MULTI-SOURCE DIRECTION-OF-ARRIVAL ESTIMATION USING GROUP-SPARSE FITTING OF STEERED RESPONSE POWER MAPS

Elisa Tengan,[1] *Thomas Dietzen,*[1] *Filip Elvander,*[2] *Toon van Waterschoot*[1]

[1] Department of Electrical Engineering (ESAT/STADIUS), KU Leuven, Leuven, Belgium
[2] Department of Information and Communications Engineering, Aalto University, Espoo, Finland

ABSTRACT

In this paper, a method is proposed for estimating the direction of arrival (DOA) of multiple broadband sound sources by solving a group-sparse optimization problem. A steered response power (SRP) map is modeled using power spectral densities (PSDs) defined on an overcomplete grid of candidate DOAs. The source DOAs are then estimated as the directions corresponding to the largest peaks of the frequency-averaged PSDs. The proposed optimization problem is iteratively solved using the alternating direction method of multipliers (ADMM), and simulation results show that the proposed method overall outperforms the frequency-domain sparse iterative covariance-based estimation (SPICE) method and performs better than or similar to the conventional SRP-PHAT method for varying levels of noise and reverberation.

***Index Terms*—** source localization, steered response power, group-sparsity

1. INTRODUCTION

In microphone array processing, the steered response power (SRP) is a widely used beamforming-based method for acoustic source localization, with one of its main advantages being the robustness to adverse acoustic conditions, such as high levels of noise and reverberation [1–3]. In addition, applying the phase transform (PHAT) weighting function when computing the generalized cross-correlation (GCC) [4] values used in the conventional SRP formulation, yielding the well-known SRP-PHAT framework [1], has particularly shown to improve performance in the localization of speech sources in reverberant environments [1–3, 5, 6].

In the case of a single sound source in the far field, estimating its direction of arrival (DOA) using the conventional SRP method is performed by finding the direction that yields the largest peak in the SRP map computed for a given set of candidate directions [3]. However, when multiple sources are simultaneously active in the acoustic scene, estimating their DOAs using conventional SRP becomes more challenging, especially in noisy and reverberant environments, where the presence of spurious peaks and the inability to resolve closely located sources may limit the DOA estimation performance [7, 8]. Moreover, using SRP-PHAT when sources have different spectral characteristics can have undesired effects on the obtained SRP map, resulting in degraded localization performance [6].

Different approaches have been previously proposed to improve the localization of multiple sources under an SRP framework, with examples including the use of postprocessing steps [9–12], clustering techniques [13–15] and sparsity-based modeling of SRP maps [16]. In the latter case, the observed broadband SRP map is modeled as a linear combination of basis functions describing ideal SRP maps, weighted by coefficients associated to each candidate source location. By solving a linear inverse problem with an l_1-norm penalty, the resulting sparse weight vector is then used to compute a new SRP map from which the source locations' estimates can be obtained [16].

Alternatively, in this paper, we propose a method for DOA estimation of multiple sources based on a group-sparse modeling framework. Group-sparse optimization has been previously used in DOA estimation of broadband sources with co-prime arrays [17], as well as with a single directional microphone [18]. Here, it is employed to model a broadband SRP map using direction-dependent power spectral densities (PSDs). In particular, the observed SRP map is modeled as a linear combination of (unknown) source PSDs, supported on an overcomplete grid of candidate DOAs. The PSDs for all candidate DOAs are concatenated and compose a vector to be estimated, in which each element is frequency and direction-dependent. As only a limited number of broadband sources are assumed to be present, the vector of concatenated source PSDs is expected to have a group-sparse structure, where each group corresponds to a source PSD. Based on this rationale, it is herein proposed to estimate DOAs by first solving a linear inverse problem augmented by a group-sparsity penalty that fits the observed SRP map to the proposed model, and then identifying the directions corresponding to the largest peaks of the resulting group-sparse PSD vector, averaged over frequency. Hence, the proposed method allows for estimating DOAs without requiring the computation of a new SRP map, while also accounting for frequency-dependent contributions from multiple sources.

We evaluate the DOA estimation performance of the proposed method through simulations with two speech sources and under different acoustic conditions. The results are compared to those obtained with the conventional SRP-PHAT method and the frequency-domain sparse iterative covariance-based estimation (SPICE) algorithm [19, 20], which, similarly to [16], is also based on sparse model fitting, but employs the observed microphone data instead of a computed SRP map. It is shown that the proposed method overall outperforms the SPICE method and performs better than or similar to the SRP-PHAT method for varying levels of noise and reverberation.

This paper is organized as follows. In Sec. 2, the signal model and the broadband SRP formulation are defined. In Sec. 3, the pro-

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of FWO Mandate SB 1S86520N. The research leading to these results has also 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.

posed method is presented. In Sec. 4, the simulation setups and their implementation are detailed, with results discussed in Sec. 5. Finally, a summary and final remarks conclude the paper in Sec. 6.

2. SIGNAL MODEL AND BROADBAND STEERED RESPONSE POWER (SRP) METHOD

2.1. Signal model

In this work, we focus on scenarios where sound sources are located in the far field, on the same horizontal plane as a planar microphone array with M elements, indexed by $m = 1, \ldots, M$. In this context, the time difference of arrival (TDOA) of a propagating sound wave arriving from an azimuth angle θ that is observed at a microphone pair (m, m') is given by [2]

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

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, $(\cdot)^\top$ denotes the transpose, $\mathbf{r}_\theta$ denotes the incident direction vector of unit length for azimuth angle θ, and c is the speed of sound. The azimuth angle of a sound source will also be referred to as its DOA in the remainder.

In the discrete Fourier transform (DFT) domain, while assuming anechoic conditions and considering a total of J sound sources located at DOAs $\theta_{S_1}, \ldots, \theta_{S_J}$, the m-th microphone signal denoted by $y_m(\omega_k)$ is modeled as

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

with ω_k denoting the discrete angular frequency, $(\cdot)^H$ the conjugate transpose and $n_m(\omega_k)$ the noise component assumed to be spatially white, i.e., uncorrelated for different microphones. The vector $\mathbf{s}(\omega_k) \in \mathbb{C}^J$ corresponds to the source signal vector at the array's first microphone, 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 source steering vector relative to the array's first microphone:

$$\mathbf{w}_m(\omega_k) = \left[e^{j\omega_k \Delta t_{m,1}(\theta_{S_1})} \ \ldots \ e^{j\omega_k \Delta t_{m,1}(\theta_{S_J})}\right]^\top. \tag{4}$$

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

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

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator and $(\cdot)^*$ the complex conjugate. 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's signal. Furthermore, from (1), one can note that $\Delta t_{m,m'}(\theta) = \Delta t_{m,1}(\theta) - \Delta t_{m',1}(\theta)$. Thus, the cross-correlation $\psi_{m,m'}(\omega_k)$ can be expressed as

$$\begin{aligned} \psi_{m,m'}(\omega_k) = &\sum_{i=1}^{J} \phi_{S_i}(\omega_k) e^{-j\omega_k \Delta t_{m,m'}(\theta_{S_i})} \\ &+ \sigma_{n_m}^2(\omega_k)\delta(m - m'), \end{aligned} \tag{7}$$

where $\sigma_{n_m}^2(\omega_k)$ denotes the noise PSD at microphone m and $\delta(\cdot)$ denotes the Kronecker delta function.

2.2. Broadband steered response power (SRP) formulation

For conciseness, in the remainder, a microphone pair (m, m') will be denoted by index p, such that a given TDOA $\Delta t_{m,m'}(\theta)$ will be denoted by $\Delta t_p(\theta)$. The total number of different microphone pairs for the M-element array is given by $P = M(M-1)/2$.

Since $\psi_p(\omega_k)$ is complex symmetric, we consider its single-sided form, with $k = 0, \ldots, K$. Assuming the DC component to be negligible ($\psi_p(\omega_0) = 0$), the broadband SRP for a given candidate DOA angle θ_l is defined as [21]

$$z(\theta_l) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} \Re\{\psi_p(\omega_k) e^{j\omega_k \Delta t_p(\theta_l)}\}, \tag{8}$$

where $\Re\{\cdot\}$ denotes the real part. The exponential term in (8) 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 θ_l [1]. Considering a total of L candidate DOAs, denoted by $\theta_1, \ldots, \theta_L$, we define the broadband SRP vector $\mathbf{z} \in \mathbb{R}^L$ as

$$\mathbf{z} = [z(\theta_1) \ \ldots \ z(\theta_L)]^\top. \tag{9}$$

The vector $\mathbf{z}$ can be expressed as a matrix-vector multiplication as follows. We define, for a microphone pair p, the matrix $\mathbf{H}_p \in \mathbb{C}^{L \times K}$ and vector $\boldsymbol{\psi}_p \in \mathbb{R}^K$ containing the steering vectors for all candidate DOAs and the cross-correlations, respectively, as

$$[\mathbf{H}_p]_{l,k} = e^{j\omega_k \Delta t_p(\theta_l)}, \tag{10}$$

$$\left[\boldsymbol{\psi}_p\right]_k = \psi_p(\omega_k), \tag{11}$$

where $[\mathbf{H}_p]_{l,k}$ denotes the element of matrix $\mathbf{H}_p$ at row l and column k, and $\left[\boldsymbol{\psi}_p\right]_k$ denotes the k-th element of vector $\boldsymbol{\psi}_p$, for $k = 1, \ldots, K$. By stacking multiple matrices $\mathbf{H}_p$ and vectors $\boldsymbol{\psi}_p$, respectively, as

$$\mathbf{H} = [\mathbf{H}_1 \ \ldots \ \mathbf{H}_P], \tag{12}$$

$$\boldsymbol{\psi} = \left[\boldsymbol{\psi}_1^\top \ \ldots \ \boldsymbol{\psi}_P^\top\right]^\top, \tag{13}$$

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

$$\mathbf{z} = 2\Re\{\mathbf{H}\boldsymbol{\psi}\}. \tag{14}$$

In the conventional SRP method, the estimation of a single source's DOA is usually performed by identifying the direction index θ_l corresponding to the maximal element of the broadband SRP map $\mathbf{z}$ computed for a predefined set of candidate DOAs $\{\theta_1, \ldots, \theta_L\}$. However, when considering the case of multiple sources, especially under adverse acoustic conditions, this approach can be subject to DOA estimation errors caused by the presence of

spurious peaks in the 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 method.

3. PROPOSED METHOD

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

$$\boldsymbol{\varphi}_{S,l} = [\phi_S(\omega_1, \theta_l) \ \ \dots \ \ \phi_S(\omega_K, \theta_l)]^\top, \tag{15}$$

where $\phi_S(\omega_k, \theta_l)$ is the source PSD value at the l-th candidate direction θ_l for frequency ω_k. We stack different $\boldsymbol{\varphi}_{S,l}$ vectors for all L candidate source directions in order to obtain the overcomplete vector $\boldsymbol{\varphi}_S \in \mathbb{R}^{KL}$ of direction-dependent source PSDs as

$$\boldsymbol{\varphi}_S = \left[\boldsymbol{\varphi}_{S,1}^\top \ \ \dots \ \ \boldsymbol{\varphi}_{S,L}^\top\right]^\top. \tag{16}$$

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

$$\mathbf{a}_{p,l} = \left[e^{-j\omega_1 \Delta t_p(\theta_l)} \ \dots \ e^{-j\omega_K \Delta t_p(\theta_l)}\right]^\top, \tag{17}$$

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

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \dots & \mathbf{A}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{P,1} & \dots & \mathbf{A}_{P,L} \end{bmatrix}, \tag{18}$$

where $\mathbf{A}_{p,l} = \operatorname{diag}(\mathbf{a}_{p,l})$ and $\operatorname{diag}(\cdot)$ denotes the creation of a diagonal matrix from the elements of a given vector. Based on (7), we model the cross-correlation vector $\boldsymbol{\psi}$ in (13) as

$$\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\varphi}_S, \tag{19}$$

so that the broadband SRP vector in (14) can be expressed as

$$\mathbf{z} = 2\Re\{\mathbf{H}\mathbf{A}\boldsymbol{\varphi}_S\} = 2\Re\{\mathbf{H}\mathbf{A}\}\boldsymbol{\varphi}_S. \tag{20}$$

The second equality in (20) is possible due to the source PSD vector $\boldsymbol{\varphi}_S$ being inherently real-valued [22]. For conciseness, we define the matrix $\mathbf{B} \in \mathbb{R}^{L \times KL}$ as

$$\mathbf{B} = 2\Re\{\mathbf{H}\mathbf{A}\}, \tag{21}$$

so that the following optimization problem is proposed to be solved:

$$\underset{\boldsymbol{\varphi}_S}{\text{minimize}} \qquad \frac{1}{2}\left\|\mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S\right\|_2^2 + \lambda \left\|\boldsymbol{\varphi}_S\right\|_{2,1} \tag{22}$$

$$\text{subject to} \qquad \boldsymbol{\varphi}_S \geq 0 \tag{23}$$

where $\|\cdot\|_2$ and $\|\cdot\|_{2,1}$ denote the l_2-norm and mixed $l_{2,1}$-norm, respectively, with $\|\boldsymbol{\varphi}_S\|_{2,1} \triangleq \sum_{l=1}^{L} \left\|\boldsymbol{\varphi}_{S,l}\right\|_2$. The constraint in (23) is employed to comply with the intrinsic nonnegativity property of PSD values [22], and the regularization term $\lambda \|\boldsymbol{\varphi}_S\|_{2,1}$ promotes group-sparsity when solving the proposed optimization problem [23], with λ as a regularization parameter. Since it is assumed that only a limited number of sound sources are present (i.e., $J \ll L$), and consequently, only a few of the subvectors $\boldsymbol{\varphi}_{S,l}$ composing $\boldsymbol{\varphi}_S$ should be nonzero, solving the proposed optimization problem allows for estimating $\hat{\boldsymbol{\varphi}}_S$, and consequently, $\hat{\boldsymbol{\varphi}}_{S,1}, \dots, \hat{\boldsymbol{\varphi}}_{S,L}$, such that the source signals' DOAs can be estimated by averaging PSD values over frequency for each subvector $\hat{\boldsymbol{\varphi}}_{S,l}$ and identifying their peak values over all L candidate directions considered [18]. For a total of J sources, the estimated DOAs $\hat{\theta}_{S_1}, \dots, \hat{\theta}_{S_J}$ are then chosen to be the candidate directions related to the J largest peaks.

4. SIMULATIONS

In order to evaluate the DOA estimation performance of the proposed method for multiple sources, simulations are carried out for different values of the noise power and the reverberation time, with a total of 150 Monte Carlo realizations considered for each specific scenario. We consider the use of a uniform circular array of $5\,\mathrm{cm}$ radius and $M = 6$ microphones, yielding $P = 15$ microphone pairs, placed on a horizontal plane inside a room with dimensions $6.3 \times 5.1 \times 2.9\,\mathrm{m}$, with the array centered at Cartesian coordinates $[3.1, 2.7, 1.5]^\top\,\mathrm{m}$. Two sound sources ($J = 2$) are placed at the same height as the microphone array and at a radial distance of $1.5\,\mathrm{m}$ from the center of the array. For each Monte Carlo realization, the source DOAs θ_{S_1} and θ_{S_2} are randomly generated over the range of 360° without constraining them to lie on the candidate direction grid. The room reverberation is simulated for different reverberation times [24], denoted by T_{60}, using the image source method as implemented in [25].

With a sampling rate of $16\,\mathrm{kHz}$, the first and second sound sources' signals correspond to randomly selected female and male speech segments [26], respectively. The broadband power level difference between the two speech sources is varied between $0\,\mathrm{dB}$, $3\,\mathrm{dB}$, and $6\,\mathrm{dB}$. The spatially white noise component is generated as speech-shaped noise and the signal-to-noise ratio (SNR) is defined as $\mathrm{SNR} = \sigma_{S_1}^2/\sigma_N^2$, where $\sigma_{S_1}^2$ and σ_N^2 denote the broadband power of the first source's reverberant signal and of the noise component, respectively, at the array's first microphone. The time-domain signals have a duration of $3\,\mathrm{s}$, and the cross-correlations in (5) are obtained by computing the short-time Fourier transform (STFT) over the microphone signals and averaging the PHAT-weighted STFT cross-correlations over all frames [1]. The STFT is implemented using a 512-point square-root Hann window with 50% overlap, resulting in a time-domain window with length of $32\,\mathrm{ms}$ and a total of 187 frames. Considering the single-sided broadband SRP formulation in (8), a total of $K = 256$ frequency bins are used.

The grid of candidate DOAs is defined in the same horizontal plane as the microphone array, such that the direction vector for a given candidate DOA θ_l is $\mathbf{r}_{\theta_l} = -[\cos\theta_l, \sin\theta_l, 0]^\top$ [2]. We consider a total of $L = 72$ candidate DOAs uniformly distributed over 360°, resulting in a 5°-resolution grid. The optimization problem in (22)-(23) is iteratively solved using the alternating direction method of multipliers (ADMM) [27], and the regularization parameter is heuristically set to $\lambda = 0.05 \max_{\ell \in \{1,\dots,L\}} \|\mathbf{B}_l^H \mathbf{z}\|_2$, where $\mathbf{B}_l$ denotes the set of K columns from matrix $\mathbf{B}$, defined in (21), associated to candidate DOA θ_l. The proposed method's DOA estimation performance is compared to that of the conventional SRP-

[1] Since the weighting operation is not accounted for in the cross-correlation model presented in Sec. 2, 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.

PHAT, as well as of the frequency-domain SPICE method [19, 20] (abbreviated as FD SPICE), which is here implemented by averaging the estimated narrowband output maps over all frequency bins considered. For all methods, the DOAs are estimated by finding the $J = 2$ largest local maxima on the corresponding maps obtained. The DOA estimation is evaluated for each method in terms of mean absolute error (MAE) averaged over both sources and all Monte Carlo realizations for each scenario considered, and is computed as

$$\text{MAE} = \frac{1}{2}\frac{1}{N_r}\sum_{i=1}^{2}\sum_{r=1}^{N_r}|\theta_{S_i}^{(r)} - \hat{\theta}_{S_i}^{(r)}|, \tag{24}$$

where N_r denotes the total number of Monte Carlo realizations and the superscript (r) denotes the realization index.

5. RESULTS AND DISCUSSION

Firstly, in order to illustrate the ability of each method considered in this study to locate multiple sources, an example of the conventional SRP-PHAT map, as well as the frequency-averaged output maps obtained with FD SPICE and the proposed method, are presented in Fig. 1, for a case where sources are off the grid, located at 77° and 97°, with $\text{T}_{60} = 0.2\,\text{s}$ and $\text{SNR} = -6\,\text{dB}$. For ease of interpretation, the peaks selected as the estimated sources' DOAs are indicated by the black circles (◯), the true DOAs are indicated by the dashed vertical lines (- - -), each method's estimated map is normalized by its maximum and the range of candidate directions shown is limited between 0° and 200°. We can observe that, in this case, the sources are too closely located to result in two separate main peaks on the conventional SRP-PHAT map and the FD SPICE frequency-averaged output, such that a secondary, spurious peak is selected as one of the estimated sources' DOA. In contrast, the output map of the proposed method has its two largest peaks located at the closest grid points to both sources' true DOAs.

The MAE obtained with each method for a set of simulations where the reverberation time is fixed at $\text{T}_{60} = 0.2\,\text{s}$ and the SNR is varied from $-24\,\text{dB}$ to $18\,\text{dB}$ is presented in Fig. 2. It can be observed that for $\text{SNR} \geq -18\,\text{dB}$, the proposed method presents DOA estimation errors that are consistently lower than those for SRP-PHAT and FD SPICE methods, indicating the benefit of the proposed approach in the case of more than one source.

Additionally, in Fig. 3, the MAE obtained with each method is presented for a set of simulations where the SNR is fixed at $-6\,\text{dB}$ and the reverberation time T_{60} is varied from $0\,\text{s}$ to $0.8\,\text{s}$, with $\text{T}_{60} = 0\,\text{s}$ corresponding to the anechoic case. It can be observed that the proposed method yields lower MAE than the FD SPICE method for all reverberation times considered, and that for $\text{T}_{60} \leq 0.6\,\text{s}$, it also performs better than or similar to SRP-PHAT.

Since the regularization parameter λ present in (22) is here heuristically set as explained in Sec. 4 and depends on setup conditions, such as the noise level and the number of target sources [28], its chosen value may be suboptimal in the simulations considered. Further tuning could be employed, for example, to investigate the proposed method's performance under higher levels of reverberation. Finally, one may note that although the proposed method currently presents a higher complexity level than SRP-PHAT, it achieves the main goal of performance improvement over both SRP-PHAT and FD SPICE in most scenarios tested.

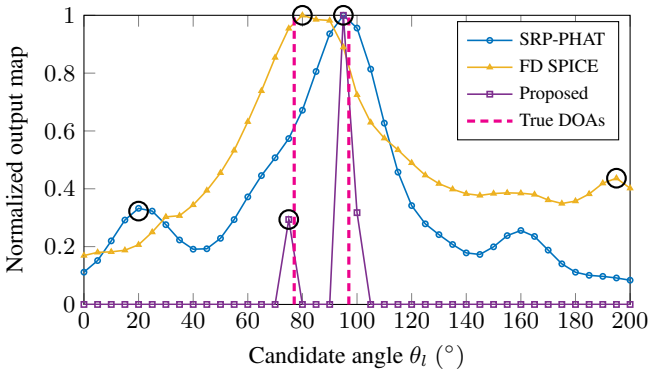

Figure 1: Normalized output maps obtained for different candidate directions in the case of two sources located at 77° and 97°, with $\text{SNR} = -6\,\text{dB}$ and $\text{T}_{60} = 0.2\,\text{s}$.

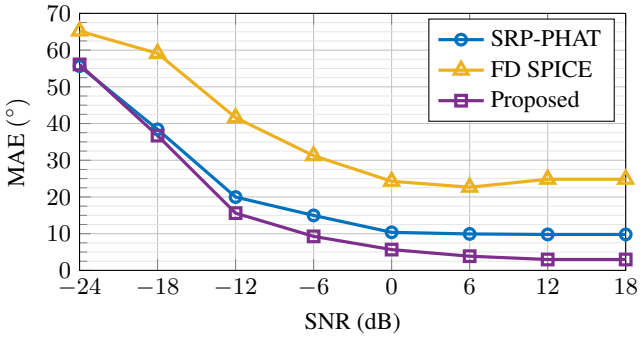

Figure 2: MAE vs. SNR, with $\text{T}_{60} = 0.2\,\text{s}$.

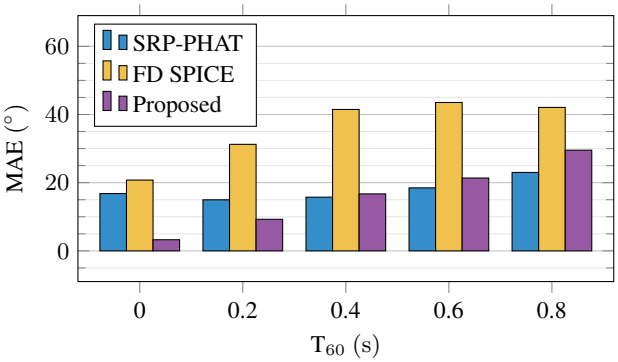

Figure 3: MAE vs. T_{60}, with $\text{SNR} = -6\,\text{dB}$.

6. CONCLUSION

In this paper, a method for estimating the direction of arrival (DOA) of multiple broadband sound sources is proposed, in which a group-sparse optimization problem modeling a broadband SRP map as a linear combination of power spectral densities (PSDs) is solved. Simulation results show that the proposed method overall outperforms the frequency-domain SPICE approach and performs better than or similar to the conventional SRP-PHAT for varying levels of noise and reverberation. Future work includes studying the use of subsampled SRP maps.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] 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, 1976.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, 2008, pp. 4349–4352.

[10] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. 2004 IEEE Int. Conf. Robotics Automation (ICRA 2004)*, vol. 1, New Orleans, LA, USA, 2004, pp. 1033–1038.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] S. Khanal and H. F. Silverman, "Multi-stage rejection sampling (MSRS): A robust SRP-PHAT peak detection algorithm for localization of cocktail-party talkers," in *Proc. 2015 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2015)*, New Paltz, NY, USA, 2015, pp. 1–5.

[16] 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.

[17] 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.

[18] 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. 2021 29th European Signal Process. Conf. (EUSIPCO 2021)*, Dublin, Ireland, 2021, pp. 221–225.

[19] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, 2011.

[20] H. Park and J. Li, "A frequency-domain SPICE approach to high-resolution time delay estimation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 360–363, 2018.

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

[22] B. Porat, *A course in digital signal processing*. New York: John Wiley & Sons, Inc., 1996.

[23] 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.

[24] H. Kuttruff and E. Mommertz, "Room acoustics," in *Handbook of engineering acoustics*. Springer, 2012, pp. 239–267.

[25] 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.

[26] 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: IEEE, 2015, pp. 5206–5210.

[27] 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.

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