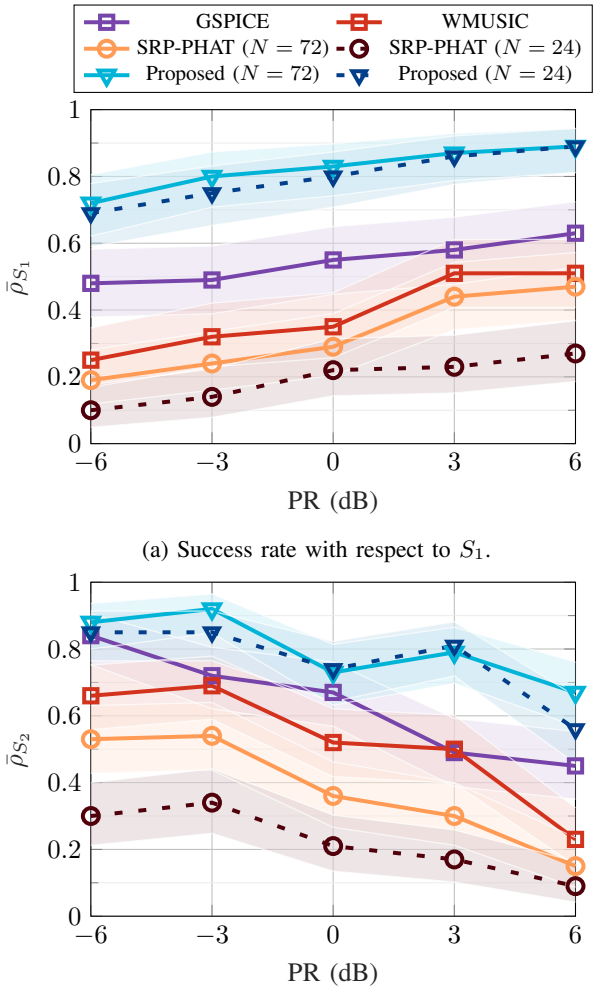


Fig. 7: Success rate obtained different reverberation times (T_{60}), with SNR $= 0\,\text{dB}$.



(a) Success rate with respect to S_1.

(b) Success rate with respect to S_2.

Fig. 8: Success rate per source obtained for different PRs, with SNR $= 0\,\text{dB}$ and $T_{60} = 0.2\,\text{s}$.

performance of the methods being compared, we consider the case of two sources (S_1 and S_2) with $\Delta\vartheta_{2,1}^{(r)} \in [15°, 30°]$, and we define a broadband power ratio (PR) between the two sources as

$$\text{PR} = \frac{\sigma_{S_1}^2}{\sigma_{S_2}^2}, \tag{55}$$

where $\sigma_{S_1}^2$ and $\sigma_{S_2}^2$ are defined as in Sec. V-A. Since the speech sources S_1 and S_2 correspond to one male and one female speaker, respectively, and naturally present different spectral characteristics, the power ratio defined in (55) is not expected to be uniform over all frequencies. The power ratio PR is varied between $-6\,\text{dB}$ and $6\,\text{dB}$, the SNR is set to $0\,\text{dB}$ and $T_{60} = 0.2\,\text{s}$. Differently from the previous evaluations presented in Sec. VI-A–VI-E, the success rate is here computed per source and denoted by $\bar{\rho}_{S_1}$ and $\bar{\rho}_{S_2}$ for sources S_1 and S_2, respectively. The results per source obtained with $N = 24$ and $N = 72$ SRP samples are shown in Fig. 8. As the value of PR increases, we can observe that, overall, the success rate for the first source increases, whereas it decreases for the second source. Moreover, we observe that the intersection between the success rates for each source and method does not necessarily occur exactly at PR $= 0\,\text{dB}$, and that the performance ranges related to S_1 and S_2 do not necessarily coincide, for the PR values considered. This is most likely due to S_1 and S_2 being a male and female speaker, respectively, and consequently presenting different spectral characteristics that may influence the DOA estimation task. Finally, we observe that the proposed method presents a more distinct advantage over the compared methods in localizing S_1 than S_2, when considering all PR values tested and for for both $N = 24$ and $N = 72$.

G. Number of sources

Up until now, only two-source scenarios have been considered while evaluating the proposed method's DOA estimation performance in comparison with SRP-PHAT, GSPICE, and WMUSIC. In order to further investigate the proposed method's performance for more challenging scenarios, we now vary the number of sources present in the simulated acoustic scene from 1 to 5. The sources' DOAs are generated as explained in Sec. V-A, with the angular separation between two subsequent sources $\Delta\vartheta_{i,i-1}^{(r)}$ set within the range of $[15°, 30°]$, while not being constrained to be equal between all neighboring sources. Each source is successively included in the simulated scenario in their corresponding order (S_1 to S_5). The SNR is set to $0\,\text{dB}$, and due to the increased number of sources in comparison to the previous sets of simulations, the reverberation time is set to $T_{60} = 0\,\text{s}$. We note that, for the SRP-PHAT method with $N = 24$, a score of $\rho_i^{(r)} = 0$ was assigned to the sources that could not be located due to an insufficient number of peaks in the observed SRP-PHAT map of the corresponding realization.

The success rates obtained with all compared methods are shown in Fig. 9. We can observe that the DOA estimation performance of all methods decrease as the number of simulated

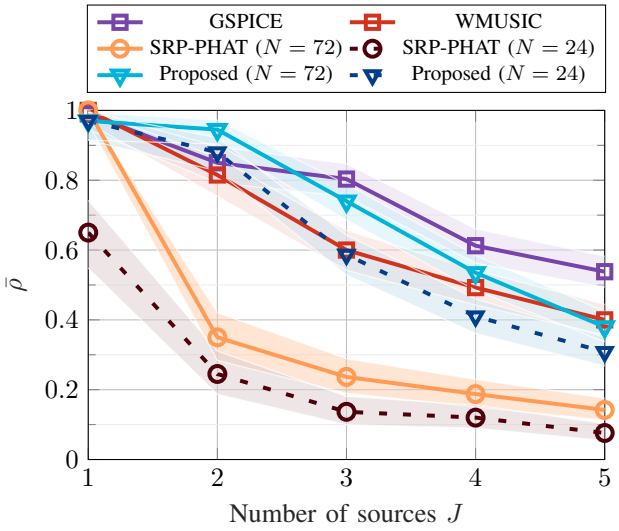


Fig. 9: Success rate obtained for different numbers of sources, with SNR $= 0\,\mathrm{dB}$ and $\mathrm{T}_{60} = 0\,\mathrm{s}$.

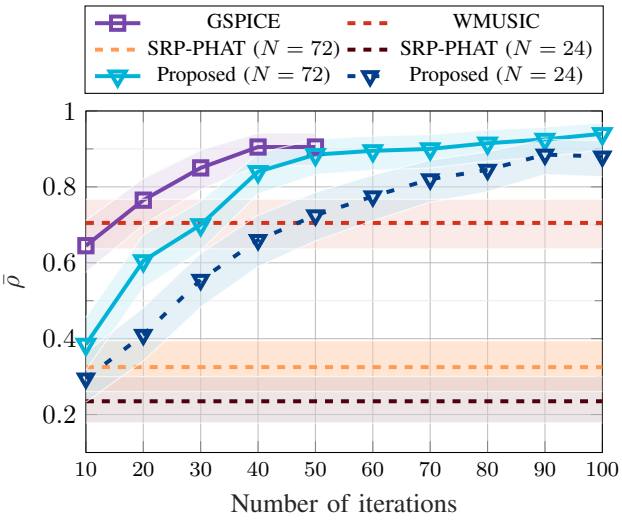


Fig. 10: Success rate obtained for different numbers of iterations, with SNR $= 0\,\mathrm{dB}$ and $\mathrm{T}_{60} = 0\,\mathrm{s}$.

sources increase, and under the anechoic conditions considered in this set of simulations, GSPICE presents overall the best DOA estimation performance in cases where more than 3 sources are present. Moreover, although the proposed method does not present an advantage over SRP-PHAT with $N = 72$ in the single-source case, it still shows an improvement in DOA estimation in the multi-source cases. Finally, when $N = 24$, the proposed method outperforms SRP-PHAT in all single and multi-source scenarios tested, presenting a trend in the success rate that is somewhat similar to WMUSIC.

H. Number of iterations

Despite the advantages in terms of DOA estimation performance of the proposed method, observed in difference scenarios tested and compared to SRP-PHAT, GSPICE and WMUSIC in the previous sections, it could be argued that the overall computational complexity resulting from the iterative procedure of the ADMM-based implementation of the proposed method may render its practical use challenging in certain applications. Therefore, in this section, we aim to investigate the number of iterations required by the proposed method to attain a comparable DOA estimation performance to SRP-PHAT, GSPICE and WMUSIC. This information could prove valuable in scenarios where setting a fixed number of iterations is feasible, potentially allowing for comparable DOA estimation performance even before reaching convergence.

Similarly to Sec. VI-A, we consider a scenario with two sources (S_1 and S_2), with an angular separation of $\Delta\vartheta_{2,1}^{(r)} \in [15°, 30°]$. The SNR is set to $0\,\mathrm{dB}$ and the simulations are performed under anechoic conditions ($\mathrm{T}_{60} = 0\,\mathrm{s}$). The number of averaged STFT cross-correlation frames corresponds to $N_f = 64$. As previously mentioned in Sec. VI-A, and given the scenario being considered, the proposed method required on average 90 iterations to reach convergence, whereas GSPICE required on average 40 iterations. In this set of simulations, however, the DOA estimation is performed while considering the estimated vector obtained after 10 to 100 iterations of the proposed method, and after 10 to 50 iterations of GSPICE.

The success rates obtained with the proposed method (for $N = 24$ and $N = 72$) and GSPICE for different numbers of iterations are presented in Fig. 10. For comparison, the success rates of SRP-PHAT and WMUSIC are represented by the dashed horizontal lines. We can observe that, for $N = 72$, the proposed method reaches a DOA estimation performance comparable to WMUSIC after 30 iterations, and comparable to GSPICE after 50 iterations. Moreover, we observe that for both $N = 72$ and $N = 24$, the proposed method outperforms SRP-PHAT after 20 iterations. These results indicate an opportunity to reduce the computational complexity of the proposed method by limiting the number of iterations computed within the ADMM scheme, while achieving a comparable performance to the other methods being tested.

I. Angular separation between two sources in a meeting room environment

In order to evaluate the DOA estimation performance of the compared methods in a realistic environment, experimental RIRs from the database detailed in [57] were used. Specifically, we employ RIRs measured in a meeting room (referred to as the *Medium Meeting Room*), with a reverberation time of $\mathrm{T}_{60} = 0.659\,\mathrm{s}$. A circular array of 60 loudspeakers and $1.5\,\mathrm{m}$ radius is placed approximately at the center of the room. We consider the case of a uniform circular array of $M = 6$ microphones and $0.1\,\mathrm{m}$ radius, placed at the center of the loudspeaker array (referred to as *Zone E*). The loudspeaker and microphone array setups are reported to be placed at a height of $1.45\,\mathrm{m}$ from the ground. In this setup, since the maximum inter-microphone distance corresponds to $d_{max} = 0.2\,\mathrm{m}$, the minimum required number of SRP samples corresponds to $N_{min} = 60$.

The DOA estimation performance is evaluated for the case of two sources (S_1 and S_2) of equal broadband power.

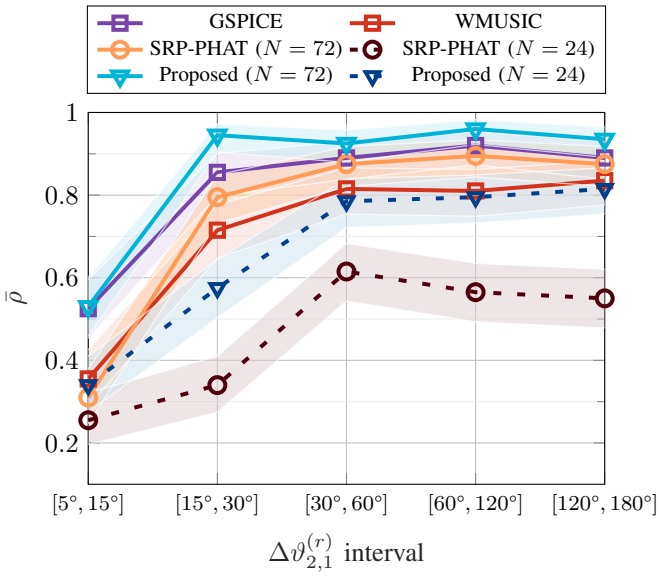


Fig. 11: Success rate obtained for different angular separations between sources in a meeting room environment, with SNR $=$ 0 dB and $T_{60} = 0.659$ s.

The angular separation between two neighboring loudspeakers from the circular array of 60 elements is approximately 6° [57], and we vary the angular separation between sources by selecting a random loudspeaker pair with an angular separation $\Delta\theta_{2,1}^{(r)}$ falling within one of the following intervals: $[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, and $[120°, 180°]$. The source signals are convolved with the measured RIRs, and diffuse babble noise [65], [70] is added to convey more realistic acoustic conditions, with SNR $= 0$ dB.

The success rates obtained with GSPICE, WMUSIC, SRP-PHAT and the proposed method are shown in Fig. 11. An improvement in DOA estimation performance can be observed for GSPICE, WMUSIC, SRP-PHAT ($N = 72$) and the proposed method ($N = 72$) when $\Delta\theta_{2,1}^{(r)} \in [15°, 30°]$ and in comparison to the results presented in Fig. 5. This improvement is likely due to the greater spacing between microphones from the circular array when compared to the scenario simulated in Sec. VI-C. However, we now observe a greater performance degradation for the proposed method with $N = 24$. This suggests that the higher spatial resolution of the circular array may induce more pronounced spatial aliasing effects when computing an SRP map with only $N = 24$ samples, thereby limiting the DOA estimation performance of the proposed method in this scenario. Nevertheless, with $N = 72$, we observe that the proposed method performs slightly better than the compared methods for angular separations above 15°, and that with $N = 24$, it performs similarly to WMUSIC for angular separations above 30°, which indicates the potential advantage of employing the proposed method in realistic scenarios, such as the meeting room environment under consideration.

VII. Conclusion

In this paper, we proposed a method for estimating the DOAs of multiple broadband sources by solving a group-sparse optimization problem that models a broadband SRP map as a linear function of PSDs, and where the grid resolution of the observed SRP map can be lower than the resolution of the DOA search grid used to model the candidate source PSDs. The estimation of the source DOAs is then accomplished by identifying peaks in the resulting spatial power density, i.e., the direction-specific PSDs integrated over frequency. An implementation using the alternating direction method of multipliers (ADMM) is presented, and the DOA estimation performance of the proposed method is evaluated with both simulated and experimental data. Results showed that, especially in reverberant scenarios, the proposed method presented an advantage in locating closely spaced sources when compared to the conventional SRP-PHAT, the group-sparse iterative covariance-based estimation (GSPICE) method, and the wideband MUSIC method with geometric averaging. Furthermore, it was observed that, for a compact microphone array, the proposed method overall maintained its performance even when using SRP maps computed with grid resolutions that are lower than the sampling requirements of the broadband SRP function. This observation, along with results obtained for limited numbers of iterations within the ADMM-scheme, indicated potential approaches to reduce the proposed method's overall computational complexity while maintaining its DOA estimation performance. Finally, results obtained with experimental data showed the validity and applicability of the proposed method in a practical meeting room environment.

Appendix
GSPICE applied to microphone cross-correlations

Let $\hat{\boldsymbol{\psi}} \in \mathbb{C}^{KP}$ denote the vector of cross-correlations estimated from the microphone array's signals, which is here modeled as

$$\hat{\boldsymbol{\psi}} = \mathbf{A}\boldsymbol{\varphi}_S + \mathbf{e}, \tag{56}$$

where $\mathbf{A}$ and $\boldsymbol{\varphi}_S$ are defined as in Sec. IV, and $\mathbf{e} \in \mathbb{C}^{KP}$ is the noise component in $\hat{\boldsymbol{\psi}}$. A covariance matrix of the measured cross-correlations is modeled as

$$\mathbf{R} = \tilde{\mathbf{A}}\mathbf{P}\tilde{\mathbf{A}}^H \in \mathbb{R}^{KP\times KP}, \tag{57}$$

where $\tilde{\mathbf{A}}$ is an augmented version of $\mathbf{A}$, defined as $\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{I}_{KP}] \in \mathbb{C}^{KP\times KL+KP}$. The matrix $\mathbf{P}$ is defined as $\mathbf{P} = \mathrm{diag}(\mathbf{p})$, with $\mathbf{p}$ denoting the vector of unknown diagonal loadings within the covariance model used:

$$\mathbf{p} = [\mathbf{p}_1^\top, \ldots, \mathbf{p}_L^\top, \sigma_1, \ldots, \sigma_{KP}]^\top, \tag{58}$$

with $\mathbf{p}_l$ corresponding to the l-th group's diagonal loadings vector:

$$\mathbf{p}_l = [p_{1,l}, \ldots, p_{K,l}]^\top, \tag{59}$$

and $\sigma_1, \ldots, \sigma_{KP}$ corresponding to the noise power for each frequency and microphone pair, represented as a total of KP single-element groups.

In the GSPICE method, it is first proposed to solve the following optimization problem [55]:

$$\begin{aligned} \underset{\mathbf{p}}{\text{minimize}} \;\; & \hat{\boldsymbol{\psi}}^H \mathbf{R}^{-1} \hat{\boldsymbol{\psi}} + \sum_{l=1}^{L+KP} \nu_l \left\| \mathbf{p}_l \right\|_p \\ \text{subject to} \;\; & \mathbf{R} = \tilde{\mathbf{A}} \mathbf{P} \tilde{\mathbf{A}}^H \\ & \mathbf{p} \geq 0 \end{aligned} \tag{60}$$

where ν_l is

$$\nu_l = \left\| \left[\|\tilde{\boldsymbol{a}}_{(l-1)K+1}\|_2^2 \; \ldots \; \|\tilde{\boldsymbol{a}}_{lK}\|_2^2 \right]^\top \right\|_2, \tag{61}$$

for $l = 1, \ldots, L$, with $\tilde{\boldsymbol{a}}_q$ denoting the q-th column of $\tilde{\mathbf{A}}$. The objective function in (60) corresponds to a relaxed version of the conventional covariance fitting criteria, depending on the choice of the l_p-norm [55], and by introducing the auxiliary variable $\boldsymbol{\beta} = [\boldsymbol{\varphi}_S^\top, \mathbf{e}^\top]^\top$, an equivalent formulation of the optimization problem can be obtained as:

$$\begin{aligned} \underset{\mathbf{p}, \boldsymbol{\beta}}{\text{minimize}} \;\; & \boldsymbol{\beta}^H \mathbf{P}^{-1} \boldsymbol{\beta} + \sum_{l=1}^{L+KP} \nu_l \left\| \mathbf{p}_l \right\|_p \\ \text{subject to} \;\; & \tilde{\mathbf{A}} \boldsymbol{\beta} = \hat{\boldsymbol{\psi}} \\ & \mathbf{p} \geq 0 \end{aligned} \tag{62}$$

which, when implemented, can be split into two convex subproblems that are iteratively solved for $\mathbf{p}$ and $\boldsymbol{\beta}$. It is shown in [55] that in the case of homoscedastic noise, i.e., the noise component presents equivariance power, another equivalent formulation of the optimization problem (proof omitted) can be obtained as:

$$\underset{\boldsymbol{\varphi}_S}{\text{minimize}} \;\; \|\hat{\boldsymbol{\psi}} - \mathbf{A} \boldsymbol{\varphi}_S\|_2 + \sum_{l=1}^{L} \sqrt{\frac{\nu_l}{KP}} \left\| \boldsymbol{\varphi}_{S,l} \right\|_{\frac{2p}{p+1}} \tag{63}$$

which, when $p \to \infty$, corresponds to the group square-root Lasso problem *without* a nonnegativity constraint.

REFERENCES

[1] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[2] F. Elvander, R. Ali, A. Jakobsson, and T. van Waterschoot, "Offline noise reduction using optimal mass transport induced covariance interpolation," in *Proc. 27th European Signal Process. Conf. (EUSIPCO 2019)*, A Coruna, Spain, 2019, pp. 1–5.

[3] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, Prague, Czech Republic, 2011, pp. 2592–2595.

[4] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimed. Tools. Appl.*, vol. 75, pp. 10 407–10 439, 2016.

[5] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Rob. Auton. Syst.*, vol. 96, pp. 184–210, 2017.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Berlin, Heidelberg: Springer, 2008, vol. 1.

[7] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[8] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 10, pp. 1481–1494, 1989.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[10] R. Roy and T. Kailath, "ESPRIT - Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[11] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. '83 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '83)*, vol. 8, Boston, MA, USA, 1983, pp. 336–339.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[13] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*. Berlin, Heidelberg: Springer, 2001, pp. 157–180.

[14] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence, RI, USA, 2000.

[15] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. 2017 IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 136–140.

[16] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, Jul. 2022.

[17] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[18] S. Fortunati, R. Grasso, F. Gini, M. S. Greco, and K. LePage, "Single-snapshot DOA estimation by using compressed sensing," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, p. 120, Dec. 2014.

[19] Y. Park, F. Meyer, and P. Gerstoft, "Sequential sparse Bayesian learning for time-varying direction of arrival," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 2089–2099, Mar. 2021.

[20] Z. Bai, L. Shi, J. R. Jensen, J. Sun, and M. G. Christensen, "Acoustic DOA estimation using space alternating sparse Bayesian learning," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 14, Dec. 2021.

[21] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion," *Signal Process.*, vol. 171, pp. 1–19, 2020.

[22] ——, "Tracking and sensor fusion in direction of arrival estimation using optimal mass transport," in *Proc. 26th European Signal Process. Conf. (EUSIPCO 2018)*, Rome, Italy, 2018, pp. 1617–1621.

[23] ——, "Non-coherent sensor fusion via entropy regularized optimal mass transport," in *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2019)*, Brighton, UK, 2019, pp. 4415–4419.

[24] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[25] A. Johansson, G. Cook, and S. Nordholm, "Acoustic direction of arrival estimation, a comparison between root-MUSIC and SRP-PHAT," in *Proc. 2004 IEEE Region 10 Conf. (TENCON 2004)*. Chiang Mai, Thailand: IEEE, 2004, pp. 629–632.

[26] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *Proc. 2007 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2007)*. New Paltz, NY, USA: IEEE, 2007, pp. 18–21.

[27] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer, 2010, pp. 307–337.

[28] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

[29] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Process.*, vol. 87, no. 7, pp. 1677–1691, 2007.

[30] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array," in *Proc. 2003 Int. Workshop Acoustic Echo Noise Control (IWAENC 2003)*, Seattle, WA, USA, 2003, pp. 315–318.

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

[32] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, pp. 1–17, 2010.

[33] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2011)*, 2011, pp. 92–97.

[34] D. Diaz-Guerra and J. R. Beltran, "Source cancellation in cross-correlation functions for broadband multisource DOA estimation," *Signal Process.*, vol. 170, p. 107442, 2020.

[35] Y. Oualil, F. Faubel, and D. Klakow, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Proc. 2012 Workshop on Statistical and Perceptual Audition (SAPA-SCALE 2012)*, Portland, OR, USA, 2012, pp. 68–73.

[36] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2010)*, Dallas, TX, USA, 2010, pp. 125–128.

[37] ——, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, 2008, pp. 301–304.

[38] Y. Oualil, F. Faubel, and D. Klakow, "A fast cumulative steered response power for multiple speaker detection and localization," in *Proc. 21st European Signal Process. Conf. (EUSIPCO 2013)*, Marrakech, Morocco, 2013, pp. 1–5.

[39] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.

[40] E. Hadad and S. Gannot, "Multi-Speaker Direction of Arrival Estimation using SRP-PHAT Algorithm with a Weighted Histogram," in *Proc. 2018 IEEE Int. Conf. Science Electrical Eng. Israel (ICSEE)*, Eilat, Israel, 2018, pp. 1–5.

[41] S. Thakallapalli, S. V. Gangashetty, and N. Madhu, "NMF-weighted SRP for multi-speaker direction of arrival estimation: Robustness to spatial aliasing while exploiting sparsity in the atom-time domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 13, 2021.

[42] Z. Chu, J. Weng, and Y. Yang, "Determination of propagation model matrix in generalized cross-correlation based inverse model for broadband acoustic source localization," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2098–2109, 2020.

[43] T. Padois, O. Doutres, F. Sgard, and A. Berry, "Time domain localization technique with sparsity constraint for imaging acoustic sources," *Mech. Syst. Signal Process.*, vol. 94, pp. 85–93, 2017.

[44] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, no. 10, pp. 13 781–13 812, 2012.

[45] Q. Shen, W. Liu, W. Cui, S. Wu, Y. D. Zhang, and M. G. Amin, "Low-complexity direction-of-arrival estimation based on wideband co-prime arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1445–1456, 2015.

[46] E. Tengan, M. Taseska, T. Dietzen, and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone," in *Proc. 29th European Signal Process. Conf. (EUSIPCO 2021)*, Dublin, Ireland, 2021, pp. 221–225.

[47] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 38, pp. 1–20, 2023.

[48] ——, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered response power maps," in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2023)*, New Paltz, NY, USA, 2023, pp. 1–5.

[49] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[50] G. García-Barrios, J. M. Gutiérrez-Arriola, N. Sáenz-Lechón, V. J. Osma-Ruiz, and R. Fraile, "Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps," *IEEE Access*, vol. 9, pp. 121 549–121 560, 2021.

[51] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, 2010.

[52] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Gonçalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.

[53] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 586–601, 2017.

[54] F. Bunea, J. Lederer, and Y. She, "The group square-root Lasso: Theoretical properties and fast algorithms," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1313–1325, 2014.

[55] T. Kronvall, S. I. Adalbjörnsson, S. Nadig, and A. Jakobsson, "Group-sparse regression using the covariance fitting criterion," *Signal Process.*, vol. 139, pp. 116–130, 2017.

[56] S. Boyd, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[57] S. Zhao, Q. Zhu, E. Cheng, and I. S. Burnett, "A room impulse response database for multizone sound field reproduction (L)," *J. Acoust. Soc. Am.*, vol. 152, no. 4, pp. 2505–2512, Oct. 2022.

[58] M. R. Azimi-Sadjadi, A. Pezeshki, and N. Roseveare, "Wideband doa estimation algorithms for multiple moving sources using unattended acoustic sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, no. 4, pp. 1585–1599, 2008.

[59] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on Nyquist-Shannon sampling," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2021)*, New Paltz, NY, USA, 2021, pp. 206–210.

[60] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, 2009.

[61] R. J. I. Marks, *Introduction to Shannon sampling and interpolation theory*. Springer, 2012.

[62] T. Kronvall and A. Jakobsson, "Hyperparameter selection for group-sparse regression: A probabilistic approach," *Signal Process.*, vol. 151, pp. 107–118, 2018.

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

[64] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 774–786, 2015.

[65] T. Dietzen, R. Ali, M. Taseska, and T. van Waterschoot, "Myriad: a multi-array room acoustic database," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 1, pp. 1–17, 2023.

[66] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, Australia, 2015, pp. 5206–5210.

[67] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[68] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: New DOA estimator for wideband signals," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1977–1989, 2006.

[69] J. Reiczigel, "Confidence intervals for the binomial parameter: some new considerations," *Stat. Med.*, vol. 22, no. 4, pp. 611–621, 2003.

[70] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.