

SOURCE LOCALIZATION BY MULTIDIMENSIONAL STEERED RESPONSE POWER MAPPING WITH SPARSE BAYESIAN LEARNING

Wei-Ting Lai, Lachlan Birnie, Xingyu Chen, Amy Bastine, Thushara D. Abhayapala, Prasanga N. Samarasinghe

Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia

ABSTRACT

We propose a method that combines Steered Response Power (SRP) with sparse optimization for localizing multiple sources. While conventional SRP is robust under adverse conditions, it struggles with scenarios involving neighboring sources, often resulting in ambiguous SRP maps. The current state-of-the-art approach optimizes observed SRP maps through group-sparse modeling, but its performance degrades in reverberant scenarios. To address this issue, we extend the framework by modeling SRP functions as a multidimensional matrix, thereby preserving time-frequency information. Additionally, we employ multi-dictionary sparse Bayesian learning as the sparse optimization method to identify source positions without prior knowledge of their quantity. We validate our method through practical experiments using a 16-channel planar microphone array and compare it against three other localization methods. Results demonstrate that our proposed method outperforms other methods, including the current state-of-the-art, in localizing closely spaced sources in reverberant environments.

Index Terms— Source Localization, Steered Response Power, Sparse Representation, Sparse Bayesian Learning

1. INTRODUCTION

Acoustic source localization in reverberant environments is an active problem in microphone signal processing. Several methods have been developed for source localization, such as subspace-based approaches like MULTiple SIgnal Classification (MUSIC) [1, 2] and ESPRIT [3], Time Differences of Arrival (TDOA) approaches like Generalized Cross-Correlation Phase Transform (GCC-PHAT) [4, 5] and Steered Response Power (SRP) [4, 6], sparsity-based approaches like Orthogonal Matching Pursuit (OMP) [7], Sparse Bayesian Learning (SBL) [8, 9], as well as learning-based approaches [10, 11].

The SRP method estimates sound source positions by the summation of the cross-correlation of all possible microphone pairs. Recently, several approaches have focused on hierarchical search [12, 13], and real-time SRP [14] to reduce errors and complexity. However, SRP still struggles with localization in scenarios where multiple sources are closely spaced in reverberant environments due to an ambiguous SRP map.

Some research has overcome this challenge by iterative grid decomposition [15] or sparse fitting [16, 17]. However, the former approach still requires careful grid selection, especially in scenarios with numerous sources. The latter performs well in low reverberant scenarios but degrades dramatically as reverberation increases.

This paper proposes a method to improve localization performance using the sparse-fitting approach. To reduce high localization errors caused by reverberant environments, we represent the obtained SRP maps as multidimensional matrices to preserve more time-frequency information. We apply multi-dictionary SBL (M-SBL) [18] as the sparse optimization method, allowing the proposed method to operate without prior knowledge of the number of sources, thereby adapting to adverse real-world conditions. We utilize two sets of candidate grids with different resolutions for SRP mapping and sparsity fitting to improve efficiency further. We validate the proposed method through practical experiments with a planar microphone array, demonstrating improvements in localizing multiple closely spaced sources in a reverberant room compared to conventional and state-of-the-art methods [17]. Results show our method enhances robustness, allowing for stable performance even when localizing on short-duration recordings.

2. PROBLEM FORMULATION

Consider J sound sources, $j = \{1, \dots, J\}$, located at positions \mathbf{y}_j that each emit signals s_j . Let there be a microphone array composed of M elements, each positioned at \mathbf{x}_m for $m = \{1, \dots, M\}$. The received signals can be expressed as

$$p_m(t) = s_j(t) * g_m(t, \mathbf{y}_j) + v_m(t), \quad (1)$$

where $p_m(t)$ denotes the m^{th} -microphone's signal at time t , $g_m(t, \mathbf{y}_j)$ denotes the impulse response from the j^{th} source to the m^{th} microphone, and v_m denotes a noise term.

The TDOA between the pair of microphones (m, m') due to the j^{th} sound source is given by [4]:

$$\tau_{m,m'}(\mathbf{y}_j) = \frac{1}{c} \begin{cases} \|\mathbf{y}_j - \mathbf{x}_m\| - \|\mathbf{y}_j - \mathbf{x}_{m'}\| & \text{for NF,} \\ (\mathbf{x}_m - \mathbf{x}_{m'}) \cdot \vec{\mathbf{y}}_j & \text{for FF,} \end{cases} \quad (2)$$

where c is the speed of sound, NF and FF denote near-field and far-field sound propagation, respectively.

TDOA-based sound source localization aims to find the source locations of \mathbf{y}_j from estimated TDOAs (2) of a microphone array.

3. SOURCE LOCALIZATION BY SRP

Two common TDOA-based localization methods are GCC-PHAT [4, 5], and SRP-PHAT [6]. While GCC-PHAT maximizes the GCC function of a microphone pair (m, m') to estimate TDOAs, SRP-PHAT maximizes the GCC functions of all microphone pairs.

The GCC function at a time-frequency frame is defined as

$$R_{m,m'}(\tau, k, t) = \Psi_{m,m'}(k, t) P_m(k, t) P_{m'}^*(k, t) e^{j2\pi k\tau}, \quad (3)$$

where $k = \{1, \dots, K\}$ and $t = \{1, \dots, T\}$ index the wave number and the time frame in time-frequency domain, respectively, K and T denote the totals, $P_m(k, t)$ is the short-time Fourier transform of $p_m(t)$, $(\cdot)^*$ is the complex conjugate, $\Psi_{m,m'}(k, t) = 1/|P_m(k, t)P_{m'}^*(k, t)|$ is the frequency-dependent weighting function for GCC-PHAT.

Assuming all sound sources belong to a set of N candidate positions $\mathbf{y}_n \in \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$, where $N > J$, the SRP function is given as

$$z_{k,t}(\mathbf{y}_n) = \Re \left(\sum_{\ell=1}^L R_{\ell}(\tau, \mathbf{y}_n, k, t) \right), \quad (4)$$

where $\ell \equiv (m, m')$ denotes a microphone pair, $L = \binom{M}{2}$ denotes the total number of microphone pairs, and $\Re(\cdot)$ denotes the real part.

The SRP map comprises the SRP functions (4) of all N candidate positions. When the obtained SRP reaches a local maximum, the candidate position will be an estimated source location. In practice, the source positions $\hat{\mathbf{y}}_j$ are estimated by averaging SRP over time and frequency as follows:

$$\hat{\mathbf{y}}_j = \arg \max_{n \in N} z(\mathbf{y}_n) = \arg \max_{n \in N} \sum_{k=1}^K \sum_{t=1}^T z_{k,t}(\mathbf{y}_n), \quad (5)$$

where $z(\mathbf{y}_n)$ denotes the averaged SRP function of the n^{th} candidate position.

The SRP method for localization is robust in adverse environments after time-frequency averaging. However, multiple sources close to each other can still lead to an ambiguous SRP map and poor source localization. In the next section, we address this issue of closely spaced sources by employing a sparsity-based SRP method with M-SBL [18].

4. PROPOSED METHOD

We propose optimizing the ambiguous SRP maps using M-SBL. We aim for a sparse solution that enforces concentrated peaks of source weights, enabling the identification of multiple closely spaced sources.

We define another set of Q candidate positions $\mathbf{y}_q \in \{\mathbf{y}_1, \dots, \mathbf{y}_Q\}$. The microphone signal model from \mathbf{y}_q to \mathbf{x}_m

can be expressed as a linear regression model by STFT as follows:

$$P_m(k, t) = G_{q,m}(k) S_q(k, t) + \mathcal{N}_m(k, t) \quad (6)$$

where $P_m(k, t)$ denotes the m^{th} microphone signal at the k^{th} wave number and the t^{th} time frame, $G_{q,m}$ denotes the dictionary matrix, $S_q(k, t)$ denotes the source weight at \mathbf{y}_q , and $\mathcal{N}_m(k, t)$ denotes the noise term.

We assume the reverberation of the environment is unknown; hence, the dictionary function $G_{q,m}(k)$ is modeled as Green's functions in the case of NF or FF scenarios, respectively, as follows:

$$G_{q,m}(k) = \begin{cases} e^{jk\|\mathbf{x}_m - \mathbf{y}_q\|}/(4\pi\|\mathbf{x}_m - \mathbf{y}_q\|) & \text{for NF,} \\ e^{-jk\vec{\mathbf{y}}_q \cdot \mathbf{x}_m} & \text{for FF.} \end{cases} \quad (7)$$

By combining (3) and (6), we can obtain the corresponding GCC from this candidate position \mathbf{y}_n as follows:

$$R_{m,m'}(\tau, k, t) \approx \frac{S_q(k, t) G_{q,m}(k) (S_q(k, t) G_{q,m'}(k))^*}{|S_q(k, t) G_{q,m}(k) (S_q(k, t) G_{q,m'}(k))^*|} e^{j2\pi k\tau}, \quad (8)$$

where the TDOA τ corresponds one-to-one with the first set of N candidate positions. It is worth noting to clarify the two sets of candidate grids in the proposed method: N for SRP mapping and Q for subsequent sparsity fitting. Following (7), the TDOA between the microphone pair (m, m') for the candidate position \mathbf{y}_n , denoted as $\tau_{m,m'}(\mathbf{y}_n)$, can be expressed as follows:

$$\tau_{m,m'}(\mathbf{y}_n) = \frac{1}{j2\pi k} \ln (G_{n,m'}(k) G_{n,m}^*(k)). \quad (9)$$

By combining (4), (8), and (9), the SRP can be reformulated as:

$$z_{k,t}(\mathbf{y}_n, \mathbf{y}_q) \approx \Re \left(\sum_{\ell=1}^L H_{n,\ell}^*(k) \frac{H_{q,\ell}(k)}{|H_{q,\ell}(k)|} \right), \quad (10)$$

where $H_{n,\ell}(k) = G_{n,m}(k) G_{n,m'}^*(k)$ is the relative transfer function of the microphone pair ℓ [19].

We define vectors $\mathbf{a}_{\ell}(k) \in \mathbb{C}^{N \times 1}$ and $\mathbf{b}_{\ell}(k) \in \mathbb{C}^{1 \times Q}$ to represent the right part of (10) for brevity, where

$$\begin{aligned} \mathbf{a}_{\ell}(k) &= \left[H_{1,\ell}^*(k), \dots, H_{N,\ell}^*(k) \right]^{\top}, \\ \mathbf{b}_{\ell}(k) &= \left[\frac{H_{1,\ell}(k)}{|H_{1,\ell}(k)|}, \dots, \frac{H_{Q,\ell}(k)}{|H_{Q,\ell}(k)|} \right]. \end{aligned} \quad (11)$$

Then, we can regard the SRP map of all candidate positions as the summation of (10) as follows:

$$\mathbf{z}_{k,t} = \Re (\mathbf{A}_k \mathbf{B}_k \mathbf{S}_{k,t}) + \mathcal{N}_{k,t}, \quad (12)$$

where $\mathbf{z}_{k,t} = [z_{k,t}(\mathbf{y}_1), \dots, z_{k,t}(\mathbf{y}_N)]^{\top} \in \mathbb{R}^{N \times 1}$ is the SRP map, matrices $\mathbf{A}_k = [\mathbf{a}_1(k), \dots, \mathbf{a}_L(k)] \in \mathbb{C}^{N \times L}$ and $\mathbf{B}_k =$

$[\mathbf{b}_1(k), \dots, \mathbf{b}_L(k)]^\top \in \mathbb{C}^{L \times Q}$ are formed by vectors \mathbf{a}_ℓ and \mathbf{b}_ℓ , respectively, $\mathbf{S}_{k,t} = [S_1(k,t), \dots, S_Q(k,t)]^\top \in \mathbb{C}^{Q \times 1}$ denotes the source weight matrix of Q candidate positions, and $\mathcal{N}_{k,t} \in \mathbb{R}^{N \times 1}$ denotes the noise term.

Since sources are sparse in number, we assume $\mathbf{S}_{k,t}$ should have mostly zero entries. Under this assumption, we define (12) as an underdetermined system where $Q \geq N$. This indicates that we represent SRP maps using a higher-resolution source weight map. Hence, (12) is a sparse representation of the SRP map for a single time frame and wave number. Such a sparse solution of $\mathbf{S}_{k,t}$ can be obtained by solving the following optimization problem:

$$\min_{\mathbf{S}_{k,t}} \frac{1}{2} \|\mathbf{z}_{k,t} - \Re(\mathbf{A}_k \mathbf{B}_k \mathbf{S}_{k,t})\|_2^2 + \lambda \|\Re(\mathbf{S}_{k,t})\|_p^p, \quad (13)$$

where $\|\cdot\|_p$ denotes the l_p -norm as $0 < p \leq 1$.

To further enhance the robustness, we consider incorporating time-frequency diversity from SRP inputs. We here represent the SRP function as a multidimensional matrix $\mathbf{Z} \in \mathbb{R}^{N \times T \times K}$ instead of the vector form $\mathbf{z} \in \mathbb{R}^{N \times 1}$ in (5). We then express a group-sparse representation of the multidimensional matrix \mathbf{Z} as

$$\mathbf{Z} = \Re(\mathbf{ABS}) + \mathcal{N}, \quad (14)$$

where $\mathbf{Z} \in \mathbb{R}^{N \times T \times K}$, $\mathbf{A} \in \mathbb{C}^{N \times L \times K}$, $\mathbf{B} \in \mathbb{C}^{L \times Q \times K}$, $\mathbf{S} \in \mathbb{C}^{Q \times T \times K}$, and $\mathcal{N} \in \mathbb{R}^{N \times T \times K}$. Equation (14) represents multidimensional matrix multiplication.

We solve the group-sparse representation (14) by introducing a multidimensional mixed-norm penalty term [20] as

$$\min_{\mathbf{S}} \frac{1}{2} \|\mathbf{Z} - \Re(\mathbf{ABS})\|_2^2 + \lambda \mathcal{J}_{p_1, p_2, p_3}(\Re(\mathbf{S})), \quad (15)$$

where p_1, p_2, p_3 denote the norms of spatial, time, and frequency domain, respectively, following a descending order where $0 < p_1 \leq p_2 \leq p_3 \leq 2$, and $\mathcal{J}_{p_1, p_2, p_3}(\Re(\mathbf{S}))$ denotes the penalty function of $\Re(\mathbf{S})$.

Here we assume source signals are non-sparse in time-frequency domain, hence we set $p_1 = p$, $p_2 = 2$, $p_3 = 2$, which means the penalty term of (15) is an $l_{p,2,2}$ -norm function.

We next apply M-SBL to optimize (15) since it is suitable for solving the underdetermined problem without the prior knowledge of the source quantities J . The M-SBL method is based on two assumptions. First, the noise term \mathcal{N} is assumed to be the zero-mean Gaussian noise term with density $\mathcal{N}(\mathcal{N}; \mathbf{0}, \sigma^2 \mathbf{I})$. Second, the source weight \mathbf{S} is assumed to be the zero-mean complex Gaussian with density $\mathcal{CN}(\mathbf{S}; \mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$ is the diagonal matrix of hyperparameters $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_Q]$.

In the M-SBL framework, the hyperparameters $\boldsymbol{\gamma}$ are assumed to be unknown and learned iteratively by maximizing the evidence function [18, eq. (16)] to reach a sparse result. Here, the evidence model is also the zero-mean Gaussian with density $\mathcal{N}(\mathbf{Z}; \mathbf{0}, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I} + \mathbf{G}_k \boldsymbol{\Gamma} \mathbf{G}_k^H$. The

maximization of $\hat{\boldsymbol{\gamma}}$ is carried out through an iterative equation derived from the derivatives of the evidence function [18, eq. (21)]. Further procedural details of M-SBL can be found in [18]. Finally, the update equation is obtained as

$$\boldsymbol{\gamma}_q^{\text{new}} = \frac{\boldsymbol{\gamma}_n^{\text{old}}}{T} \frac{\sum_{k=1}^K \|\mathbf{Z}_k^H \boldsymbol{\Sigma}_k^{-1} \mathbf{D}_{k,q}\|_2^2}{\sum_{k=1}^K \mathbf{D}_{k,q}^H \boldsymbol{\Sigma}_k^{-1} \mathbf{D}_{k,q}}, \quad (16)$$

where $\mathbf{D}_{k,q}$ denotes the q^{th} column of the dictionary matrix $\mathbf{D}_k = \Re(\mathbf{A}_k \mathbf{B}_k)$. The estimated source locations will be the candidate positions corresponding to the largest peaks of $\hat{\boldsymbol{\gamma}}$. It is worth noting that M-SBL terminates iterations when the error falls below the convergence threshold, thus removing the need to determine the number of sources.

Compared to the current state-of-the-art sparsity-based modeling method for SRP maps [17], the proposed method represents obtained SRP maps as a multidimensional matrix \mathbf{Z} instead of a time-frequency-averaged vector \mathbf{z} to improve localization by retaining time-frequency information. Additionally, we set two candidate grids for SRP mapping and sparse fitting with different resolutions for efficiency. In the next section, we will compare these two methods and other localization methods through experiments.

5. EXPERIMENTAL VALIDATION

We conducted recordings in a large meeting room with a hard floor and ceiling at 3.3 m height, with a $T_{60} \approx 0.7$ s ($T_{20} = 113$ ms measured). We used a $M = 16$ -channel MEMS microphone planar array (MiniDSP UMA-16 v2) and $J = 3$ target loudspeakers (Genelec 8030C) positioned in the room as detailed in Fig. 1. The loudspeakers played speech signals from the MS-SNSD dataset [21]. The recording and localization used a 48 kHz sampling frequency and 1024 STFT frame length with 50% overlap. We note that the speakers' ground truth positions and DOAs in Fig. 1 were obtained manually and are thus prone to human error.

We evaluate the DOA estimation performance of the proposed method through this experiment. We, therefore, use FF propagation for (2) and (7). The candidate DOAs are defined over a hemisphere since the planar array struggles to distinguish sources from the front and back. We select $N = 247$ candidate DOAs for the SRP map, with a 15° resolution over elevation range $\theta \in [-90^\circ, 90^\circ]$ and a 10° resolution over azimuth range $\phi \in [0^\circ, 180^\circ]$, $Q = 8281$ candidate DOAs for the M-SBL, with a 2° resolution for elevation θ and a 2° -resolution for azimuth ϕ . The grids are with respect to the center of the microphone array. We set the convergence error of M-SBL as 10^{-3} . We denote the proposed method as SRP-SBL in the results for brevity.

We compare SRP-SBL to three methods: conventional SRP-PHAT [6], M-SBL [18], and SRP-sparsity (SRP-S) [17], the current state-of-the-art sparsity-based modeling method for SRP maps. The candidate DOAs for these methods are

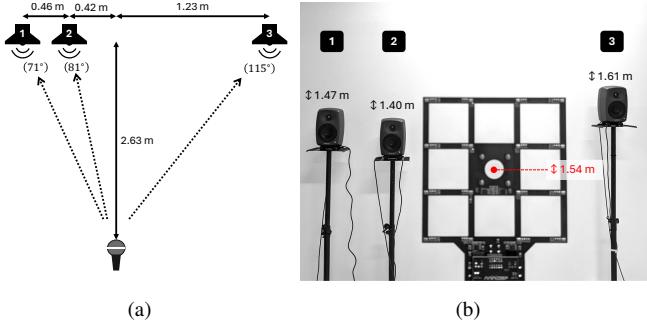


Fig. 1. Illustration of experiment setup. Loudspeaker distance and azimuth positions are drawn in (a). Loudspeaker and microphone heights are pictured in (b). The DOAs of the loudspeakers y_1, y_2, y_3 correspond to the numbers 1, 2, 3 indicated in the figures, respectively.

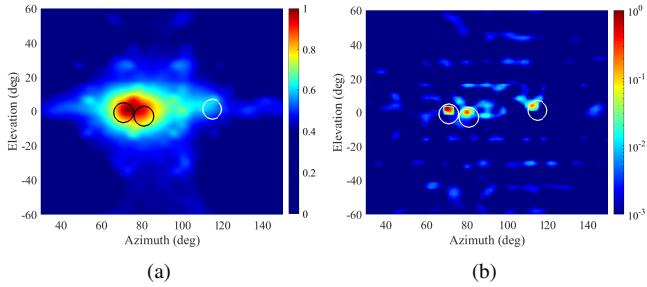


Fig. 2. Normalized output maps of (a) SRP-PHAT and (b) SRP-SBL. Note that the color scale of (b) is logarithmic to illustrate the localized peaks better. The true DOAs are denoted by (○).

the same as SRP-SBL with $Q = 8281$. Note that SBL utilizes microphone measurements \mathbf{P} for sparse optimization, while SRP-SBL uses the SRP maps \mathbf{Z} . The sparse optimization method for SRP-S in [17] utilized the ADMM solver. However, since the regularization parameter in ADMM requires careful selection under adverse conditions, we instead employed Simultaneous OMP [22] as an alternative optimization method in our evaluation.

In Fig. 2, we demonstrate the output maps of conventional SRP-PHAT and our proposed SRP-SBL. In Fig. 2(a), we observe that the generated SRP-PHAT map exhibits an ambiguous peak at $(\theta, \phi) = (0^\circ, 75^\circ)$, making it challenging to distinguish all three sources. In contrast, the SRP-SBL map in Fig. 2(b) clearly distinguishes the three individual sources and their relative locations. We note that the color scale of this SRP-SBL map is logarithmic.

Fig. 3 presents an azimuth slice of the Fig. 2 results along with the results of M-SBL and SRP-S. We observe that M-SBL is unable to localize the sources correctly. Whereas, SRP-S correctly localizes y_1 and closely finds y_2 but misses y_3 . However, the peaks estimated by SRP-SBL nearly align with the true DOAs.

We evaluate the impact of recording duration on local-

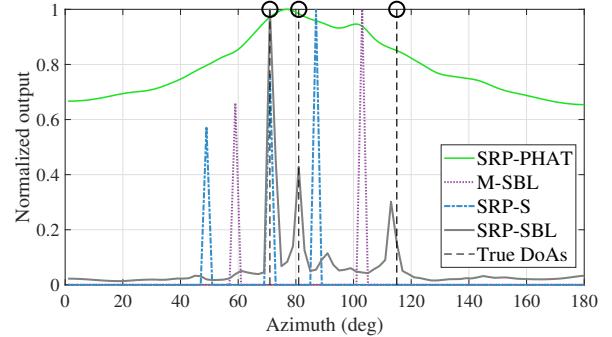


Fig. 3. Normalized output maps of the four methods and ground truth along the azimuth angles. The true DOAs are denoted by (○).

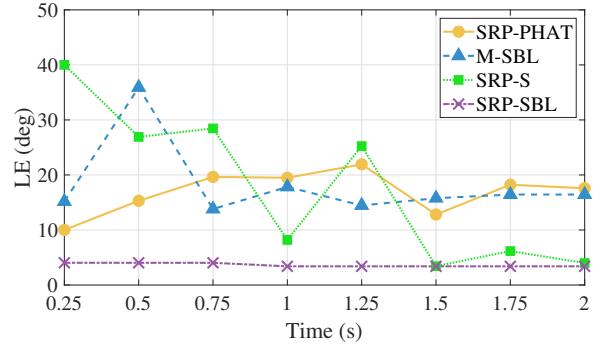


Fig. 4. LE for varying recording durations.

ization accuracy in Fig. 4. To avoid re-selecting the same peak, we here impose a constraint that each estimated DOA should be spaced 3° apart. The localization error is given as the average angles between the estimated DOA $(\hat{\theta}_j, \hat{\phi}_j)$ of the J highest peaks and the nearest true DOA (θ_j, ϕ_j) as $\text{LE} = \frac{1}{J} \sum_{j=1}^J \cos^{-1}(\sin(\hat{\theta}_j) \sin(\theta_j) \cos(\hat{\phi}_j - \phi_j) + \cos(\hat{\theta}_j) \cos(\theta_j))$. It can be observed that both SRP-PHAT and M-SBL yield significant LE for all recording duration. The SRP-S method performs well for recording duration greater than 1s but exhibits instability for duration ≤ 1 s. In contrast, the proposed SRP-SBL method is robust across various recording durations because we maintain time and frequency diversity.

6. CONCLUSION

This paper proposes a sparsity-based optimized SRP method for localizing multiple neighboring sources. The method utilizes a multidimensional SRP matrix as input and optimizes it through multidimensional SBL (M-SBL). Practical experiment results indicate that the proposed method outperforms conventional SRP, M-SBL, and the current state-of-the-art SRP sparsity-based method (SRP-S), maintaining stable localization performance in reverberant environments. In contrast, multiple closely spaced sources cause SRP and M-SBL to consider all sources as a single point. While SRP-S enhances localization performance, it loses its robustness in low-recording-duration scenarios.

7. REFERENCES

- [1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
- [2] L. Birnie, T. D. Abhayapala, H. Chen, and P. N. Samarasinghe, "Sound source localization in a reverberant room using harmonic based music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 651–655.
- [3] 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.
- [4] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," *Ph. D. Thesis*, 2000.
- [5] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, pp. 508–519, 2022.
- [6] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.
- [7] G. Chardon, T. Nowakowski, J. de Rosny, and L. Daudet, "A blind dereverberation method for narrowband source localization," *IEEE J. Selected Topics Signal Process.*, vol. 9, no. 5, pp. 815–824, 2015.
- [8] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [9] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust ocean acoustic localization with sparse bayesian learning," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 49–60, 2019.
- [10] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* IEEE, 2017, pp. 136–140.
- [11] A. S. Roman, I. R. Roman, and J. P. Bello, "Robust doa estimation from deep acoustic imaging," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2024, pp. 1321–1325.
- [12] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 499–508, 2004.
- [13] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, F. M. Gonçalves, A. Said, B. Lee, et al., "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.
- [14] H. Do, H. F. Silverman, and Y. Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2007, vol. 1, pp. I–121.
- [15] A. Martí, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2627–2630, 2013.
- [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. 13781–13812, 2012.
- [17] E. Tengan, T. Dietzen, F. Elvander, and T. Van Waterschoot, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered response power maps," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* IEEE, 2023, pp. 1–5.
- [18] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse bayesian learning for doa," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1469–1473, 2016.
- [19] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.
- [20] N. Murata, S. Koyama, N. Takamune, and H. Saruwatari, "Sparse representation using multidimensional mixed-norm penalty with application to sound field decomposition," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3327–3338, 2018.
- [21] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Inter-speech*, pp. 1816–1820, 2019.
- [22] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.