Sensor signal processing

Masked-Aware Directional Attention Network for DOA Estimation Under Sensor Failure Conditions

Anuj Kumar Mishra[1,2*], Aditya Srivastava[3], and Ripul Ghosh[1,2**]

[1]*Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*
[2]*CSIR-Central Scientific Instruments Organisation, Chandigarh 160030, India*
[3]*School of Electrical and Electronics Engineering, VIT Bhopal University, Madhya Pradesh 466114, India*
[*]*Graduate Student Member, IEEE*
[**]*Member, IEEE*

Manuscript received 30 October 2025; accepted 4 December 2025. Date of publication 22 December 2025; date of current version 29 January 2026.

Abstract— We investigate direction-of-arrival (DoA) estimation for airborne sound sources using a tetrahedral microphone array, addressing key challenges in drone detection and situational awareness applications. Practical deployment of DoA systems is often constrained by computational demands, stringent calibration requirements, and susceptibility to sensor failures. To address these limitations, we propose masked-aware directional attention network (MADANet), a lightweight signal processing pipeline coupled with a pairwise attention-based neural architecture designed for robust performance under sensor failure scenarios. The architecture extracts magnitude, phase, and geometric features for each microphone pair from resampled acoustic signals (4–20 kHz), selects active frames via energy gating, and embeds these features through a shared multilayer perceptron before applying multihead self-attention for adaptive fusion. A structured sensor dropout mechanism is introduced to mask feature pairs from randomly deactivated microphones and normalize attention weights accordingly. Experiments conducted in a semianechoic chamber demonstrate that downsampling to 4 kHz and using four attention heads minimize mean spherical error, approaching the Cramér–Rao lower bound for the given array geometry. The model exhibits strong generalization to single- and double-microphone failure scenarios, maintaining subdegree accuracy.

Index Terms—3-D direction-of-arrival estimation, airborne acoustic source, deep neural network, localization, sensor failure, tetrahedral array.

I. INTRODUCTION

Microphone array-based sound source localization (SSL) is essential in aeroacoustics, indoor monitoring, and structural fault detection [1]. Classical approaches extract spatial cues, such as time delays, interchannel phase/level differences, and steered response power with phase transform [2]. Although computationally efficient, their accuracy degrades under noise, reverberation, and multipath due to dependence on fixed array topology. Beamforming remains a baseline but suffers from poor low-frequency resolution constrained by the Rayleigh limit and spatial aliasing [3]. Adaptive methods, such as minimum variance distortionless respons (MVDR), multiple signal classification (MUSIC), and estimation of signal parameters via rotational invariance techniques (ESPRIT), enhance resolution [4], yet require precise calibration, known source count, and high signal-to-noise ratio (SNR) conditions rarely met in airborne or outdoor settings. These challenges have driven data-driven approaches. Convolutional neural networks learn localization cues directly from spectrograms or interchannel features [5], improving robustness over analytical models but capturing only local dependencies. Recent works adopt attention and Transformer-based architectures that adaptively weight spatial cues and encode global context, achieving superior performance in reverberant environments [6].

A critical but underexplored challenge in SSL is robustness to sensor failures. In field deployments, microphones may become unreliable due to wind, mechanical damage, or communication dropouts, yet most prior work assumes fully operational arrays. Recent graph neural network (GNN) approaches for distributed microphone networks [7] highlight the benefit of modeling pairwise relations between sensors as edges in a graph, naturally accommodating irregular topologies and missing channels. Building on this insight, MADANet integrates structured sensor dropout into training, where one or more microphones are masked to mimic failure scenarios. Combined with pairwise spatial embeddings and multihead attention fusion, this strategy enables reliable DoA estimation even when parts of the array are compromised. In this study, we make the following three specific advances over prior SSL work.

1) We demonstrate that low-bandwidth (300–2000 Hz) operation reduces computation while achieving near Cramér–Rao lower bound (CRB) accuracy.
2) Propose an attention-based fusion framework combining spectral and geometric cues for robust 3-D localization.
3) Validated resilience under single- and double-microphone failures with graceful performance degradation.

II. METHODOLOGY

This section outlines the experimental setup and framework for direction-of-arrival (DoA) estimation. A tetrahedral microphone array comprising four PCB HT 378B02 sensors was employed, along with a DEWE43 V multichannel 24-bit data acquisition system. Acoustic signals were played using a Bose SoundLink Revolve II speaker and recorded in a semi-anechoic chamber measuring 5 m × 5 m × 4 m. The dataset was curated from real-world acoustic signatures of jet

Corresponding author: Ripul Ghosh (e-mail: ripul.ghosh.csio@csir.res.in).
Associate Editor: Karthick Thiyagarajan.
Digital Object Identifier 10.1109/LSENS.2025.3646924

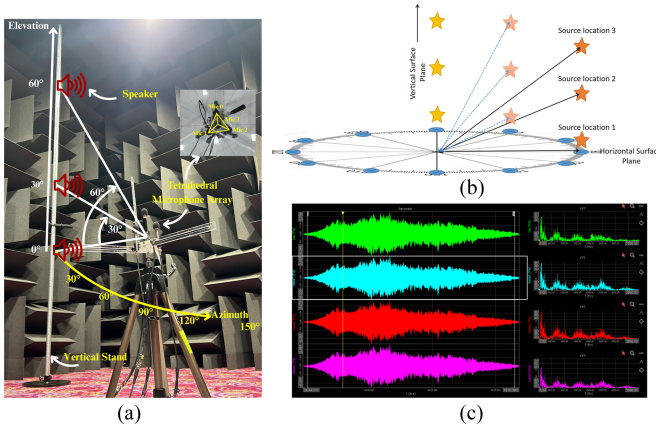

Fig. 1. Representation of (a) experimental setup (b) experimental design (c) graphical user interface of DEWE-43 V acquisition.

aircraft, tandem-rotor helicopters, drones, whistle blowing, and clap instances [8]. Each source position is defined by an azimuth angle $\phi \in \{0^\circ, 30^\circ, \ldots, 330^\circ\}$ and an elevation angle $\theta \in \{0^\circ, 30^\circ, 60^\circ\}$, as shown in Fig. 1, and a tetrahedral microphone array placed at the center of the room at a vertical height of 0.74 m. Let the true source direction be parameterized by

$$\mathbf{u}(\phi, \theta) = \begin{bmatrix} \cos\theta \cos\phi \\ \cos\theta \sin\phi \\ \sin\theta \end{bmatrix} \tag{1}$$

where ϕ and θ are expressed in radians. The tetrahedral array consists of $C = 4$ sensors with known 3-D coordinates $\{\mathbf{p}_c \in \mathbb{R}^3 \mid c = 0, \ldots, 3\}$. For a source signal $s(t)$, the observed signal at microphone c is modeled as

$$x_c(t) = a_c\, s(t - \tau_c) + \epsilon_c(t) \tag{2}$$

where a_c is a channel-dependent attenuation factor, $\tau_c = \frac{1}{c}\|\mathbf{p}_c - \mathbf{p}_s\|$ is the propagation delay given the source position $\mathbf{p}_s$, and $\epsilon_c(t)$ represents additive noise and residual reverberation. The recorded multichannel signals are sampled at rate f_s, segmented into short-time frames of 512 sample points for an audio recording of 200 s for each source location, and transformed into time–frequency features. From these features, the task is to learn a nonlinear mapping

$$f_\Theta : \{x_c(t)\}_{c=0}^{C-1} \;\mapsto\; \left(\hat{\phi}, \hat{\theta}\right) \tag{3}$$

parameterized by neural network weights Θ.

A. Array Geometry and Time–Frequency Analysis

A tetrahedral microphone array of side length $d = 0.08$ m is used, with sensor positions defined as

$$\mathbf{p}_0 = \begin{bmatrix} 0 \\ -\frac{d}{\sqrt{3}} \\ 0 \end{bmatrix}, \quad \mathbf{p}_1 = \begin{bmatrix} \frac{d}{2} \\ \frac{d}{2\sqrt{3}} \\ 0 \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} -\frac{d}{2} \\ \frac{d}{2\sqrt{3}} \\ 0 \end{bmatrix}, \quad \mathbf{p}_3 = \begin{bmatrix} 0 \\ 0 \\ \sqrt{\frac{2}{3}}\, d \end{bmatrix}. \tag{4}$$

The four signals are resampled at f_s and processed using a short-time Fourier transform (STFT) with Hanning window $w[m]$ of length L, hop size H, and DFT size N. For the tth frame and frequency bin k, the STFT of channel c is given by

$$X_c(k, t) = \sum_{m=0}^{L-1} x_c[m + tH]\, w[m]\, e^{-j\frac{2\pi km}{N}}. \tag{5}$$

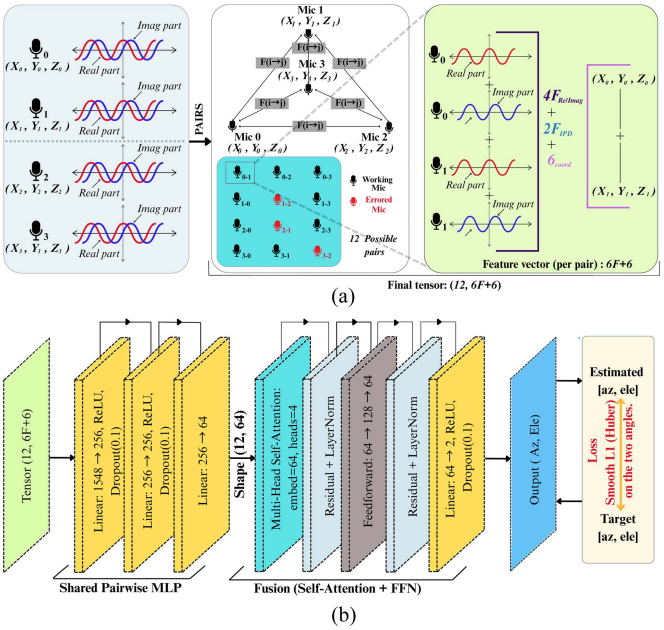

Fig. 2. Representation of (a) feature extraction process pipeline and (b) masked-aware directional attention network architecture.

To eliminate silent windows, an energy-based detector has been used, if the frame's mean channel's rms energy exceeds a fixed threshold chosen empirically $\tau_E = 10^{-4}$ dB.

B. Feature Representation

Spatial cues are encoded on a per-pair basis so that for each ordered pair of microphones (i, j), we extract a feature vector composed of the real and imaginary parts of both spectra, trigonometric encodings of the interchannel phase difference, and the 3-D coordinates of both microphones. Further, these feature vectors $f_{ij}(t) \in \mathbb{R}^D$ concatenate phasor and geometric information, and stacking all 12 ordered pairs yields the final pairwise feature tensor $F(t) \in \mathbb{R}^{12\times D}$, as shown in Fig. 2(a).

C. Pairwise Embedding and Attention Fusion

Each pairwise feature vector is transformed into a latent embedding through a shared multilayer perceptron

$$z_{ij}(t) = \mathrm{MLP}_{\mathrm{pair}}\left(f_{ij}(t)\right) \in \mathbb{R}^E, \qquad E = 64. \tag{6}$$

The resulting set of embeddings $Z(t) \in \mathbb{R}^{12\times E}$ is then passed to a multihead self-attention module with h heads, followed by residual and feedforward blocks. Attention enables the network to weight microphone-pair relational contributions adaptively. Global pooling produces a single frame-level representation

$$\bar{z}(t) = \frac{1}{12}\sum_{p=1}^{12} \tilde{z}_p(t) \tag{7}$$

which is finally mapped to the predicted azimuth and elevation via a regression head, as shown in Fig. 2(b)

$$\hat{y}(t) = \begin{bmatrix} \hat{\phi}(t) \\ \hat{\theta}(t) \end{bmatrix} = \mathbf{W}_2\, \sigma(\mathbf{W}_1 \bar{z}(t)) + \mathbf{b}_2. \tag{8}$$

D. Structured Sensor Dropout for Robustness

To ensure robustness to sensor outages, we introduce a structured dropout mechanism during training. At each iteration, a random subset of microphones $F \subseteq \{0, 1, 2, 3\}$ is declared failed, with probabilities chosen so that 0, 1, or 2 failures occur with predefined frequencies of (50:35:15). Let $\mathbf{F}(t) \in \mathbb{R}^{12\times D}$ represent pairwise tensor extracted for frame t, where each row $\mathbf{f}_{ij}(t) \in \mathbb{R}^D$ corresponds to the concatenated spectral and spatial descriptors for the microphone pair (i, j). Any pair of features involving a failed microphone is masked out, producing

$$\tilde{\mathbf{F}}(t) = \mathbf{F}(t) \odot \left(\mathbf{m}\mathbf{1}_D^\top\right) \tag{9}$$

where $\odot$ denotes elementwise multiplication, $\mathbf{m} \in \{0, 1\}^{12}$ is a binary mask indicating valid microphone pairs, and $\mathbf{1}_D$ is a D-dimensional vector of ones used for broadcasting across feature dimensions.

E. Training and Evaluation

The model is trained by minimizing the Smooth-L_1 loss between predictions and ground-truth DoAs over all active frames

$$\mathcal{L} = \frac{1}{|T|}\sum_{t\in T}\left(\mathrm{SL1}(\hat{\phi}(t), \phi(t)) + \mathrm{SL1}(\hat{\theta}(t), \theta(t))\right) \tag{10}$$

where T denotes the set of active frames. Model training employs stratified 70(train)/15(val)/15(test) splits using per-frame pairwise features extracted from four-channel representations (resampled to 4 kHz, $n_{\text{fft}} = 512$, win_length $= 512$, hop_length $= 128$, Hann window), i.e., in total 1,45,473 / 31,173 / 31,173 frames. Models were optimized with AdamW ($lr = 1 \times 10^{-3}$, weight_decay $= 1 \times 10^{-4}$) using Smooth-L1 loss (PyTorch's $\beta = 1.0$). Training ran for up to 150 epochs with early stopping (patience $= 10$) monitored on validation loss and an improvement threshold of 1×10^{-4} with a minibatch size of 16 samples. Two sets of experiments were conducted 1) a sampling rate variation evaluating the effect of $f_s \in \{20, 16, 8, 4$ kHz$\}$ at fixed four attention heads, and 2) an attention head variation studying $h \in \{0, 1, 2, 4, 8\}$ at fixed $f_s = 4$ kHz, while the sensor dropout strategy ensured stable performance even under 1–2 microphone failures.

1) Cramér–Rao Lower Bound (CRB): In addition to empirical evaluation, the proposed DoA estimator is benchmarked against the CRB, defining the theoretical minimum variance achievable by any unbiased estimator for a given array geometry and noise condition. For broadband signals, information across all frequency bins contributes to the wideband Fisher information matrix

$$\mathbf{J}_{\text{WB}} = \sum_{k\in\mathcal{F}} \gamma_k \left(\frac{2\pi f_k}{c_0}\right)^2 \mathbf{G}(\phi, \theta) \tag{11}$$

where γ_k is the SNR weight of frequency bin f_k, and $\mathbf{G}(\phi, \theta)$ depends solely on array geometry. The wideband CRB is

$$\mathrm{CRB}_{\text{WB}} = \mathbf{J}_{\text{WB}}^{-1} \tag{12}$$

yielding one-sigma angular deviations

$$\sigma_\phi = \sqrt{[\mathrm{CRB}_{\text{WB}}]_{11}} \qquad \sigma_\theta = \sqrt{[\mathrm{CRB}_{\text{WB}}]_{22}}. \tag{13}$$

With $N = 512$-point STFT at $f_s = 4$ kHz, the 300–2000 Hz band corresponds to $k = 39$–256 (218 bins). Assuming flat SNR (γ_k constant), the resulting wideband CRB for the tetrahedral array yields standard deviations of $\sigma_\phi \approx 0.087^\circ$ and $\sigma_\theta \approx 0.076^\circ$.

III. RESULTS AND DISCUSSION

Following the methodology described in Section II, we evaluate the proposed wideband pairwise attention DoA estimator. Performance is reported in terms of mean absolute error (MAE) for azimuth and elevation, as well as the percentage of frames whose error falls within multiples of the CRB.

Table 1. Effect of Sampling Rates for different DoA Error Ranges (°)

Plane	f_s (kHz)	[0,0.01]	[0,0.05]	[0,0.1]	[0,0.5]	[0,1]	[0,5]
Az	20	0.76	4.11	8.25	51.35	84.01	94.14
	16	2.45	11.65	22.90	75.35	95.06	96.13
	8	2.08	10.70	21.15	73.79	93.12	96.38
	4	**3.93**	**19.59**	**40.11**	**98.13**	**98.81**	**99.21**
El	20	6.70	16.89	29.43	99.26	**99.96**	**99.98**
	16	21.13	66.92	87.90	**99.90**	99.95	99.97
	8	20.07	40.65	65.80	90.68	99.78	99.84
	4	**39.94**	**73.71**	**96.64**	99.77	99.78	99.83

Table 2. Effect of Attention Heads for different DoA Error Ranges (°)

Plane	Heads	[0,0.01]	[0,0.05]	[0,0.1]	[0,0.5]	[0,1]	[0,5]
Az	H0	3.49	17.06	33.09	95.98	97.99	98.68
	H1	3.46	16.47	28.73	84.63	97.27	98.55
	H2	4.15	19.96	37.89	94.46	97.69	99.00
	H4	**3.93**	**19.59**	**40.11**	**98.13**	**98.80**	**99.20**
	H8	2.48	11.62	22.64	87.06	98.31	99.17
El	H0	35.38	58.81	77.88	99.64	99.70	99.80
	H1	15.28	45.81	73.90	**99.79**	**99.82**	**99.87**
	H2	24.99	52.15	83.87	99.77	99.79	99.85
	H4	**39.94**	**73.71**	**96.64**	99.77	99.79	99.83
	H8	21.01	36.12	52.96	99.75	99.78	99.82

A. Sampling Rate Analysis

Table 1 summarizes the distribution of absolute DoA errors when the proposed network is trained and evaluated at four sampling frequencies with a fixed attention configuration of four heads. For each frequency, the percentages of frames whose azimuth error falls within increasingly wide error bins are reported. Lower sampling rates consistently yield superior localization accuracy. At 4 kHz, almost 40% of azimuth frames have $\leq 0.1^\circ$ error and 98.8% have $\leq 1^\circ$, whereas at 20 kHz only 8.25% lie within 0.1° and 84% within 1°; moreover, 4.53% of frames exceed 5° error. Elevation estimates are even more precise, at 4 kHz, 96.6% of frames have $\leq 0.1^\circ$ error and only 0.17% exceed 5°. The results indicate that downsampling increases the effective aperture relative to wavelength, improving spatial resolution and pushing accuracy closer to the CRB for airborne acoustic objects.

B. Response of Attention Heads

The number of attention heads h controls how many pairwise interactions the network learns to weight independently. Table 2 lists the percentages of frames whose errors fall within the same thresholds when $h \in \{0, 1, 2, 4, 8\}$ and the sampling rate is fixed at 4 kHz. A moderate number of heads improve accuracy with four heads: more than 40% of azimuth frames have $\leq 0.1^\circ$ error and only 0.8% exceed 5°. In contrast, no-attention (H0) and too many heads (H8) yield poorer performance. Elevation performance follows a similar trend, with four heads giving the highest proportion of small errors.

C. Robustness to Sensor Failures

The structured sensor dropout technique, as given in Section II-D, trains the model to handle missing channels. We evaluated the trained network under all single- and double-microphone failure scenarios. Fig. 3(a) shows the percentage of frames whose azimuth errors fall within multiples of the azimuth CRB (0.087°) for each failure case. Even with two microphones disabled, more than half of the frames fall within $5 \times$ CRB and roughly one-third within $3 \times$ CRB. Single-microphone failures retain high coverage. It is observed that with the

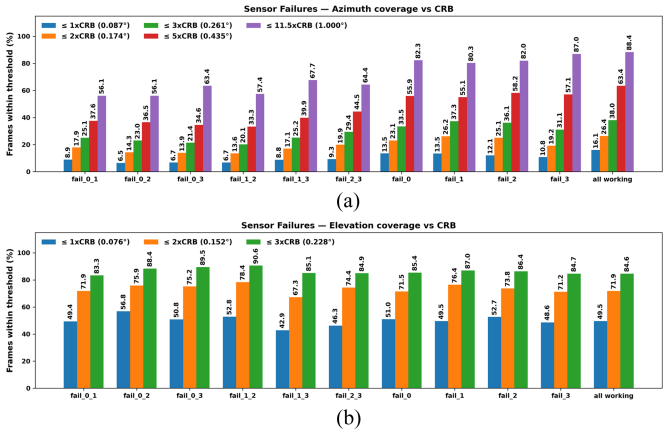

Fig. 3. Representation of (a) azimuth coverage and (b) elevation coverage under sensor failures.

Table 3. Comparison of MADANet with State-of-the-Art Methods

Model	$MAE_{spherical}$ Angle (°)			Inf.	Size
	SSD0	SSD1	SSD2		
MobileNetV2[9]	1.07	2.11 ± 0.43	8.98 ± 2.37	0.20	1.2 MB
AST[10]	0.98	1.63 ± 0.49	8.39 ± 3.02	0.57	3.0 MB
FAST[11]	1.33	2.05 ± 0.42	7.63 ± 2.43	0.58	3.4 MB
MADANet (ours)	**0.67**	**1.62 ± 0.37**	**7.22 ± 2.42**	**0.18**	**2.0 MB**

*SSD0, SSD1, and SSD2 denote all-sensor working, single-failure, and double-sensor failure, Inf. as inference time (ms)

microphone 2 missing, 26% of frames fall within 2 × CRB and 82% within 11.5 × CRB. Complete operation (no failure) yields the best performance, with 16% of frames within CRB, 26% within 3 × CRB, and 87% within 11.5 × CRB.

Fig. 3(b) displays the elevation coverage under failures. Elevation estimation is markedly more resilient, even under two-microphone outages, more than 70% of frames are within 2 × CRB and roughly 85% within 3 × CRB. For all single-sensor failures, at least 70% of frames stay within 2 × CRB. To further validate the MADANet, we benchmarked its performance against three state-of-the-art models, *MobileNetV2* [9], *AST* (Audio Spectrogram Transformer) [10], and *FAST* (Fast Audio Spectrogram Transformer) [11]. As in Table 3, MADANet consistently achieves superior performance across all conditions. The overall angular MAE (0.67°) is notably lower than that of AST (0.98°) and FAST (1.33°), while retaining robustness under single and double-ensor failure scenarios. Furthermore, it requires fewer parameters (0.52M) and achieves the shortest inference latency (0.18 ms), supporting its suitability for real-time embedded deployment.

Considering to previous state-of-the-art studies, there are limited literature directly addressing DoA estimation under sensor failure conditions. In [7] a GNN-inspired relational framework is used to infer source direction from distributed microphones with variable topology conceptually related to the sensor-failure scenario. However, the present work formulates the localization task through a pairwise attention fusion mechanism directly operates on microphone-pair embeddings and incorporates explicit geometric relationships. Similarly, [8] investigated the spectral–spatial latent representations of acoustic sources, emphasizing feature abstraction across dimensional spaces supported in deciding the necessary feature responsible for real-time directional regression. MADANet extends these foundations to a masked-aware attention mechanism capable of dynamically reweighting pairwise contributions under partial sensor loss.

Further evaluating performance under varying reverberation and additive white Gaussian noise conditions at ($RT_{60} \in \{0.01\text{–}0.03$ s}, SNR $\in \{20,\ 10,\ 5$ dB}) without explicitly training on SSD, the model produces an MAE $\leq 1.13°$ for $RT_{60} \leq 0.03$ s and an MAE $\leq 9.59°$ for the lowest SNR of 5 dB, whereas MAE $< 13.0°$ is observed for the combined case ($RT_{60} - 0.03$ s, SNR $-$ 5 dB). Under successive single- and double-sensor failures, the MAE ranges from approximately 14° to 31° and 33° to 51°, respectively, exhibiting consistent performance trends across both reverberant and noisy conditions.

IV. CONCLUSION

The frequency band 300–2000 Hz is selected to balance aliasing constraints and information content. The maximum usable frequency without spatial aliasing is determined by $f_{\text{alias}} \leq \frac{c_0}{2d}$, where $d = 0.08$ m is the microphone spacing. Substituting $c_0 \approx 343$ m/s gives $f_{\text{alias}} \approx 2.14$ kHz. Thus, restricting the band to 2000 Hz avoids phase wrapping across all microphone pairs. Section II E derived wideband CRBs for the azimuth and elevation estimators. Comparing our empirical MAEs with the CRB shows that the proposed network operates near this theoretical limit. At 4 kHz and $h = 4$, more than 80% of frames have azimuth errors within 5 × CRB and nearly 90% of elevation frames are within 3 × CRB. While two-microphone failures increase error, the system still localizes within a few degrees and retains high coverage when errors are measured relative to the CRB. These results validate the proposed approach as a practical solution for real-time, low-bandwidth sound source localization offering high accuracy, resilience to hardware failures, and near-optimal performance relative to the fundamental limits set by the Cramér–Rao bound.

ACKNOWLEDGMENT

This work was supported in part by CSIR-CSIO under Grant Nos. HCP-17 and MMP045201-WPT47A. The first author received fellowship from CSIR-HRDG under Grant 31/GATE/31(40)/2021-EMR-I.

REFERENCES

[1] L. Chen et al., "Application of the improved fast iterative shrinkage-thresholding algorithms in SSL," *Appl. Acoust.*, vol. 180, 2021, Art. no. 108101.
[2] F. Chen et al., "A DOA estimation algorithm based on eigenvalues ranking problem," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 9501315.
[3] T. Suzuki, "L1 generalized inverse beam-forming algorithm resolving coherent/incoherent, distributed and multipole sources," *J. Sound Vib.*, vol. 330, no. 24, pp. 5835–5851, 2011.
[4] R. Ghosh et al., "Estimation of direction of arrival of a moving target using subspace based approaches," in *Proc. SPIE Autom. Target Recognit XXVI*, vol. 9844, 2016, Art. no. 189.
[5] S. Y. Lee et al., "Deep learning-enabled high-resolution and fast sound source localization in spherical microphone array system," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2506112.
[6] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 8–21, Mar. 2019.
[7] E. Grinstein et al., "Graph neural networks for sound source localization on distributed microphone networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
[8] A. K. Mishra et al., "A walkthrough to airborne acoustic source's hidden patterns in lower and higher dimensional space," in *Proc. IEEE Space Aerosp. Defence Conf.*, 2024, pp. 174–177.
[9] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[10] Y. Gong et al., "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 4510–4520.
[11] A. Naman and G. Zhang, "FAST: Fast audio spectrogram transformer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025, pp. 1–5.