Sound Source Localization using Multi-Dictionary Orthogonal Matching Pursuit in Reverberant Environments

Wei-Ting Lai, Lachlan Birnie, Thushara Abhayapala, *Fellow, IEEE*, Amy Bastine, and Prasanga Samarasinghe, *Senior Member, IEEE*

***Abstract*—Sound source localization in reverberant environments remains a challenging problem, particularly when precise position estimation is required. Existing DOA estimation methods, while effective in determining sound direction, often fall short of accurate position estimation in adverse acoustic conditions due to their inherent constraints. In this paper, we propose Multi-Dictionary Orthogonal Matching Pursuit (MD-OMP), extending sparse recovery methods by jointly exploiting spatial, temporal, and spectral information for source position estimation. MD-OMP iteratively selects candidate grid entries that maximize joint correlation across time frames and frequency bins, thereby enhancing localization accuracy under reverberant conditions. We also introduce a generalized extension of the algorithm to accommodate an arbitrary number of multidimensional domains, thereby supporting advanced applications such as higher-order source modeling and scenarios with dominant frequency components. To validate the effectiveness of MD-OMP, we arrange microphones along the room walls to form a perimeter array and conduct both extensive simulations and practical experiments in a conference room setting. Numerical simulations under various acoustic conditions—including different reverberation times, candidate grid densities, frequency ranges, source types, and room dimensions—show that MD-OMP consistently outperforms group-sparsity-based and TDOA-based benchmark methods in localization accuracy. Experimental results in a realistic office environment further verify its practical applicability. Overall, MD-OMP achieves accurate and robust source position estimation across diverse and challenging scenarios.**

***Index Terms*—source localization, source position estimation, sparse representation, matching pursuit, spatio-temporal-spectral fusion.**

I. Introduction

A. Literature Review

SOUND source localization is a fundamental task in acoustic signal processing, with applications in speech enhancement [1], sound field reconstruction [2], [3], and sound event detection [4]. This task can be categorized into two main problems: direction-of-arrival (DOA) estimation and position estimation. DOA estimation focuses on determining the azimuth and elevation angles of a source, whereas position estimation seeks its precise three-dimensional location.

Over time, many studies have defaulted to treating source localization as DOA estimation, leading to the development of numerous methods. Early approaches include subspace-based methods such as Multiple Signal Classification (MUSIC) [5]–[7] and Estimation of Signal Parameters via Rotational Invariance Techniques [8]–[10]; Time Difference of Arrival (TDOA) methods like Generalized Cross-Correlation with Phase Transform [11], [12] and Steered Response Power with Phase Transform (SRP-PHAT) [13]–[16]; sparsity recovery methods such as Lasso regression [17], [18], Sparse Bayesian Learning (SBL) [19]–[22], Orthogonal Matching Pursuit (OMP) [3], [23]–[25], and Compressive Sampling Matching Pursuit (CoSaMP) [25]; and learning-based methods [4], [26]–[29].

While these methods perform well for DOA estimation, they often face challenges in near-field position estimation due to inherent constraints. For example, MUSIC relies on the far-field assumption; Lasso-based methods struggle with multiple sources; and learning-based methods depend on extensive datasets, which are typically designed for DOA estimation. To overcome these limitations, several variants of these methods have been developed specifically for position estimation, such as Near-Field MUSIC [30], [31], Harmonic-Domain MUSIC [32], Near-Field SRP-PHAT (NF-SRP-PHAT) [33]–[35], TDOA with sparse regularization [36], and Iteratively Reweighted SRP-PHAT (IR-SRP-PHAT) [37]. Extended learning-based approaches have also been explored for position estimation [38]–[41]. Nevertheless, these methods still face challenges: conventional approaches degrade severely in reverberant environments, where near-field conditions introduce strong mutual interference between source signals, while learning-based methods lack adaptability because their performance depends heavily on the microphone arrays, room configurations, and source signals used during training.

Geometry-based methods represent an alternative approach to position estimation [42]–[46]. These methods infer the three-dimensional source positions by intersecting bearings from DOA triangulation [47]–[52] or hyperbolic surfaces from TDOA multilateration [44], [45], [53], [54]. Despite their low computational cost, they are highly sensitive to the accuracy of the DOA/TDOA estimates [42], [46]. In reverberant multi-source scenarios, incorrect cross-array pairing yields ghost intersections, making the true source positions difficult to identify [51], [54].

In contrast to these methods, sparse recovery approaches stand out due to their flexibility in handling both DOA estimation and source position estimation [14], [17]–[25], [55]–[59]. These sparsity-based methods assume that sound sources are located at predefined candidate spatial grids and impose a sparsity condition, where most of the source weights are zero. This assumption improves localization accuracy and reduces the number of microphones required. While effective,

This work was supported by the Taiwan-Australian National University Scholarship.

The authors are with the Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, ACT 2601, Australia (e-mail: wei-ting.lai@anu.edu.au; lachlan.birnie@anu.edu.au; thushara.abhayapala@anu.edu.au; amy.bastine@anu.edu.au; prasanga.samarasinghe@anu.edu.au).

sparsity-based methods for sound source position estimation are constrained by two primary factors. First, source position estimation typically requires a significantly higher number of microphones than DOA estimation. For example, previous studies employed a 48-channel spherical array [57] and a 42-channel random array [58]. However, while improving accuracy, these methods impose substantial hardware costs, limiting practical deployment in resource-constrained scenarios.

Second, standard sparsity-based formulations are inherently limited to narrowband signals, making source position estimation particularly challenging. To address this issue, some advanced approaches, including Group Iteratively Reweighted Least Squares (G-IRLS) [56], Multi-Dictionary SBL (MD-SBL) [60], [61], and SBL with Principal Component Analysis [57], apply a group-sparsity framework to integrate multi-frequency information. Nonetheless, these methods still face challenges, such as repeated selection errors and increased computational complexity due to the use of multi-frequency dictionaries.

OMP [23], [24], [62] is a widely used greedy algorithm for sparse recovery that overcomes the computational challenges posed by other sparsity methods. It iteratively selects entries with the highest correlation to the residual signal. Over time, variants have emerged to address multidimensional inputs. For example, Simultaneous OMP (S-OMP) [56], [63] processes multidimensional measurements concurrently, while Multidimensional OMP (M-OMP) [64] reduces computational cost by sequentially refining search grids along individual coordinate axes. However, both approaches are limited by fixed domain configurations. S-OMP is confined to a two-dimensional framework, and M-OMP is restricted to a single domain and cannot handle inputs from multiple domains simultaneously.

OMP has demonstrated strong performance in sound source position estimation by iteratively selecting the most relevant dictionary atoms and projecting onto their corresponding subspaces, thereby avoiding repeated selection errors [58], [59]. Nonetheless, the extension of OMP to concurrently integrate spatial, temporal, and spectral information remains unexplored, which constrains its practical applications in position estimation.

B. Contributions

In this paper, we propose Multi-Dictionary OMP (MD-OMP) for sound source position estimation. The main contributions of this paper are summarized as follows:

1) *Proposal of MD-OMP*

To the best of our knowledge, MD-OMP is the first OMP-based method in audio signal processing that jointly integrates spatio-temporal-spectral information. It exploits a group-sparsity framework to address key localization challenges, including multiple sources and reverberant environments. To identify source positions, MD-OMP employs multidimensional dictionaries and iteratively selects the entry with the highest joint correlation across different time frames and frequency bins. While it does not require prior environmental knowledge, it assumes that the relative positions of the microphones are known.

2) *Generalization to Arbitrary Multidimensional Domains*

We introduce a generalization that allows MD-OMP to accommodate an arbitrarily defined number of multidimensional domains—each representing a distinct set of variables (e.g., spatial, temporal, or spectral features)—to further enhance its applicability. This flexibility enables the method to adapt to complex acoustic conditions, including higher-order sources, dominant frequency components, and moving microphones. By formulating the corresponding signal model based on domain characteristics, such as sparsity and dictionary relevance, MD-OMP improves source position estimation accuracy and robustness.

3) *Validation with Design of a Practical Microphone Array Configuration*

We evaluate MD-OMP in a conference room scenario with multiple talkers, reflecting real-world applications. To ensure accurate localization while minimizing the number of microphones, we design a perimeter microphone array arranged along the room walls, which is specifically optimized for localizing multiple sources in reverberant environments. We validate MD-OMP through extensive numerical simulations—varying reverberation times, grid densities, frequency ranges, source signals, and room dimensions—and practical experiments in an office room setting. The results demonstrate that MD-OMP outperforms existing group-sparsity-based and TDOA-based methods in both localization accuracy and robustness.

The remainder of this paper is organized as follows: Section II presents the problem formulation. Section III reviews the standard OMP algorithm. Section IV details the proposed MD-OMP algorithm. Section V and Section VI provide the numerical simulations and practical experiments, respectively. Section VII discusses the localization results and the evaluation of computational time. Finally, Section VIII concludes the paper.

II. Problem Formulation

In this section, we formulate the problem of localizing near-field sound sources in reverberant environments using a group-sparse representation.

Consider an arbitrary reverberant acoustic environment containing J point sources distributed within a predefined Region of Interest (ROI) Ω, with each source positioned at $\mathbf{y}_j \equiv (x_j, y_j, z_j)$ in Cartesian coordinates for $j = 1, 2, \ldots, J$. A total of M microphones are placed at positions $\mathbf{x}_m \equiv (x_m, y_m, z_m)$ for $m = 1, 2, \ldots, M$. A schematic illustration of the setup is shown in Fig. 1.

Let $S_j(t, f)$ denote the signal emitted by the j^{th} source in short-time Fourier transform (STFT) domain, where $t \in \{1, \cdots, T\}$ and $f \in \{1, \cdots, F\}$ index the time frame and the frequency bin, respectively. The sound pressure $P_m(t, f)$ received by the m^{th} microphone is given by:

$$P_m(t, f) = \sum_{j=1}^{J} G(f, \mathbf{x}_m, \mathbf{y}_j) S_j(t, f) + \mathcal{N}_m(t, f), \quad (1)$$

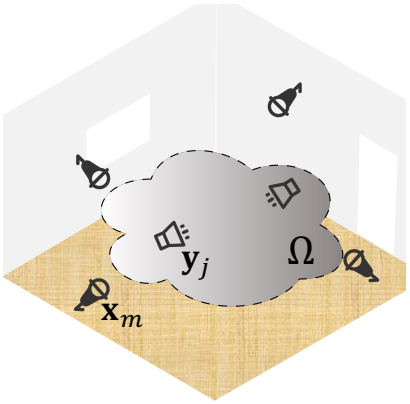

Fig. 1. Illustration of the source localization setup in a reverberant room, depicting sound sources positioned inside Ω, with microphones placed outside this region.

where $G(f, \mathbf{x}_m, \mathbf{y}_j)$ denotes the transfer function from the j^{th} source to the m^{th} microphone, and $\mathcal{N}_m(t, f)$ denotes the noise term.

In the source localization problem, the acoustic environment and room geometry are generally unknown. As a result, $\mathcal{N}_m(t, f)$ incorporates sensor noise, early reflections, and late reverberation. Under this assumption, the transfer function $G(f, \mathbf{x}_m, \mathbf{y}_j)$ is modeled using the free-field Green's function of a point source:

$$G(f, \mathbf{x}_m, \mathbf{y}_j) = \frac{e^{ik_f \|\mathbf{x}_m - \mathbf{y}_j\|}}{4\pi \|\mathbf{x}_m - \mathbf{y}_j\|}, \tag{2}$$

where k_f denotes the f^{th} wavenumber.

Assuming all J sound sources are located within a discrete set of N candidate positions in the region Ω, denoted as $\mathbf{y}_n \in \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$, where $N \gg J$, the microphone signal model can be reformulated in matrix form as:

$$\mathbf{P}_{t,f} = \mathbf{G}_f \mathbf{S}_{t,f} + \boldsymbol{\mathcal{N}}_{t,f}, \tag{3}$$

where $\mathbf{P}_{t,f} = [P_1(t,f) \ \cdots \ P_M(t,f)]^\top \in \mathbb{C}^M$ and $\mathbf{S}_{t,f} = [S_1(t,f) \ \cdots \ S_N(t,f)]^\top \in \mathbb{C}^N$ denote the microphone signals and source weights of the f^{th} frequency bin and t^{th} time frame, respectively, $\boldsymbol{\mathcal{N}}_{t,f} \in \mathbb{C}^M$ denotes the noise term, and $\mathbf{G}_f \in \mathbb{C}^{M \times N}$ denotes the dictionary matrix for the f^{th} frequency bin, where its (m, n) entry represents the transfer function from the n^{th} candidate position to the m^{th} microphone, given by:

$$\mathbf{G}_f = \begin{bmatrix} G(f, \mathbf{x}_1, \mathbf{y}_1) & \cdots & G(f, \mathbf{x}_1, \mathbf{y}_N) \\ \vdots & \ddots & \vdots \\ G(f, \mathbf{x}_M, \mathbf{y}_1) & \cdots & G(f, \mathbf{x}_M, \mathbf{y}_N) \end{bmatrix}. \tag{4}$$

Since the sound sources are assumed to be sparsely distributed in Ω, with $N \gg J$, $\mathbf{S}_{t,f}$ will have only a few non-zero entries. Therefore, such a solution of $\mathbf{S}_{t,f}$ can be solved using the following sparse optimization model:

$$\min_{\mathbf{S}_{t,f}} \frac{1}{2} \|\mathbf{P}_{t,f} - \mathbf{G}_f \mathbf{S}_{t,f}\|_2^2 + \mu \|\mathbf{S}_{t,f}\|_1, \tag{5}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 and ℓ_1 norms, respectively, and μ denotes the regularization parameter.

Solving (5) yields source estimates $\hat{\mathbf{S}}_{t,f}$, from which source positions $\hat{\mathbf{y}}_j$ can be inferred by selecting the non-zero entries. However, relying solely on a single time-frequency bin provides limited information for localization, necessitating the use of a large number of microphones to achieve effective and robust performance. Additionally, the spatial sampling of the microphone array must adhere to the Nyquist criterion for spatial frequencies [65], which imposes strict constraints on array design to cover a wide frequency range.

To overcome these limitations, it is essential to incorporate spatio-temporal-spectral information across multiple time-frequency bins. Accordingly, the microphone signals $\mathbf{P}_{t,f}$, source weights $\mathbf{S}_{t,f}$, and transfer functions $\mathbf{G}_f$ in (3) are stacked over T time frames and F frequency bins, represented as multidimensional matrices:

$$\mathbf{P} = \begin{array}{l} \boxed{\mathbf{P}_{1,F}, \cdots, \mathbf{P}_{T,F}} \\ \boxed{\mathbf{P}_{1,2}, \cdots, \mathbf{P}_{T,2}} \ \cdots \\ \boxed{\mathbf{P}_{1,1}, \cdots, \mathbf{P}_{T,1}} \end{array} \in \mathbb{C}^{M \times T \times F}, \tag{6}$$

$$\mathbf{S} = \begin{array}{l} \boxed{\mathbf{S}_{1,F}, \cdots, \mathbf{S}_{T,F}} \\ \boxed{\mathbf{S}_{1,2}, \cdots, \mathbf{S}_{T,2}} \ \cdots \\ \boxed{\mathbf{S}_{1,1}, \cdots, \mathbf{S}_{T,1}} \end{array} \in \mathbb{C}^{N \times T \times F}, \tag{7}$$

$$\mathbf{G} = \begin{array}{l} \boxed{\mathbf{G}_F} \\ \boxed{\mathbf{G}_2} \ \cdots \\ \boxed{\mathbf{G}_1} \end{array} \in \mathbb{C}^{M \times N \times F}. \tag{8}$$

Hence, the multidimensional signal model can be formulated as:

$$\mathbf{P} = \mathbf{G} \times \mathbf{S} + \mathbf{N}, \tag{9}$$

where $\mathbf{N} \in \mathbb{C}^{M \times T \times F}$ denotes the noise, and the operator $\times$ represents slice-wise matrix multiplication, defined by:

$$\mathbf{P}_{:,t,f} = \mathbf{G}_{:,:,f} \mathbf{S}_{:,t,f} + \mathbf{N}_{:,t,f}, \tag{10}$$

where $\mathbf{P}_{:,t,f} \triangleq \mathbf{P}_{t,f}$, $\mathbf{S}_{:,t,f} \triangleq \mathbf{S}_{t,f}$, and $\mathbf{G}_{:,:,f} \triangleq \mathbf{G}_f$, so that each slice of the multidimensional signal model in (9) corresponds exactly to the single time-frequency bin model in (3).

In many practical localization scenarios, source signals (e.g., speech, music) are typically wideband and active over time. Therefore, the source distribution is assumed to be non-sparse in time and frequency domains, whereas the spatial distribution of sound sources is assumed to be sparse. To account for this varying sparsity across multiple domains, (9) can be solved using the following group-sparse optimization model:

$$\min_{\mathbf{S}} \frac{1}{2} \|\mathbf{P} - \mathbf{G} \times \mathbf{S}\|_2^2 + \mu \mathcal{J}_{1,2,2}(\mathbf{S}), \tag{11}$$

where $\mathcal{J}_{1,2,2}(\mathbf{S})$ denotes the $\ell_{1,2,2}$-norm penalty term, with $(1, 2, 2)$ corresponding to the norms of spatial, time, and frequency domains, respectively.

The goal of this paper is to estimate the source positions $\hat{\mathbf{y}}_j$ from the measured signals $\mathbf{P}$. This is achieved by solving (11) to obtain the source weights $\hat{\mathbf{S}}_{t,f}$, followed by selecting the entries corresponding to the largest J values among the N candidate positions. To this end, we extend OMP to incorporate

group sparsity. In the following section, we first provide a brief review of the conventional OMP framework as a foundation for our proposed extension.

III. Orthogonal Matching Pursuit

In this section, we introduce the conventional OMP, a widely used greedy algorithm for sparse approximation. It makes step-by-step decisions, selecting the option that minimizes the error at each stage, resulting in a locally optimal solution. OMP consists of three main steps in each iteration: atom selection, support augmentation, and orthogonal projection.

For a narrowband localization problem as in (3), OMP is initialized with an empty support set $\mathbf{\Lambda}^{(0)}$, an empty support index vector $\mathbf{\Psi}^{(0)}$, and an initial residual $\mathbf{r}^{(0)} = \mathbf{P}_{t,f}$. The algorithm then follows an iterative procedure to update these three parameters in order to identify source positions.

In the first step, atom selection, the current residual vector $\mathbf{r}^{(j-1)}$ is projected onto the columns of the dictionary matrix $\mathbf{G}_f$. This involves computing the correlation between $\mathbf{r}^{(j-1)}$ and each column of $\mathbf{G}_f$ to determine the column most aligned with the residual, as follows:

$$\lambda^{(j)} = \arg \max_{n \in \{1,2,\ldots,N\}} |\langle \mathbf{r}^{(j-1)}, \mathbf{g}_n^{(f)} \rangle|, \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\mathbf{g}_n^{(f)} \in \mathbb{C}^M$ is the n^{th} column (atom) of $\mathbf{G}_f$.

In the second step, support augmentation, the selected index $\lambda^{(j)}$ is added to the support index vector $\mathbf{\Psi}^{(j-1)}$, and the corresponding column of the dictionary $\mathbf{g}_{\lambda^{(j)}}^{(f)}$ is added to the support set $\mathbf{\Lambda}^{(j-1)}$. To prevent reselection in subsequent iterations, the selected column in $\mathbf{G}_f$ is nullified.

In the final step, orthogonal projection, OMP updates the residual $\mathbf{r}^{(j)}$ by projecting the measurement $\mathbf{P}_{t,f}$ onto the subspace spanned by the updated support $\mathbf{\Lambda}^{(j)}$. This orthogonal projection step ensures that the new residual is orthogonal to all previously selected columns, facilitating the selection of new indices in subsequent iterations. By repeating this process iteratively, OMP gradually constructs an accurate approximation of the original signal.

The stopping criterion for OMP can be set depending on whether the value of J is known a priori. If J is known, OMP stops after completing J iterations. Otherwise, the algorithm stops based on a convergence parameter ϵ, such that if $\|\mathbf{r}^{(j)}\| < \epsilon$, the process is terminated.

Overall, OMP provides both fast computation and high accuracy but is inherently designed for single-dimensional signals. To tackle the challenges posed by multidimensional data, we propose an advanced OMP algorithm in the next section.

IV. Proposed MD-OMP Algorithm

This section presents the proposed Multi-Dictionary OMP (MD-OMP) algorithm, which addresses wideband signals over multiple time frames and incorporates a group-sparse model to control sparsity across multiple dimensions. We begin with a detailed formulation and implementation of the core algorithm, then extend MD-OMP to a generalized multi-domain framework, and finally provide case studies illustrating its flexibility.

A. Core Algorithm of MD-OMP

Our MD-OMP builds upon the conventional OMP framework discussed in Section III, following an iterative procedure that incorporates group sparsity. Each iteration consists of three main steps: group-atom selection, group-support augmentation, and group-orthogonal projection. Similar to OMP, we initialize empty sets for the support $\mathbf{\Lambda}^{(0)}$ and indices $\mathbf{\Psi}^{(0)}$, but we initialize a multidimensional residual to accommodate the additional domains as $\mathbf{R}^{(0)} = \mathbf{P}$.

The objective of MD-OMP is to solve the group-sparse optimization problem formulated in (11), which aims to estimate the source weights $\hat{\mathbf{S}}$ by jointly considering spatio-temporal-spectral correlations. As discussed in Section II, we assume that the source distribution is sparse in spatial domain but non-sparse in time and frequency domains.

In group-atom selection, we compute joint correlation across all frequency bins and time frames to identify the most relevant grid position. The joint correlation function is defined as:

$$z_n = \sum_{f=1}^{F} \sum_{t=1}^{T} |\langle \mathbf{R}_{:,t,f}^{(j-1)}, \mathbf{G}_{:,n,f} \rangle|, \tag{13}$$

where z_n denotes the joint correlation of the n^{th} grid position, $\mathbf{R}_{:,t,f}^{(j-1)}$ denotes the multidimensional residual at the t^{th} time frame and f^{th} frequency bin for the $(j-1)^{\text{th}}$ iteration.

The candidate grid index with the maximum joint correlation is selected as:

$$\lambda^{(j)} = \arg\max_n(z_n), \tag{14}$$

where $\lambda^{(j)}$ denotes the selected grid index at the j^{th} iteration.

In group-support augmentation, the selected index $\lambda^{(j)}$ is appended to the support index set $\mathbf{\Psi}^{(j-1)}$, and the corresponding dictionary entries $\mathbf{G}_{:,\lambda^{(j)},:}$ are added to the support set $\mathbf{\Lambda}^{(j-1)}$. To prevent reselection in later iterations, these entries are set to zero in the dictionary.

In group-orthogonal projection, we estimate the multidimensional residual across all time frames and frequency bins. The multidimensional residual at the t^{th} time frame and f^{th} frequency bin for the j^{th} iteration is computed as follows:

$$\mathbf{R}_{:,t,f}^{(j)} = \mathbf{P}_{:,t,f} - \mathbf{\Lambda}_{:,:,f}^{(j)} \mathbf{\Lambda}_{:,:,f}^{(j)\dagger} \mathbf{P}_{:,t,f}, \tag{15}$$

where $\mathbf{\Lambda}_{:,:,f}^{(j)}$ denotes the support set at the f^{th} frequency bin for the j^{th} iteration. The algorithm terminates when either the number of iterations reaches J or the residual is below a predefined threshold ϵ.

The procedure of the proposed MD-OMP is outlined in Algorithm 1. In the context of the source localization problem discussed in this paper, MD-OMP identifies the candidate grid position $\mathbf{y}_n$ with the highest joint correlation in each step and removes its contribution from the measurements to calculate the residual. This process allows MD-OMP to detect even sources with very low sound levels in subsequent iterations.

In each iteration, the group-atom selection requires $\mathcal{O}(MNTF)$ for correlation computation, while the group-orthogonal projection requires $\mathcal{O}(MJTF)$ for the least-squares projection. Assuming that J is known, the overall computational complexity of MD-OMP is

Algorithm 1 Proposed MD-OMP

Input: $\mathbf{P}$, $\mathbf{G}$, sparsity J
Output: estimated source positions $\hat{\mathbf{Y}}$, estimated sparse coefficients S, estimated non-zero indices $\mathbf{\Psi}^{(J)}$

```
Initialize: Λ^(0) = ∅, Ψ^(0) = ∅, R^(0) = P, j = 1
while stopping criterion is not met do
    z_n ← Σ_{f=1}^{F} Σ_{t=1}^{T} |⟨R^(j−1)_{:,t,f}, G_{:,n,f}⟩|
    λ^(j) ← arg max_n(z_n)
    Ψ^(j) ← Ψ^(j−1) ∪ {λ^(j)}
    Λ^(j) ← Λ^(j−1) ∪ {G_{:,λ^(j),:}}
    G_{:,λ^(j),:} ← 0
    for t = 1 to T do
        for f = 1 to F do
            R^(j)_{:,t,f} = P_{:,t,f} − Λ^(j)_{:,:,f} Λ^(j)†_{:,:,f} P_{:,t,f}
        end for
    end for
    j ← j + 1
end while
Ŝ_{Ψ^(J),t,f} ← Λ^(J)†_{:,:,f} P_{:,t,f}
Ŷ ← { y_n | n ∈ Ψ^(J) }
```

$\mathcal{O}\big(MNTFJ + MJ^2TF\big)$. In typical source localization scenarios, where N is much larger than M and J, the term $\mathcal{O}\big(MNTFJ\big)$ dominates the overall complexity.

B. Generalized Multi-Domain Framework

In this subsection, we extend the MD-OMP framework to accommodate an arbitrary number of multidimensional domains and allow the sparsity adjustments in each domain based on prior knowledge, thereby offering flexibility for diverse acoustic scenarios.

Let us consider a more complicated scenario where the measurements are represented by an $(A+1)$-dimensional matrix $\mathbf{P} \in \mathbb{C}^{M\times L_1\times L_2\times\cdots\times L_A}$. Correspondingly, the dictionary matrix $\mathbf{G}$ is a $(B+2)$-dimensional matrix, where $\mathbf{G} \in \mathbb{C}^{M\times N\times L_1\times L_2\times\cdots\times L_B}$. We denote these domains by two sets, $\mathbf{A} = \{L_1, L_2, \cdots, L_A\}$ for the measurements and $\mathbf{B} = \{L_1, L_2, \cdots, L_B\}$ for the dictionary, where the Green's functions vary with values in the $\mathbf{B}$ and spatial domains.

We then define a domain set $\mathbf{C} = \mathbf{A} \cup \mathbf{B}$, and the sparse coefficients are denoted as a $(C+1)$-dimensional matrix $\mathbf{S} \in \mathbb{C}^{N\times L_1\times L_2\times\cdots\times L_C}$. Accordingly, the sparse optimization function for this generalized model is expressed as:

$$\min_{\mathbf{S}} \frac{1}{2}\left\|\mathbf{P} - \mathbf{G}\times\mathbf{S}\right\|_2^2 + \mu\mathcal{J}_{1,\mathfrak{p}_1,\mathfrak{p}_2,\cdots,\mathfrak{p}_C}(\mathbf{S}), \tag{16}$$

where $(1, \mathfrak{p}_1, \mathfrak{p}_2, \cdots, \mathfrak{p}_C)$ denote the norms of spatial and $L_1, L_2, \cdots, L_C$ domains. If domain L_c is sparse, where $c \in \{1, 2, \cdots, C\}$, we set the corresponding norm $\mathfrak{p}_c = 1$. Otherwise, we set $\mathfrak{p}_c = 2$.

Suppose D domains are sparse, denoted as $\mathfrak{p}_1 = \mathfrak{p}_2 = \cdots = \mathfrak{p}_D = 1$, and E domains are non-sparse, denoted as $\mathfrak{p}_1 = \mathfrak{p}_2 = \cdots = \mathfrak{p}_E = 2$. We define two subsets $\mathbf{D} = \{L_1, L_2, \cdots, L_D\} \subseteq \mathbf{C}$ and $\mathbf{E} = \{L_1, L_2, \cdots, L_E\} \subseteq \mathbf{C}$, where $\mathbf{D} \cap \mathbf{E} = \varnothing$.

The first step in this generalized multidimensional OMP involves calculating the joint correlation across all E dense domains:

$$z_{n,l_1,l_2,\cdots,l_D} = \sum_{l_1}^{L_1}\cdots\sum_{l_E}^{L_E}|\langle\mathbf{R}^{(j-1)}_{:,l_1,l_2,\cdots,l_A}, \mathbf{G}_{:,n,l_1,l_2,\cdots,l_B}\rangle|, \tag{17}$$

where l_i denotes the element within the L_i domain.

Upon obtaining the joint correlation, we identify the index with the maximum joint correlation across all D sparse domains for group-atom selection:

$$\lambda^{(j)} = \arg\max_{n,l_1,l_2,\cdots,l_D}(z_{n,l_1,l_2,\cdots,l_D}). \tag{18}$$

The algorithm continues similarly to MD-OMP in Subsection IV-A, ensuring robust adaptation to multidimensional data.

C. Case Studies of the Generalized MD-OMP

To illustrate the flexibility of the generalized MD-OMP framework, we discuss three representative acoustic scenarios. Detailed evaluation and analysis are presented for Case 1 in Subsection V-F, whereas Cases 2 and 3 are included as conceptual illustrations to show the general applicability of the method.

1) Case 1: Dominant Frequency Component Sources: We consider reverberant environments where target sound sources consist of only a few dominant frequency components, such as sinusoidal, tonal, or harmonic signals. In such cases, the source distribution in frequency domain is expected to be sparse. Assuming this prior knowledge, we set $D = 1$ including frequency domain and $E = 1$ including time domain.

The joint correlation is then computed as

$$z_{n,f} = \sum_{t=1}^{T}|\langle\mathbf{R}^{(j-1)}_{:,t,f}, \mathbf{G}_{:,n,f}\rangle|. \tag{19}$$

By selecting the index that maximizes $z_{n,f}$ instead of z_n, the algorithm can more effectively localize sources with dominant frequency components.

2) Case 2: Higher-Order Sources: Real-world sound sources often exhibit directional patterns, particularly in the near-field. In such cases, higher-order spherical harmonics (SH) can be exploited to model the transfer function from a source at $\mathbf{y}_n$ to a microphone at $\mathbf{x}_m$ as:

$$G(f, \mathbf{x}_m, \mathbf{y}_n) = \sum_{\gamma=0}^{\Gamma}\sum_{\alpha=-\gamma}^{\gamma} b_{\gamma,\alpha}h_\gamma^{(2)}(k_f\left\|\mathbf{x}_m - \mathbf{y}_n\right\|)Y_\gamma^\alpha(\theta_{m,n}, \phi_{m,n}), \tag{20}$$

where $b_{\gamma,\alpha}$ is the coefficient of the higher-order sources and $h_\gamma^{(2)}(\cdot)$ is the spherical Hankel function of the second kind, $Y_\gamma^\alpha(\cdot)$ is the SH function, and $\theta_{m,n}$ and $\phi_{m,n}$ are the angular components based on the relative position of $\mathbf{x}_m$ and $\mathbf{y}_n$.

Building on this, the signal model can be extended to include the SH domain, leading to $A = 2$, $B = 2$, $C = 3$. The parameter E is set to 3 to reflect the non-sparse characteristics of source distributions in the SH domain.

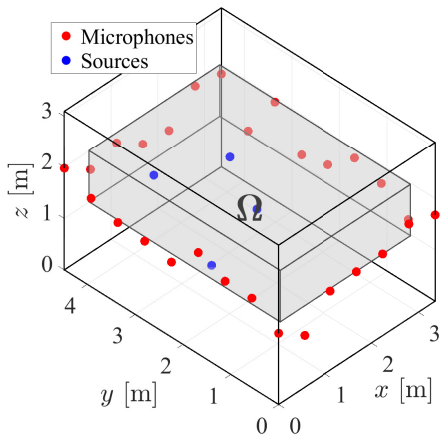

Fig. 2. Schematic of the simulation setup. The simulation is conducted in a rectangular reverberant room. Microphones are placed along the surrounding walls, and sources are randomly positioned within the ROI Ω.

3) Case 3: Moving Microphones: Consider a scenario where measurements are obtained from several participants wearing smart glasses equipped with microphones. In this case, participants may be stationary, moving, crouching, or jumping, implying that the positions of the microphones $\mathbf{x}_m$ may change over time. Consequently, the Green's function varies across spatial, time, and frequency domains, resulting in $A = B = C = 2$.

In this context, microphone motion introduces dynamic spatial sampling, mitigating localization issues related to spatial aliasing or sparse measurements in fixed microphone arrays.

V. Simulation Analysis

This section evaluates the localization performance of the proposed MD-OMP through numerical simulations. Based on Section II, we simulate a scenario involving multiple participants conversing in a conference room equipped with a surrounding microphone array. The proposed method is validated using Monte Carlo simulations under various acoustic conditions.

A. Numerical Simulation Settings

We utilize the Room Impulse Response (RIR) generator toolbox [66], [67], based on the image-source model, to simulate a $3.36 \times 4.48 \times 4.1$ m reverberant shoebox room. The origin of the coordinate system is set at the front-left-bottom corner of the room. The sampling frequency is set at 16 kHz, the image source order is set at 20, and Gaussian white noise is added to the microphone signals, resulting in an SNR of 10 dB.

An overview of the simulation setup is provided in Fig. 2. As illustrated, $M = 28$ microphones are distributed along the room walls, forming a perimeter array with 7 microphones on each wall parallel to the x-axis and 9 on each wall parallel to the y-axis, at heights ranging from 1.0 m to 2.0 m. $J = 4$ sources are randomly positioned within the ROI Ω, defined by $0.28 < x < 3.08$ m, $0.24 < y < 4.24$ m, and $1.3 < z < 1.7$ m. We assume the sources are speech signals, with two male and two female voices selected from the MS-SNSD dataset [68].

Each source signal has a duration of 10 seconds and is normalized to have equal power. We select a 2-second segment from the 4^{th} to 6^{th} second of the recordings. For STFT analysis, a periodic Hamming window of length 1024 samples with a 50% overlap is applied, and the FFT length is set to 2048. The frequency range $125 < f < 2000$ Hz is chosen for localization, except in Subsection V-E, as it captures the majority of speech energy while minimizing the effects of low-frequency noise and high-frequency sensitivity to environmental factors. Although the full frequency range produces $F = 241$ frequency bins, we uniformly select $F = 61$ bins to reduce computational cost. Additionally, $T = 61$ time frames are generated for temporal analysis.

The evaluation of the proposed method's localization performance is conducted under various acoustic conditions, including differing reverberation times, candidate grid densities, frequency ranges, source signals, and room dimensions. For each scenario, $J = 4$ sources are randomly positioned within the ROI over $H = 100$ Monte Carlo simulations. We define the averaged success rate $\bar{\rho}$ for evaluation as:

$$\bar{\rho} = \frac{1}{HJ} \sum_{h=1}^{H} \sum_{j=1}^{J} \rho_j^{(h)}, \tag{21}$$

where $\rho_j^{(h)} = 1$ if the estimated source position $\hat{\mathbf{y}}_j$ is within 0.2 m of the true source position $\mathbf{y}_j$, i.e., $\|\hat{\mathbf{y}}_j - \mathbf{y}_j\|_2 < 0.2$, and $\rho_j^{(h)} = 0$ otherwise.

Consequently, $\bar{\rho}$ ranges between 0 and 1 and serves as a quantitative measure of localization accuracy, with $\bar{\rho} = 1$ denoting perfect localization.

B. Compared Methods

We compare the performance of the proposed MD-OMP with five methods: Wideband OMP (WB-OMP), G-IRLS [56], MD-SBL [60], NF-SRP-PHAT [34], [35], and IR-SRP-PHAT [37].

WB-OMP, a variant of MD-OMP, employs the same methodology but is limited to considering only frequency and spatial domains, with measurements averaged over time frames. The purpose of WB-OMP is to provide a benchmark for evaluating the improvement in localization performance achieved by MD-OMP when multiple time frames are utilized.

For G-IRLS, we set the norm $\mathsf{p} = 1$, the maximum number of iterations to 20, and the minimum ℓ_2-norm change in $\hat{\mathbf{S}}$ after an iteration to 10^{-4}, following the parameter settings in [56]. MD-SBL is configured with a threshold of 10^{-4}, an update exponent in the fixed-point iteration of 1, and a convergence parameter $\epsilon = 10^{-4}$. For IR-SRP-PHAT, the SNR threshold and the phase-distance threshold are set to $\lambda_{\text{TH}} = 5$ dB and $\rho_{\text{TH}} = 0.5$ rad, respectively, in accordance with [37]. In our simulations, we set the minimum-distance merging threshold as 0.2 m and the coherence threshold as 0.4, as these adjustments yield improved performance under our testing conditions.

To simplify evaluation, we assume the number of sources J is known. Accordingly, the iterations of MD-OMP, WB-OMP, and IR-SRP-PHAT are terminated after J iterations. For G-IRLS, MD-SBL, and NF-SRP-PHAT, the J largest values in the averaged source weights $\bar{\mathbf{S}} = \mathbb{E}_{T,F}[\hat{\mathbf{S}}]$, where $\mathbb{E}_{T,F}[\cdot]$ is the

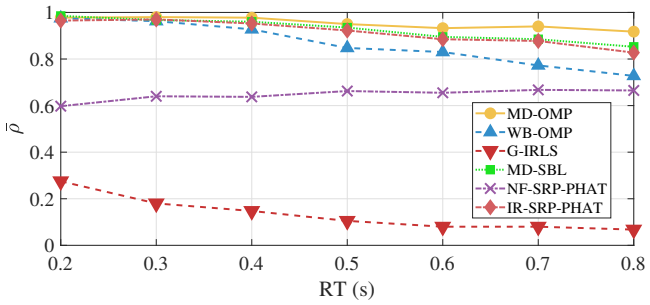

Fig. 3. Success rates for varying reverberation times from 0.2 to 0.8 s averaged over 100 Monte Carlo tests.

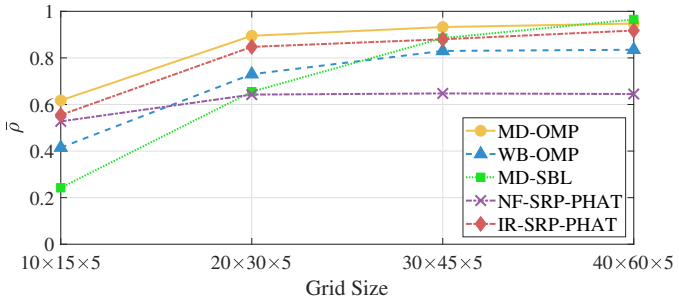

Fig. 4. Success rates for varying grid sizes averaged over 100 Monte Carlo tests.

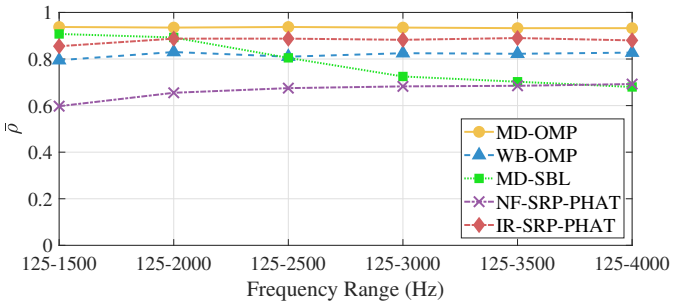

Fig. 5. Success rates for varying frequency ranges averaged over 100 Monte Carlo tests.

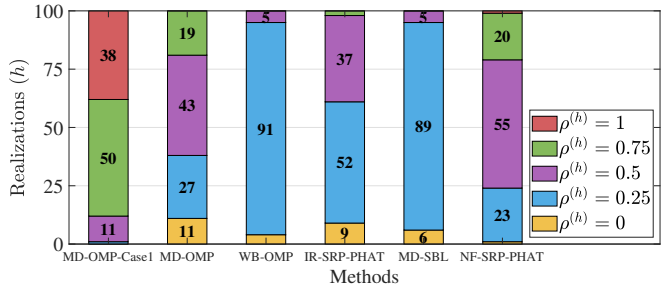

Fig. 6. Success rates for monotone signal localization, based on 100 Monte Carlo tests, where $\rho^{(h)} = \frac{1}{J}\sum_{j=1}^{J}\rho_j^{(h)}$.

expectation over time and frequency, are selected using Non-Maximum Suppression (NMS) [69] with a 0.2 m threshold to ensure spatial separation. MD-OMP and WB-OMP do not require NMS due to the orthogonal projection step.

Notably, none of the methods requires prior information, except for the relative microphone positions and the number of sources.

C. Performance Evaluation over Varying Reverberation Time

First, we aim to investigate the localization performance under different reverberant environments. We control the room reflection coefficients in the image source model such that the reverberation time (RT) ranges from 0.2 s to 0.8 s. We consider $N = 6750$ uniform candidate grid points in the ROI, 30-by-45-by-5 along the $x-y-z$ axes.

Figure 3 illustrates the success rates achieved by each method under comparison. The proposed MD-OMP consistently maintains a success rate exceeding 0.9 across all RTs, demonstrating its robustness. The success rate of WB-OMP remains above 0.8 for RTs up to 0.4 s; however, it declines dramatically as the RT increases. MD-SBL and IR-SRP-PHAT maintain a success rate above 0.8 across all RTs, although their overall performance is slightly inferior to MD-OMP, particularly for RT $\geq$ 0.6 s. Conversely, NF-SRP-PHAT struggles across all conditions, with success rates consistently below 0.7.

G-IRLS fails to achieve reliable localization under any RT and is therefore excluded from further evaluations due to its poor performance.

D. Performance Evaluation over Varying Candidate Grid Densities

Second, we compare localization performance across different candidate grid densities, considering four grid configurations: $10\times15\times5$, $20\times30\times5$, $30\times45\times5$, $40\times60\times5$ along the $x-$, $y-$, and $z-$axes in the ROI. The RT is fixed at 0.6 s.

The results, shown in Fig. 4, reveal that the lowest-resolution grid, $10\times15\times5$, results in poor success rates across all methods, indicating insufficient spatial resolution for accurate localization.

MD-SBL, initially the weakest performer at low densities, significantly improves with finer grids, achieving near-perfect accuracy at the highest resolution. However, its strong dependence on grid density suggests limitations in resource-constrained scenarios. In contrast, MD-OMP and IR-SRP-PHAT show robustness across all but the coarsest grid, achieving success rates above 0.8 for the remaining configurations.

E. Performance Evaluation over Varying Frequency Ranges

Next, we analyze the impact of different frequency ranges on localization performance. We fix the lower bound at 125 Hz and vary the upper bound from 1500 Hz to 4000 Hz. The candidate grid configuration remains $30\times45\times5$, and the RT is set to 0.6 s.

Figure 5 shows that the success rates of MD-OMP, WB-OMP, NF-SRP-PHAT, and IR-SRP-PHAT are stable across different frequency ranges, with MD-OMP consistently achieving the highest success rate.

Conversely, MD-SBL demonstrates a strong dependence on the selected frequency range. While its success rate is approximately 0.9 in the $125-1500$ Hz range, it steadily declines as the upper bound increases, dropping below 0.7 at $125-4000$ Hz. This result highlights the importance of carefully selecting the frequency range when applying MD-SBL in different scenarios.

F. Performance Evaluation of Dominant Frequency Component Sources

In Subsection IV-B, we introduced multiple case studies that demonstrate the flexibility of MD-OMP. To investigate the

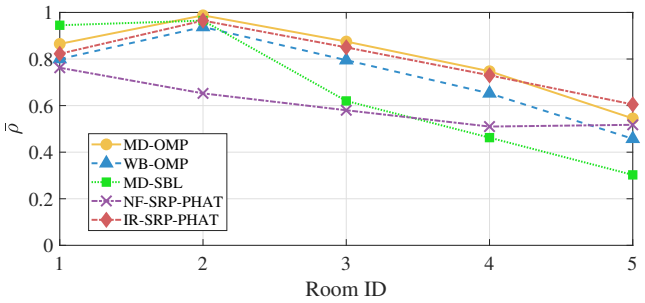

Fig. 7. Success rates for varying room dimensions averaged over 100 Monte Carlo tests. Room IDs correspond to increasing room sizes from Room 1, the smallest, to Room 5, the largest.

performance of the generalized formulation, we select Case 1: Dominant Frequency Component Sources for evaluation.

For this validation, we replace speech signals with monotone signals at 800 Hz, 900 Hz, 1000 Hz, and 1100 Hz for each respective source. We set the RT to 0.2 s and the candidate grid configuration as $30\times45\times5$. Based on Subsection IV-B for Case 1, we introduce an MD-OMP variant, termed MD-OMP-Case1, which assumes source signals sparsely distributed in frequency domain.

Figure 6 presents the success rates of all evaluated methods. For WB-OMP and MD-SBL, the success rate remains at 0.25 in most realizations, indicating that only one source is typically localized. IR-SRP-PHAT performs slightly better, achieving a success rate of 0.5 in 37 realizations, but still struggles to localize three or more sources. MD-OMP and NF-SRP-PHAT show improved performance, reaching success rates of 0.75 in 20 realizations. However, successful localization of all four sources is rarely achieved by any method.

In contrast, MD-OMP-Case1 achieves a success rate of 1.0 in 38 realizations and 0.75 in 50 realizations, demonstrating its effectiveness in identifying dominant frequency component sources where other methods struggle.

G. Performance Evaluation over Varying Room Dimensions

To analyze the influence of room dimensions while maintaining the same number of microphones, we consider five shoebox rooms with dimensions:

$$\begin{aligned}
&\text{Room 1: } 1.68 \times 2.24 \times 4.1 \text{ m},\\
&\text{Room 2: } 3.36 \times 4.48 \times 4.1 \text{ m},\\
&\text{Room 3: } 5.04 \times 6.72 \times 4.1 \text{ m},\\
&\text{Room 4: } 6.72 \times 8.96 \times 4.1 \text{ m},\\
&\text{Room 5: } 8.40 \times 11.20 \times 4.1 \text{ m}.
\end{aligned}$$

The RT is fixed at 0.4 s for all rooms. Both the microphone array ($M = 28$) and the candidate grid ($N = 6750$) remain fixed in number but are scaled proportionally with the room dimensions, such that the microphone spacing, grid spacing, and ROI scale consistently with room size.

As shown in Fig. 7, MD-SBL achieves near-perfect performance in the smallest room but degrades most rapidly as the room size increases, indicating high sensitivity to room geometry. In contrast, MD-OMP and IR-SRP-PHAT exhibit stronger robustness. MD-OMP attains the highest success rates in Room 2, Room 3, and Room 4, while IR-SRP-PHAT surpasses MD-OMP in the largest room. Nevertheless, the success rates of all methods drop below 0.7 in Room 5, demonstrating that $M = 28$ microphones provide insufficient spatial sampling for such a large room.

Fig. 8. Photograph of the experimental setup. The experiment was conducted in an office room with reflective surfaces and furniture. 14 microphones were positioned along the room perimeter, and 4 loudspeakers were placed within the central region for source localization evaluation.

VI. Practical Experiments

In this section, we investigate the localization performance of MD-OMP using real recordings conducted in a practical environment.

A. Experimental Settings

We experimentally examine the localization performance by identifying multiple loudspeaker positions in a reverberant office room at the Australian National University. The room dimensions are $4.35 \times 3 \times 3$ m, with an RT of approximately 730 ms, measured using T_{20}[1]. The room contains typical office furniture, including chairs, tables, and boxes, which introduce scattering effects and serve as environmental noise, enhancing the realism of the test scenario.

We place a total of $M = 14$ condenser microphones and $J = 4$ loudspeakers as shown in Fig. 8. The configuration details of the microphones and loudspeakers in Cartesian coordinates are provided in Table I. All measurements are made with respect to the origin, which is defined as the center of the room. It is important to note that the ground-truth positions of both the microphones and loudspeakers were determined manually and are, therefore, subject to potential human error. The loudspeakers played the same normalized speech signals as in Section V. We define the ROI as $-1.20 < x < 1.20$ m, $-1.00 < y < 1.00$ m, and $-0.2 < z < 0.2$ m. We consider $N = 4500$ uniform candidate grid points in the ROI, 30-by-30-by-5 along the $x-y-z$ axes. The sampling frequency of the recordings is 16 kHz. Other settings are consistent with those in Section V, including STFT parameters, the selected recording duration, and the selected frequency bins, resulting in $F = 61$ frequency bins and $T = 61$ time frames.

We compare MD-OMP with WB-OMP, MD-SBL, and IR-SRP-PHAT. All these methods use the same parameter settings as in Section V.

[1]As defined in ISO 3382-2:2008, T_{20} is the reverberation time extrapolated from the -5 to -25 dB decay slope to a 60 dB decay.

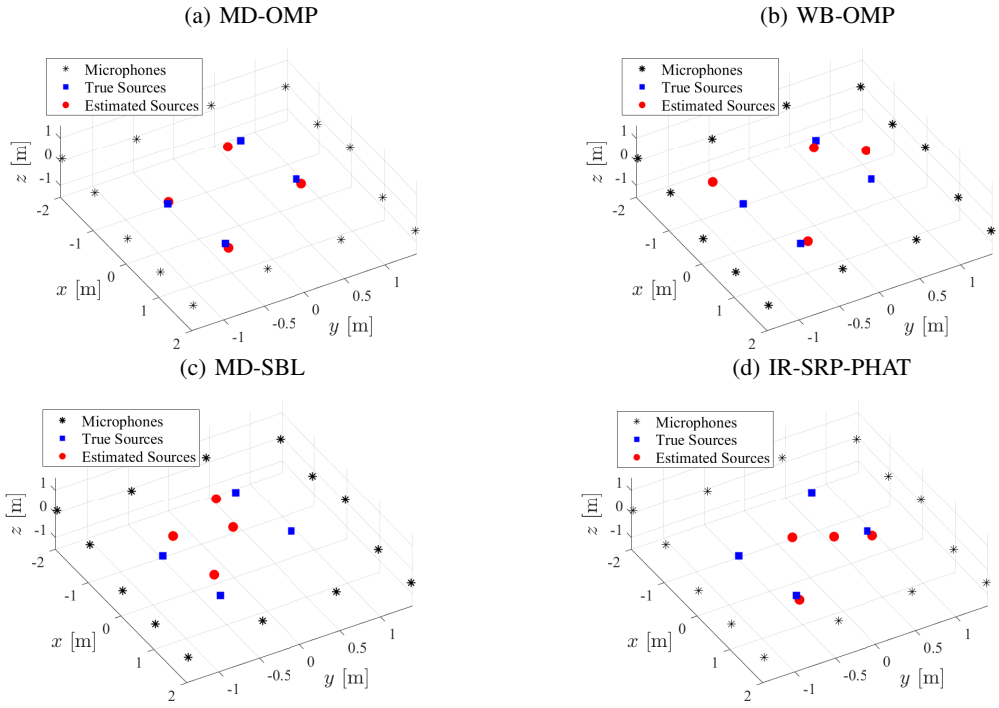

Fig. 9. Results of the four localization methods for the experiment conducted in an office room. Each subfigure depicts the spatial distribution of microphones, true sources, and estimated sources.

TABLE I
COORDINATES OF MICROPHONES AND LOUDSPEAKERS WITH RESPECT TO THE ROOM CENTER (IN METERS)

Type	ID	x (m)	y (m)	z (m)
Microphone	$\mathbf{x}_1$	-2.00	-1.38	0.10
	$\mathbf{x}_2$	-1.00	-1.38	0.05
	$\mathbf{x}_3$	0.00	-1.38	-0.46
	$\mathbf{x}_4$	1.00	-1.38	-0.46
	$\mathbf{x}_5$	2.00	-1.38	-0.46
	$\mathbf{x}_6$	2.00	-0.46	0.01
	$\mathbf{x}_7$	2.00	0.46	0.17
	$\mathbf{x}_8$	2.00	1.38	-0.52
	$\mathbf{x}_9$	1.00	1.38	-0.52
	$\mathbf{x}_{10}$	0.00	1.38	0.19
	$\mathbf{x}_{11}$	-1.00	1.38	-0.22
	$\mathbf{x}_{12}$	-2.00	1.38	-0.08
	$\mathbf{x}_{13}$	-2.00	0.46	0.22
	$\mathbf{x}_{14}$	-2.00	-0.46	-0.12
Loudspeaker	$\mathbf{y}_1$	-1.00	0.42	0.20
	$\mathbf{y}_2$	-0.30	-0.76	-0.15
	$\mathbf{y}_3$	1.10	-0.62	0.00
	$\mathbf{y}_4$	0.26	0.60	0.12

B. Experimental Results

Figure 9 illustrates the localization results for the proposed methods. In Fig. 9(a), MD-OMP successfully identifies four sources, demonstrating its robustness in practical reverberant environments. Conversely, as shown in Fig. 9(b) and (d), WB-OMP and IR-SRP-PHAT only detect two sources, while the remaining two sources are localized inaccurately. MD-SBL in Fig. 9(c) fails to localize all sources, though three estimated positions are near the true positions but not close enough for correct identification.

VII. DISCUSSION

This section provides a detailed discussion of the proposed MD-OMP in comparison with several baselines. The analysis begins with localization performance in both simulations and experiments, followed by a computational time evaluation.

Across different RTs, as shown in Fig. 3, MD-OMP consistently achieves success rates above 0.9. In practical experiments within a furnished office with strong scattering as shown in Fig. 9, the method also localizes sources accurately. The proposed approach considers reverberation as part of the noise term, so localization under strong reverberation is equivalent to operating at low SNR. For instance, at RT $= 0.8\,$s where the SNR is about $0\,$dB, MD-OMP remains robust to both reverberation and noise.

WB-OMP performs comparably to MD-OMP when RT $\leq 0.4\,$s, but its performance degrades substantially as RT increases. The difference lies in group-atom selection. MD-OMP integrates spatio-temporal-spectral information to suppress reflections, whereas WB-OMP considers only spatio-spectral information and is more sensitive to reverberation.

When sources contain dominant frequency components, as shown in Fig. 6, the variant MD-OMP-Case1 achieves success rates of 1.0 or 0.75 in most realizations, outperforming other baselines. These results show that incorporating frequency-domain sparsity in the MD-OMP framework improves performance for sparse frequency component sources, highlighting the importance of designing adaptive models for different acoustic scenarios.

MD-SBL, although widely regarded as a representative method for DOA estimation, is unstable in source localization and strongly affected by the environment. For example, it ranks highest with the finest candidate grid but lowest with the coarsest, and performs best in the smallest room but worst in the largest, as shown in Fig. 4 and Fig. 7. In contrast, MD-OMP, WB-OMP and IR-SRP-PHAT maintain relatively robust success rates. This stability is explained by the iterative source contribution removal mechanism in these methods,

TABLE II
AVERAGE RUNTIMES UNDER THE SIMULATION SETUP.

Method	Avg. runtime (s)
MD-OMP	0.5608
WB-OMP	0.0604
NF-SRP-PHAT	0.2187
IR-SRP-PHAT	4.0413
G-IRLS	108.5550
MD-SBL	18.7229

which prevents repeated selection of the same dominant source position and is particularly useful in near-field scenarios where sources are located within the microphone array.

IR-SRP-PHAT is generally the strongest baseline after MD-OMP across most acoustic conditions, although there are notable exceptions. In Fig. 7, it outperforms MD-OMP in the largest room. Fig. 9 indicates that in practical experiments it detects only two sources. As shown in Fig. 6, its performance degrades in scenarios with dominant frequency components.

We further evaluate the computational time of all algorithms under the simulation setup described in Subsection V-C, with $\mathrm{RT} = 0.6\,\mathrm{s}$. The average computational time is computed over 100 realizations, with a one-time warm-up performed on the first realization. Evaluations are conducted on a CPU (Intel Core i7-12700, 12 cores) with 9.0 GB RAM using MATLAB R2024a, and execution time is measured via MATLAB's `tic`/`toc` functions. The results are summarized in Table II. It is worth noting that NF-SRP-PHAT[2]and MD-SBL [61] follow publicly available implementations, whereas the other algorithms are implemented for this study. Our proposed method utilizes MATLAB's `pagemtimes` function to compute joint correlations, thereby accelerating the computation. Consequently, the reported computational times should be interpreted as relative comparisons rather than absolute benchmarks.

VIII. CONCLUSION

In this paper, we propose the Multi-Dictionary Orthogonal Matching Pursuit (MD-OMP) algorithm, which enhances sound source localization by introducing group sparsity across time, frequency, and spatial domains. We evaluate MD-OMP through extensive simulations and real-room experiments under various acoustic conditions, comparing it with representative group-sparsity and TDOA-based methods. The results show that MD-OMP achieves higher accuracy and robustness, even in environments with reverberation and sparse microphone configurations. Furthermore, we showcase the algorithm's flexibility in handling multidimensional inputs, enabling its application to diverse scenarios. In future work, we aim to extend MD-OMP to dynamic source modeling and real-time applications.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[2]https://github.com/WenzheLiu-Speech/sound-source-localization-algorithm_DOA_estimation

[2] D. Y. Hu, H. B. Li, Y. Hu, and Y. Fang, "Sound field reconstruction with sparse sampling and the equivalent source method," *Mech. Syst. Signal Process.*, vol. 108, pp. 317–325, 2018.

[3] S. Xu, J. A. Zhang, T. D. Abhayapala, A. Bastine, W. T. Lai, and P. N. Samarasinghe, "Sparse sound field representation using complex orthogonal matching pursuit," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1336–1340.

[4] Q. Wang, J. Du, H. X. Wu, J. Pan, F. Ma, and C. H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1251–1264, 2023.

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

[6] J. C. Mosher and R. M. Leahy, "Source localization using recursively applied and projected (RAP) MUSIC," *IEEE Trans. Signal Process.*, vol. 47, no. 2, pp. 332–340, 1999.

[7] B. Friedlander, "The root-MUSIC algorithm for direction finding with interpolated arrays," *Signal Process.*, vol. 30, no. 1, pp. 15–29, 1993.

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

[9] S. Shahbazpanahi, S. Valaee, and M. H. Bastani, "Distributed source localization using ESPRIT algorithm," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2169–2178, 2001.

[10] A. Hu, T. Lv, H. Gao, Z. Zhang, and S. Yang, "An ESPRIT-based approach for 2-D localization of incoherently distributed sources in massive MIMO systems," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 996–1011, 2014.

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

[12] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508–519, 2022.

[13] 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.

[14] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[15] 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.

[16] E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor, "Steered response power for sound source localization: A tutorial review," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, p. 59, 2024.

[17] P.-A. Gauthier, P. Lecomte, and A. Berry, "Source sparsity control of sound field reproduction using the elastic-net and the LASSO minimizers," *J. Acoust. Soc. Amer.*, vol. 141, no. 4, pp. 2315–2326, 2017.

[18] A. Panahi and M. Viberg, "Fast LASSO based DOA tracking," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, 2011, pp. 397–400.

[19] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun, pp. 211–244, 2001.

[20] A. Xenaki, J. Bünsow Boldt, and M. G. Christensen, "Sound source localization and speech enhancement with sparse Bayesian learning beamforming," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3912–3921, 2018.

[21] 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.

[22] W. T. Lai, L. Birnie, X. Chen, A. Bastine, T. D. Abhayapala, and P. N. Samarasinghe, "Source localization by multidimensional steered response power mapping with sparse Bayesian learning," in *Int. Workshop Acoust. Signal Enhanc.*, 2024, pp. 31–35.

[23] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[24] M. Emadi, E. Miandji, and J. Unger, "OMP-based DOA estimation performance analysis," *Digit. Signal Process.*, vol. 79, pp. 57–65, 2018.

[25] A. Aich and P. Palanisamy, "On application of OMP and CoSaMP algorithms for DOA estimation problem," in *Proc. IEEE Int. Conf. Commun. Signal Process.*, 2017, pp. 1983–1987.

[26] 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.*, 2017, pp. 136–140.

[27] 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.*, 2024, pp. 1321–1325.

[28] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2814–2818.

[29] Y. Shul and J.-W. Choi, "CST-Former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 8686–8690.

[30] X. Zhang, W. Chen, W. Zheng, Z. Xia, and Y. Wang, "Localization of near-field sources: A reduced-dimension MUSIC algorithm," *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1422–1425, 2018.

[31] B. Wang, Y. Zhao, and J. Liu, "Mixed-order MUSIC algorithm for localization of far-field and near-field sources," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 311–314, 2013.

[32] L. I. Birnie, T. D. Abhayapala, and P. N. Samarasinghe, "Reflection assisted sound source localization through a harmonic domain music framework," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 279–293, 2019.

[33] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, F. M. Gonçalves, A. Said, and B. e. a. 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.

[34] H. Do and H. F. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 301–304.

[35] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 134, no. 4, pp. 2627–2630, 2013.

[36] X. Dang, W. Ma, E. A. P. Habets, and H. Zhu, "TDOA-based robust sound source localization with sparse regularization in wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1108–1123, 2022.

[37] X. Dang and H. Zhu, "An iteratively reweighted steered response power approach to multisource localization using a distributed microphone network," *J. Acoust. Soc. Amer.*, vol. 155, no. 2, pp. 1182–1197, 02 2024.

[38] Y. Zhao, Y. He, H. Chen, Z. Zhang, and Z. Xu, "Three-dimensional grid-free sound source localization method based on deep learning," *Appl. Acoust.*, vol. 227, p. 110261, 2025.

[39] A. Kujawski, G. Herold, and E. Sarradj, "A deep learning method for grid-free localization and quantification of sound sources," *J. Acoust. Soc. Amer.*, vol. 146, no. 3, pp. EL225–EL231, 2019.

[40] A. Kujawski and E. Sarradj, "Fast grid-free strength mapping of multiple sound sources from microphone array data using a Transformer architecture," *J. Acoust. Soc. Amer.*, vol. 152, no. 5, pp. 2543–2556, 2022.

[41] X. Dang, A. Herzog, S. R. Chetupalli, E. A. P. Habets, and H. Liu, "SepLocNet: Multi-speaker localization with separation-guided TDOA estimation in wireless acoustic sensor networks," *Appl. Acoust.*, vol. 231, p. 110488, 2025.

[42] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Lang.*, vol. 11, no. 2, pp. 91–126, 1997.

[43] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[44] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, 1987.

[45] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, 1987.

[46] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, 2001.

[47] L. M. Kaplan, Q. Le, and N. Molnar, "Maximum likelihood methods for bearings-only target localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, 2001, pp. 3001–3004.

[48] M. Swartling, B. Sällberg, and N. Grbić, "Source localization for multiple speech sources using low complexity non-parametric source separation and clustering," *Signal Process.*, vol. 91, no. 8, pp. 1781–1788, 2011.

[49] M. Compagnoni, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 1964–1975, 2012.

[50] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, "Localizing multiple audio sources in a wireless acoustic sensor network," *Signal Process.*, vol. 107, pp. 54–67, 2015.

[51] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation in wireless acoustic sensor networks using DOA estimates: The data-association problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 342–356, 2017.

[52] M. Hahmann, E. Fernandez-Grande, H. Gunawan, and P. Gerstoft, "Sound source localization using multiple ad hoc distributed microphone arrays," *JASA Express Letters*, vol. 2, no. 7, p. 074801, 2022.

[53] K. Aström, M. Larsson, G. Flood, and M. Oskarsson, "Extension of time-difference-of-arrival self calibration solutions using robust multilateration," in *Proc. Eur. Signal Process. Conf.*, Dublin, Ireland, 2021, pp. 870–874.

[54] J. Yang, X. Zhong, W. Chen, and W. Wang, "Multiple acoustic source localization in microphone array networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 334–347, 2021.

[55] Z. He, A. Cichocki, R. Zdunek, and S. Xie, "Improved FOCUSS method with conjugate gradient iterations," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 399–404, 2008.

[56] S. Koyama, N. Murata, and H. Saruwatari, "Sparse sound field decomposition for super-resolution in recording and reproduction," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3780–3795, 2018.

[57] G. Ping, E. Fernandez-Grande, P. Gerstoft, and Z. Chu, "Three-dimensional source localization using sparse Bayesian learning on a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 147, no. 6, pp. 3895–3904, 2020.

[58] G. Chardon, T. Nowakowski, J. de Rosny, and L. Daudet, "A blind dereverberation method for narrowband source localization," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 815–824, 2015.

[59] W. T. Lai, L. Birnie, T. Abhayapala, A. Bastine, S. Xu, and P. Samarasinghe, "A two-step approach for narrowband source localization in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops*, 2024, pp. 490–494.

[60] S. Nannuru, K. L. Gemba, P. Gerstoft, W. S. Hodgkiss, and C. F. Mecklenbräuker, "Sparse Bayesian learning with multiple dictionaries," *Signal Process.*, vol. 159, pp. 159–170, 2019.

[61] 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.

[62] S. Ganguly, I. Ghosh, R. Ranjan, J. Ghosh, P. K. Kumar, and M. Mukhopadhyay, "Compressive sensing based off-grid DOA estimation using OMP algorithm," in *Proc. IEEE Int. Conf. Signal Process. Integr. Netw.*, 2019, pp. 772–775.

[63] 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.

[64] J. Palacios, N. González-Prelcic, and C. Rusu, "Multidimensional orthogonal matching pursuit: Theory and application to high accuracy joint localization and communication at mmWave," *arXiv preprint arXiv:2208.11600*, 2022.

[65] M. A. Doron and E. Doron, "Wavefield modeling and array processing. i. spatial sampling," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2549–2559, 1994.

[66] E. A. P. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[67] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 04 1979.

[68] 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. Interspeech*, pp. 1816–1820, 2019.

[69] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 679–698, 1986.