Leveraging Deep Bayesian Inference for Efficient Multiple Source Localization with Spherical Harmonics

Priyadarshini Dwivedi, Gyanajyoti Routray*, and Rajesh M. Hegde
Indian Institute of Technology Kanpur, SRM University AP Amaravati*
Email: {priyadw, rhegde}@iitk.ac.in, gyanajyoti.r@srmap.edu.in*

***Abstract*—Accurately localizing multiple sound sources in noisy and reverberant environments poses significant challenges. To address these, a two-stage framework is developed that integrates a multi-label, multi-class convolutional neural network (M-CNN) with a Bayesian prior. In the first stage, the M-CNN classifies directions of arrival (DOAs) based on the spherical harmonics decomposition (SHD) of signals captured by a spherical microphone array (SMA). In the second stage, a sparse Bayesian learning (SBL) algorithm is applied to the SHD signals using a high-resolution search grid, corresponding to the DOA classes identified in the first stage, to accurately localize multiple acoustic sources. The M-CNN model demonstrates robustness in noisy and reverberant conditions, while the SBL algorithm excels in high-resolution localization of sparse sources. This hybrid model enhances DOA estimation up to 1° accuracy and reduces root mean square error (RMSE) by 80%. Extensive simulations and real-world experiments validate the accuracy and performance of the proposed method.**

***Keywords*—** DOA estimation, spherical harmonics domain, Bayesian machine learning, CNN

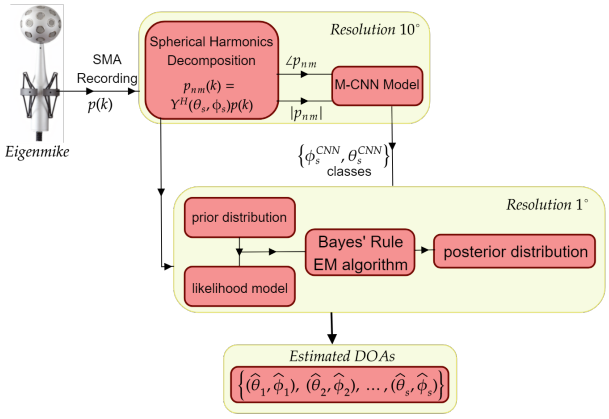

Fig. 1: Illustration of the proposed hybrid M-CNN-SBL framework for acoustic source localization in SH domain.

I. Introduction

Accurately localizing multiple acoustic sources is crucial across various applications, and as the complexity of environments with multiple simultaneous sources grows, the demand for robust and efficient localization techniques increases. Spherical microphone arrays (SMA) are ideal for 3D direction-of-arrival (DOA) estimation, as they capture sound sources with uniform spatial resolution in all directions [1], [2]. Furthermore, spherical harmonics (SH) representation has become a powerful approach for analyzing and processing sound fields, especially in three-dimensional environments. However, precise localization of multiple sources in this domain remains challenging, particularly when sources are closely spaced or overlapping. Learning-based methods have shown superior robustness and accuracy for acoustic source localization compared to traditional techniques, particularly in noisy and reverberant environments [3], [4]. However, precise DOA estimation leads to high complexity and demands for more training data, making generalization difficult. To mitigate this, recent research has emphasized reducing model complexity while preserving performance [5], [6].

Despite advancements, localizing multiple sources in reverberant environments remains challenging. Several methods have been proposed to address this issue. The Multiple Signal Classification (MUSIC) method in the SH domain [7] is a subspace-based approach for DOA estimation. K-means clustering of PIVs has been used to estimate multiple DOAs [8], while higher-order SHs refine DOAs through peak detection in histograms [9], [10]. A more noise-robust Subspace Pseudointensity Vector (SSPIV) formulation from the SH domain was introduced in [11]. Convolutional recurrent neural networks (CRNNs) were employed for multi-source localization using first-order Ambisonics, with improved performance in [12], [13]. CNN-based methods have also been used to learn modal coherence patterns from SH coefficients [14]. Recently, SH domain features like relative harmonic coefficients have been explored for higher-order arrays [15], and a decoupled relative harmonic coefficients (DRHC) DOA estimator reducing complexity in reverberant environments was proposed in [16]. However, these methods struggle in highly noisy and reverberant conditions with multiple sources.

In this work, a framework for the super-resolution DOA estimation in the SH domain has been proposed as represented in Fig. 1. The major contributions of the proposed two-stage hybrid model M-CNN-SBL method are:

- The proposed work introduces a two-stage method to localize multiple sound sources in highly reverberant and noisy environments using SH. In the first stage, M-CNN provides a coarse DOA estimation, which is then fine-tuned in second stage by a SBL method to achieve super-resolution.

• The model incorporates a probabilistic layer from the SBL method, which uses the output from the CNN as prior knowledge to enhance the overall estimation process with a statistically robust framework. This two-stage implementation offers the dual benefit of reducing computational complexity while simultaneously improving localization accuracy.
• A major advantage of the framework is its robustness; CNN is trained on diverse datasets, enabling it to generalize well to several acoustic conditions, and SBL leverages this capability for effective performance across different environments.

II. System Model and Problem Formulation

Consider a spherical sensor array with radius r_a and number of omnidirectional microphones M. The position of mth microphone of the array is $\mathbf{r}_m = (r_m, \theta_m, \phi_m)^T$, where $(\cdot)$ T denotes the transpose operator, and θ_m and ϕ_m represent the elevation and azimuth of the mth sensor, respectively. The sound scene consists of $\mathcal{S} > 1$, far-field sources and the angle of arrival of the sth plane wave is $\Omega_s == (\theta_s, \phi_s)$. $k = 2\pi f/c$ is the wavenumber corresponding to frequency f, c is the speed of sound, and θ_s and ϕ_s are the incident elevation and the azimuth angle, respectively. The sound pressure at the l^{th} microphone of the array is

$$p(k, r, \theta_l, \phi_l) = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_n^m(k) b_n(kr) Y_n^m(\theta_l, \phi_l)$$

where, the radial component $b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j_n^{'}(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr) \right]$ represent the mode strength. $a_n^m(k)$ is the incident sound field composed of plane waves given as $a_n^m(k) = [Y_n^m(\theta_s, \phi_s)]^* s(k)$. The radial function $b_n(kr)$ represents the mode strength, where $j_n(\cdot)$, and $h_n^{(2)}(\cdot)$ represents the order n spherical Bessel of first kind and Hankel function of second kind. The frequency f is represented in term of wave number $k = 2\pi f/c$, c being the velocity of sound. $(\cdot)'$ represents the derivative and $(\cdot)^*$ the complex conjugate. N represents the SMA order, n and m is the order and degree of the SH basis function. The SH basis function $Y_n^m(\theta, \phi)$ is expressed as [4]

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\ \theta) e^{im\phi} \quad (1)$$

where P_n^m is the associated Legendre function of order n and degree m. The total pressure captured by the SMA is written in matrix form given as

$$\mathbf{p}(k) = \mathbf{Y}(\Omega)\mathbf{B}(kr) [Y(\theta_s, \phi_s)]^* s(k) + \eta(k) \quad (2)$$

where $B(kr) = \mathrm{diag}[\, b_0(kr), b_1(kr), b_1(kr), \ldots, b_N(kr)]$ is matrix representing the mode strengths. $\eta(k)$ represents the noise. $\mathbf{Y}(\Omega) \in \mathbb{C}^{(N+1)\times M}$ represents the SH basis matrix corresponding to the location of the microphones

$$\mathbf{Y}(\Omega) = [Y_n^m(\theta_1, \phi_1) \quad Y_n^m(\theta_2, \phi_2) \ldots Y_n^m(\theta_M, \phi_M)] \quad (3)$$

The sound pressure $\mathbf{p}_{nm}(k) = \mathbf{Y}^H(\Omega)\mathbf{p}(k)$ in the spatial domain is transformed into SH domain by using SHD and expressed as

$$\mathbf{p}_{nm}(k) = \mathbf{B}(kr) [Y(\theta_s, \phi_s)]^* s(k) + \eta_{nm}(k) \quad (4)$$

where $\eta_{nm}(k) = \mathbf{Y}^H(\Omega)\eta(k)$ represents the SHD of the noise. For the known configuration of the SMA the mode strength matrix can be eliminated, to make the signal independent of the radial component and expressed as

$$\alpha_{nm}(k) = [Y(\theta_s, \phi_s)]^* s(k) + \tilde{\eta}_{nm}(k) \quad (5)$$

where $\alpha_{nm}(k) = B^{-1}(kr)p_{nm}(k)$ and $\tilde{\eta}_{nm}(k) = B^{-1}(kr)\eta_{nm}(k)$. Since the speech signals are time varying, it is analysed by finding the short time Fourier transform (STFT) of (5) and is expressed as

$$\alpha_{nm}(\tau, \nu) = \alpha_{nm}^s(\tau, \nu) + \eta_{nm}^n(\tau, \nu) \quad (6)$$

where α_{nm}^s and η_{nm}^n represent the signal and noise components respectively. τ represents the frame index, and ν represents the frequency bin obtained from STFT. The spherical harmonic phase and magnitude features are extracted from $\alpha_{nm}(\tau, \nu)$, to train the CNN network that maps the features with the DOA classes. The orthogonality property of SH motivates using a sparse dictionary matrix with SH basis functions, where the maximally aligned column gives the direction, leading to combining M-CNN with SBL approach for super-resolution DOA estimates.

III. Hybrid CNN-SBL Framework for Super Resolution DOA Estimation

This section discusses two proposed methods: first, a unified M-CNN for DOA estimation, and second, a hybrid M-CNN and SBL-based model that provides enhanced, super-resolution estimates.

A. M-CNN Model for DOA Estimation in SH Domain

This stage proposes to estimate both azimuth (ϕ) and elevation (θ) using a single multi-label multi-class CNN model. The M-CNN architecture that maps input features to the ϕ and θ classes with 10° separation between the classes. The advantage of the proposed network is to reduce the runtime and training with fewer amount of dataset. The model comprises a learning network consisting of three convolutional layers consisting of 8 filters having kernel size (2×2) followed by batch normalization and max-pooling of size (1×2). The learning network is followed by dense layer and the final layer. The activation function used at each layer is rectified linear unit (ReLU) and the softmax activation function is used in the final layer to get the posterior probabilities of each class. The network's output is classified into ϕ and θ classes. The proposed M-CNN-based learning model computes the posterior probability for the classification of DOAs.

B. SBL Approach for Super-resolution DOA Estimation

This stage enhances the DOA obtained from the above M-CNN model. A dictionary matrix Y_D is obtained such that each column represents the SH basis function corresponding to the angle of the search grid. The search grid is defined by sampling the spatial region represented as $[\theta_1^{CNN} \pm \Delta\theta \cup \theta_2^{CNN} \pm \Delta\theta \cup \cdots \cup \theta_s^{CNN} \pm \Delta\theta, \phi_1^{CNN} \pm \Delta\phi \cup \phi_2^{CNN} \pm \Delta\phi \cup \cdots \cup \phi_s^{CNN} \pm \Delta\phi]$ with 1° sampling interval where $\cup$ integrates the grid for multiple sound sources. $\theta_1^{CNN}, \theta_2^{CNN}, \ldots, \theta_s^{CNN}$ and $\phi_1^{CNN}, \phi_2^{CNN}, \ldots, \phi_s^{CNN}$ represent DOA classes corresponding to multiple sound sources estimated by M-CNN model; $\Delta\theta$ and $\Delta\phi$ are the range around estimated DOA. The advantage of such a grid is to obtain a finite region where actual sources lie. So the search space is minimized and the resolution is

improved with a small sampling interval. To estimate DOAs using sparse recovery method [17], the potential azimuth and elevation space of the incident signals is sampled to create a DOA set using the dictionary matrix. Input signal α_{nm} is then reformulated using dictionary matrix $Y_D \in \mathbb{C}^{(N+1)^2 \times (4\Delta^2 \times S)}$ as a multiple measurement vector model.

$$\alpha_{nm} = Y_D \mathcal{R} + \mathcal{Z}(k) \tag{7}$$

$\mathcal{R} \in \mathbb{C}^{(4\Delta^2 \times S) \times K}$ is the solution matrix and $\mathcal{Z}$ denotes additive white Gaussian noise. Since the number of sources is very less that the directions (column of Y_D) in the grid, $\mathcal{R}$ is sparse and the nonzero row index corresponds to the DOA. The solution to (7), begins with assigning prior probability γ_t for tth row of $\mathcal{R}$, $t = 1, 2, \ldots, (4\Delta^2 \times S)$. Since the rows are independent and assuming prior probability follows zero-mean Gaussian distribution with an unknown variance γ_t, $p(\mathcal{R}_t; \gamma_t) \sim \mathcal{N}(0, \gamma_t I)$, where γ_t controls the row sparsity of $\mathcal{R}$. The noise term $\mathcal{Z}$ also follows a Gaussian distribution, i.e., $p(\mathcal{Z}; \lambda) \sim \mathcal{N}(0, \lambda I)$, where λ is the unknown noise variance.

An expectation-maximization (EM) algorithm is used to estimate the unknown parameters γ_t and λ. In the E-step, we treat $\mathcal{R}$ as hidden variables and use the prior distributions joint distribution of α_{nm} and $\mathcal{R}$ as

$$p(\alpha_{nm}, \mathcal{R}; \lambda, \gamma) = p(\alpha_{nm} | \mathcal{R}; \lambda) p(\mathcal{R}; \gamma) \tag{8}$$

The posterior distribution of $\mathcal{R}$ is then given by:

$$p(\mathcal{R} | \alpha_{nm}; \lambda, \gamma) \sim \mathcal{N}(\mu_z, \Sigma_z) \tag{9}$$

where the mean μ_z and covariance Σ_z are:

$$\mu_z = \Gamma Y_D^T (\lambda I + Y_D \Gamma Y_D^T)^{-1} \alpha_{nm} \;;\; \Sigma_z = \left(\Gamma^{-1} + \frac{1}{\lambda} Y_D^T Y_D \right)^{-1} \tag{10}$$

Here, $\Gamma = \text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_{4\Delta^2})$ is the diagonal matrix of variance parameters γ_t. In the M-step, the hidden parameters γ_t and λ are updated by maximizing the expected log-likelihood of the complete data and given as $\gamma_t = \frac{1}{K} \left(\mathcal{R}_t^T \mathcal{R}_t + (\Sigma_z)_{tt} \right)$ where $(\Sigma_z)_{tt}$ is the t-th diagonal element of the covariance matrix, and K is a normalization constant. The noise variance λ is updated as:

$$\lambda = \frac{1}{UK} \| \alpha_{nm} - Y_D \mathcal{R} \|_F^2 + \frac{\lambda}{U} \text{tr} \left[Y_D \Gamma Y_D^T (\lambda I + Y_D \Gamma Y_D^T)^{-1} \right] \tag{11}$$

where $U = (4\Delta^2 + 1)^2$, $\| \cdot \|_F$ is the Frobenius norm, and $\text{tr}(\cdot)$ denotes the trace of a matrix. The EM algorithm updates μ_z, Σ_z, and the parameters γ_t and λ iteratively until convergence. Upon convergence, the estimated DOA (θ, ϕ) are determined by finding the S maximum peaks corresponding to S sources in the power spectrum of μ_z, with the corresponding indices in the dictionary matrix Y_D providing the estimated DOAs.

IV. Simulations and Experiments

This section presents simulations and experimental results. The performance of the proposed technique is compared to state-of-the-art methods: MUSIC [7], SSPIV [11], and DRHC [16] and M-CNN (proposed).

A. Simulation Setup

Simulations were conducted using the Spherical Microphone Array Impulse Response (SMIR) generator [18], which simulated the room impulse responses (RIRs) for a spherical array Eigenmike with 32 channels [19]. Different room sizes were used which varied in the range $(5 \pm 2 \times 6 \pm 2 \times 7 \pm 2)$ meters, with reverberation time (RT_{60}) randomly chosen between 0.2 and 1 seconds. The signals for the array were generated by convolving source signals with the corresponding RIRs. Speech sources are selected from the Librispeech [20] dataset, and white noise was added at each channel with a signal-to-noise ratio (SNR) between $[5 - 30]$ dB. Received signals are processed in the frequency domain via short-time Fourier transform (STFT), with a 512-length Hanning window and 50% overlap. The distance between the spherical microphone array (SMA) and the source ranges from 2 to 5 meters, with DOA classes of 10° in both azimuth ($\phi \in \{0^\circ, 10^\circ, \ldots, 350^\circ\}$) and elevation ($\theta \in \{10^\circ, 20^\circ, \ldots, 170^\circ\}$). Three active sources were randomly placed in 30 different positions, with an angular separation of 30° in 3D space, to evaluate the proposed method's performance against the baseline methods.

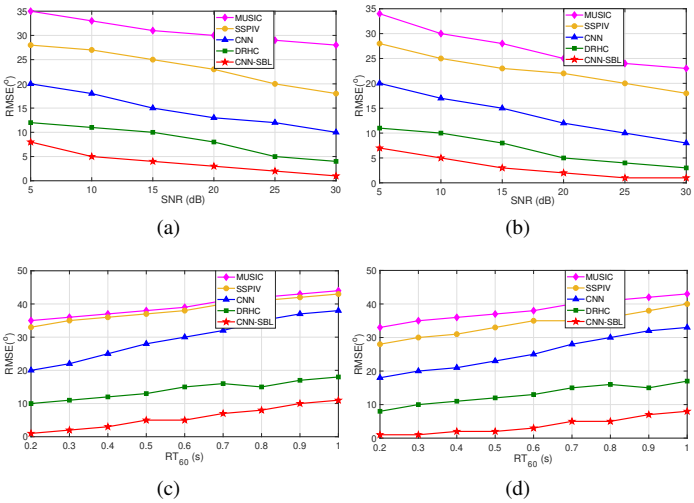

Fig. 2: RMSE plots for azimuth and elevation estimates by varying (a-b) SNR and (c-d) RT_{60}.

B. Simulation Results

Figures 2 and 3 show the RMSE and energy plot results across different methods. In Fig. 2, each result is averaged over 300 trials, where the source positions, speech signals, and noise are selected between $[5 - 30]$ dB in range of 5 dB. All methods experience some degradation as RT_{60} increases 0.2 to 1 s, with the MUSIC method performing worst under reverberation. SSPIV performs slightly better than MUSIC but lags behind the learning-based method M-CNN and DRHC. The DRHC method shows improved results by incorporating high-resolution DOA estimation, although it fails to accurately detect all sources. In contrast, the proposed method consistently yields an RMSE lower than 1° and goes maximum up to 10° even in case if higher noise and reverberation, outperforming all other approaches by accurately predicting and providing high-resolution DOA estimates. Also from Fig. 3, a test case is simulated for three source locations; $(30^\circ, 60^\circ)$, $(70^\circ, 145^\circ)$, and $(140^\circ, 280^\circ)$ with RT_{60}=0.3s, and

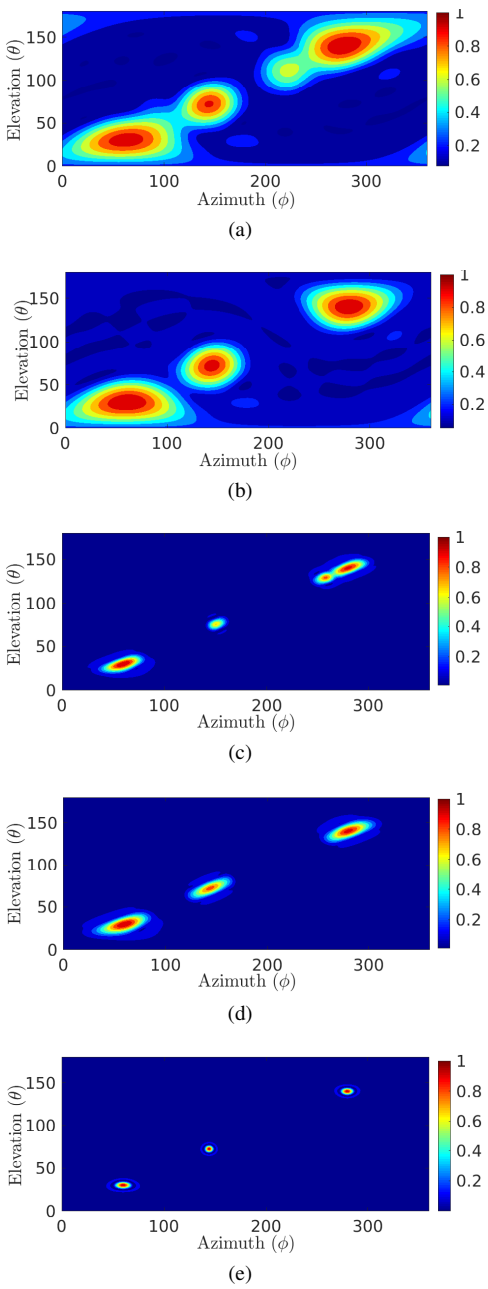

Fig. 3: Energy plots showing the DOA estimates by various methods.

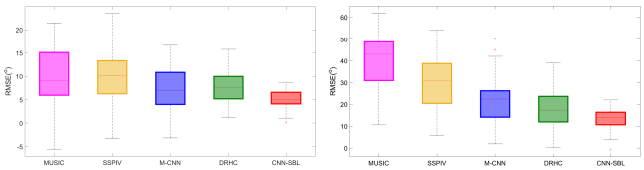

Fig. 4: Box plots showing the experimental results on (a) LOCATA Task-2 dataset and (b) on real datasets recorded in the lab environment.

SNR=5 dB, The DOAs are estimated from all the methods and the proposed method provides the sharpest and most accurate DOA estimates for all the sources.

V. Experimental Results

The performance of the SH-M-CNN-SBL method is compared with the other methods using the real datasets from LOCATA [21], [22] challenge and lab recordings.

A. LOCATA Dataset

The LOCATA dataset, part of the IEEE-AASP challenge, focuses on acoustic source localization and tracking, with this study using the Eigenmike setup in Task 2 to benchmark sound source localization with a stationary spherical microphone array and multiple static audio sources. The recordings were conducted in a laboratory space measuring approximately 7.1 × 9.8 × 3 meters, with $\mathrm{RT}_{60} = 0.55$ s. The dataset provides both audio recordings and corresponding ground truth annotations for evaluation.The results are presented in Fig. 4(a). The figure demonstrates that the SH-M-CNN-SBL method significantly outperforms other DOA estimation techniques.

B. Real-Time Experiments in Lab Environment

To further validate the proposed method, experiments were conducted in a meeting room with dimensions of 5 m × 6 m × 7 m and a RT_{60} of approximately 0.8 s. An Eigenmike array with 32 microphones was used to record signals emitted from three loudspeakers placed in the room. The DOAs for the sources were set at random positions selected to test the method's performance in detecting sources. The sources were placed 1.5 m away from the microphone array, and the SNR of the recorded signals was about 15 dB. The experimental results, shown in Fig. 4 (b), indicate the average performance and standard deviations in the RMSE obtained for various sources for each method. The proposed method delivers the lowest RMSE and standard deviation.

Fig. 5: Experimental setup for DOA estimation on real data in the lab environment.

VI. Conclusion

The proposed framework effectively addresses the challenges of localizing multiple sound sources in complex acoustic environments. The integration of SH decomposition with SBL results in a notable improvement, with up to 1° accuracy in direction of arrival estimation and an 80% reduction in RMSE. The results show the methods's effectiveness and reliability, making it a promising approach.

Acknowledgement

This work was supported by the ANRF under Project No. ANRF/ARG/2025/007629/ENS.

REFERENCES

[1] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 2, pp. II–1949–II–1952.

[2] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. II–1781–II–1784.

[3] Priyadarshini Dwivedi, Gyanajyoti Routray, and Rajesh M. Hegde, "Octant spherical harmonics features for source localization using artificial intelligence based on unified learning framework," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3845–3857, 2024.

[4] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[5] Priyadarshini Dwivedi, Gyanajyoti Routray, and Rajesh M. Hegde, "Sparse bayesian integrated cnn framework for enhanced acoustic source localization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] Priyadarshini Dwivedi, Raj Prakash Gohil, Gyanajyoti Routray, Vishnuvardhan Varanasiy, and Rajesh M. Hegde, "Joint doa estimation in spherical harmonics domain using low complexity cnn," in *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, 2022, pp. 1–5.

[7] Dima Khaykin and Boaz Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[8] Christine Evers, Alastair H Moore, and Patrick A Naylor, "Multiple source localisation in the spherical harmonic domain," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 258–262.

[9] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasias Mouchtaris, "3d doa estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 96–100.

[10] Sina Hafezi, Alastair H Moore, and Patrick A Naylor, "3d acoustic source localization in the spherical harmonic domain based on optimized grid search," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 415–419.

[11] Alastair H Moore, Christine Evers, and Patrick A Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 178–192, 2016.

[12] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[13] Pierre-Amaury Grumiaux, Srdjan Kitić, Laurent Girin, and Alexandre Guérin, "Improved feature extraction for crnn-based multiple sound source localization," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 231–235.

[14] Abdullah Fahim, Prasanga N Samarasinghe, and Thushara D Abhayapala, "Multi-source doa estimation through pattern recognition of the modal coherence of a reverberant soundfield," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 605–618, 2019.

[15] Yonggang Hu and Sharon Gannot, "Closed-form single source direction-of-arrival estimator using first-order relative harmonic coefficients," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 726–730.

[16] Yonggang Hu, Prasanga N Samarasinghe, Sharon Gannot, and Thushara D Abhayapala, "Decoupled multiple speaker direction-of-arrival estimator under reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 3120–3133, 2022.

[17] D.P. Wipf and B.D. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004.

[18] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[19] "The microphone array". "url- mhacoustics.com," .

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[21] Heinrich W. Löllmann, Christine Evers, Alexander Schmidt, Heinrich Mellmann, Hendrik Barfuss, Patrick A. Naylor, and Walter Kellermann, "The locata challenge data corpus for acoustic source localization and tracking," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018, pp. 410–414.

[22] Christine Evers, Heinrich W. Löllmann, Heinrich Mellmann, Alexander Schmidt, Hendrik Barfuss, Patrick A. Naylor, and Walter Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.