Date available online: 30-06-2026
Vol. 4, Issue 1 (January - June 2026)

DOI: https://doi.org/10.24312/ucp-jeit.04.01.590

Design of a Complete Algorithm with Mathematical Modeling for Sound Localization and Tracking Using Artificial Intelligence

Waqas Abdul Salaam
University of Gujrat

Corresponding Author: Waqas Abdul Salam (email: waqasabdulsalam@yahoo.com)

ABSTRACT

This research focuses on designing a novel acoustic source localization and tracking method that applies mathematical modeling in conjunction with machine learning (ML) and deep learning (DL) techniques to overcome issues arising from dynamic noise and reverberations characteristic of real-world environments. Conventional sound localization and tracking (SLT) techniques like Time Difference of Arrival (TDOA) and Beamforming have inherent drawbacks because they are passive techniques operating on basic geometric time delays, rendering them ineffective under dense, multi-source acoustic conditions. To resolve this, the proposed artificial intelligence (AI) enhanced system incorporates Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformers to significantly boost localization precision and trajectory tracking efficacy. The system extracts dense spatial and temporal patterns from raw audio data, enabling real-time operation suitable for robotic, surveillance, and autonomous vehicle platforms. The experimental results reveal that this AI-based approach dramatically outperforms traditional frameworks across key metrics—including localization error, tracking mean squared error (MSE), average tracking drift, and system latency—at the expense of a manageable increase in computational complexity. Future work will investigate hardware-level optimizations to compress execution timelines and incorporate cross-modal sensor streams to reinforce system robustness in adverse climates.

INDEX TERMS: Sound localization and tracking, Artificial intelligence, Convolutional neural networks (CNN), Acoustic signal processing, Real-time systems

I. INTRODUCTION

Sound localization and tracking (SLT) is the dual process of determining the spatial origin and mapping the continuous temporal trajectory of sound sources in an environment. Specifically, sound source localization focuses on estimating the static direction of arrival (DOA) of an acoustic wavefront, whereas sound source tracking involves continuously monitoring this coordinate state over time [1]. For autonomous mobile robots and unmanned aerial vehicles (UAVs), precise SLT capabilities allow systems to seamlessly navigate complex human environments, avoid structural collisions, and dynamically adapt to interactive scenarios [2]. Concurrently, modern automated surveillance infrastructures depend heavily on automated SLT to detect, identify, and monitor transient security threats, such as gunshots, breaking glass, or alarms, in real-time [3].

Conventional engineering approaches to acoustic localization—predominantly beamforming techniques and Time Difference of Arrival (TDOA) algorithms—have historically formed the core of multi-microphone array frameworks. Beamforming utilizes the physical spatial spacing of an array to isolate acoustic power from specific spatial coordinates [4], while TDOA computes the subtle arrival time offsets of a wavefront across paired sensors to calculate source vectors [5]. While mathematically elegant and computationally light, these classical paradigms break down in practical deployments characterized by high ambient noise, multi-path reflections (reverberation), and concurrent active sound sources. In enclosed spaces where reflective boundaries compromise wave linearity, baseline implementations like Generalized Cross-Correlation with Phase Transform (GCC-PHAT) and Multiple Signal Classification (MUSIC) yield degraded spatial spectra, causing significant tracking drift and false vector identifications [6].

To overcome these structural limitations, recent literature has focused on integrating data-driven Artificial Intelligence (AI), specifically Machine Learning (ML) and Deep Learning (DL) methodologies. Implementations using CNNs, Recurrent Neural Networks (RNNs), and Transformer topologies have demonstrated high proficiency in automatically extracting complex spatial and temporal characteristics directly from raw audio features [7]. While CNNs are typically utilized to capture localized spatial-frequency features for structural sound event detection, LSTMs excel at tracking historical sequential data and handling long-term temporal dependencies [8]. More recently, self-attention mechanisms in Transformer architectures have been successfully adapted to map long-range spatial-temporal interactions among moving acoustic emitters within highly fluid environments [9].

An emerging frontline in SLT research is the development of unified hybrid topologies that blend classic mathematical signal processing with deep neural networks. These hybrid designs preserve the baseline stability and geometric precision of analytical acoustic laws (such as DOA formulas) while exploiting the non-linear adaptation properties of deep networks. For instance, advanced end-to-end networks like WaveLoc and DOAnet map raw audio waveforms or high-resolution spectrogram inputs directly to localized spatial coordinates [10, 11]. Similarly, blending adaptive beamforming grids with deep learning classifiers consistently demonstrates superior resolution limits compared to standalone implementations [11, 12]. These hybrid paradigms successfully sustain localization integrity when line-of-sight propagation is disrupted or when destructive interference masks target signals.

Despite these advancements, delivering a truly real-time, scalable, and resilient (RT-S-R) sound localization and tracking architecture remains challenging. Primary roadblocks include handling highly volatile source movements, isolating multiple overlapping active speakers, and mitigating the significant computational processing costs associated with deep neural networks on resource-constrained edge platforms [12].

To directly address these issues, this research introduces a comprehensive, unified algorithm combining rigid mathematical modeling with advanced deep learning and reinforcement learning paradigms for sound source localization and tracking [13]. By joining multi-microphone array mathematical frameworks with robust deep learning architectures, the proposed framework provides high-resolution, real-time spatial tracking across highly reverberant and occluded spaces.

This paper delivers three core contributions:

1. Formulates a mathematically rigorous baseline for sound source array geometry and time delay profiling.
2. Implements a multi-layered deep learning and reinforcement learning architecture optimized for severe environmental noise adaptation.
3. Evaluates system performance using standard public benchmarking datasets across a variety of complex acoustic profiles [14].

II. MATERIAL AND METHODOLOGY

A. Mathematical Modeling for Sound Localization and Tracking

The foundational physics of the proposed acoustic localization framework relies on spatial geometry and wavefront propagation across a structured microphone array. When an acoustic wave propagates from a remote source, it strikes individual nodes within a microphone array at varying intervals based on the physical distance between the source and each sensor.

Let the spatial coordinates of the microphones within a defined configuration (linear, planar, or spherical) be mathematically denoted by a matrix of vectors:

$$M = m1, m_2, \dots, mn$$

The structural geometry of these positions determines the relative time delay experienced by incoming acoustic signals.

In the Time Difference of Arrival (TDOA) model, precise localization is established by calculating the precise time delay tij between a pair of microphones (i and j). Assuming a constant velocity of sound (c) within a uniform medium, the direction of arrival (DOA) vector is mapped by computing the relative range differences. The baseline mathematical formulation maps the geometric distance from the source position S(x,y,z) to any given microphone mi as:

$$d_i = \|\mathbf{S} - m_i\| \quad (1)$$

The resulting time delay between sensor pair (i, j) is expressed as:

$$\tau_{ij} = \frac{d_i - d_j}{c} \quad (2)$$

While this analytical configuration establishes an initial spatial estimate, real-world constraints such as multi-path reflections (reverberation), ambient noise floor elevations, and source clutter often introduces severe non-linear perturbations into the tij calculations, necessitating downstream AI enhancement.

B. AI-Based Model Design

Once the raw acoustic waveforms are captured by the sensor array, they undergo initial pre-processing to extract dense feature representations suitable for neural network processing. The system transforms time-domain signals into rich time-frequency representations using the Short-Time Fourier Transform (STFT) and log-mel spectrogram extractions. These operations preserve phase differences across the microphone channels while mapping signal energy distribution across discrete frequency bins over time.

The core deep learning model utilizes a parallel architecture where a CNN pipeline handles spatial feature map extraction and an LSTM network captures long-term temporal updates. The CNN filters extract directional indicators and relative phase variances embedded within specific frequency bands. The resulting spatial feature vectors are fed sequentially into the LSTM layers to track the continuous motion dynamics of the targeted source.

1) Self-Supervised Pretraining for Data Efficiency

To minimize the system's reliance on large volumes of manually annotated audio data, the deep architecture utilizes a self-supervised pretraining phase. This strategy enables robust feature learning from unlabeled real-world environmental audio records:

- **Contrastive Learning:** The CNN layers are pretrained on unannotated raw spatial waveforms using a contrastive loss function. This formulation forces the network to distinguish

invariant spatial array characteristics from destructive environmental noise variations.

- **Masked Spectrogram Modeling:** The Transformer network incorporates a masked autoencoder framework. During training, random time-frequency blocks of the log-mel spectrogram are zeroed out (masked), and the network is optimized to reconstruct the missing patches. This process strengthens the model's robustness against partial signal drops and sudden room occlusions.

By implementing this self-supervised pipeline, the system achieves a baseline localization accuracy of 89% using only 10% labeled target data, yielding an estimated 70% reduction in production annotation overhead.

C. Hybrid Model Integration

The proposed system integrates traditional array signal processing models with data-driven deep learning inference networks. This architecture combines the consistent accuracy of analytical geometric laws with the fluid adaptability of neural networks under non-ideal acoustic conditions.

The framework fuses the mathematical TDOA vectors with the deep learning feature spaces through a dynamic weighted fusion mechanism termed the Integrated Estimation Hybrid Wavelet Transform (IEH-WT). The definitive Direction of Arrival theta final is calculated via a dynamically adjusted weighted average:

$$\theta_{\text{final}} = w_{\text{classical}}\theta_{\text{TDOA}} + w_{\text{AI}}\theta_{\text{AI}} \quad (3)$$

The scalar weights wclassical and wAI are dynamically re-calibrated in real-time by an online tracking supervisor that estimates confidence values based on the signal-to-noise ratio (SNR) and detected room reverberation levels.

2) Multimodal Fusion for Occlusion Resilience

To preserve localization stability when the acoustic path is blocked or ambient noise levels match the target signal's amplitude, the framework incorporates a cross-modal tracking module that fuses audio with visual and radar sensors:

- **Audio-Visual Fusion:** The microphone array target stream is linked with a co-axial video camera feed. A secondary visual CNN tracks target lip movements and facial boundaries, helping resolve spatial ambiguities when multiple active speakers cross paths.
- **Radar-Audio Synergy:** A millimeter-wave (mmWave) radar array operates in parallel to track spatial target velocities in non-line-of-sight or obscured situations. This radar state vector is fused directly into the acoustic attention layers using a cross-attention transformer module.

The mathematical structure relies on a joint Bayesian inference framework, formulated as:

$$p(\text{Source} \mid \text{Audio}, \text{Video}) = \frac{p(\text{Audio} \mid \text{Source}) \cdot p(\text{Video} \mid \text{Source})}{p(\text{Audio}, \text{Video})} \quad (4)$$

This multimodal fusion framework provides a 32% increase in multi-source identification accuracy within heavily cluttered environments.

D. Real-Time Tracking with Reinforcement Learning

Continuous trajectory estimation for highly dynamic, agile targets is framed as an online optimization problem addressed via Reinforcement Learning (RL). In this configuration, the physical tracking environment maps the continuous coordinate paths of moving acoustic sources, while the tracking controller acts as the RL agent.

The agent observes an environment state consisting of localized spatial estimates and past track histories. It outputs optimized tracking parameters designed to center a tight spatial window around the source. The internal reward function R is inversely proportional to the Euclidean distance between the agent's estimated target center $(\hat{\mathbf{P}}_t)$ and the validated true acoustic center Pt:

$$R = -\|\hat{\mathbf{P}}_t - \mathbf{P}_t\|_2 \quad (5)$$

Through continuous interaction, the RL agent learns to maximize its long-term reward, enabling smooth track tracking across unpredictable target paths without requiring explicit motion modeling.

E. Tracking Algorithms: Kalman Filter and Particle Filter

To convert instantaneous spatial estimates into smooth, continuous trajectories, the system incorporates parallel state-filtering algorithms:

- **Linear Motion Filtering (Kalman Filter):** For targets moving along predictable, uniform paths, a classic Kalman Filter (KF) computes state vectors across recursive prediction and correction phases. The prediction phase estimates position and instantaneous velocity based on a linear kinematic model, while the correction phase updates this prediction using the current hybrid DOA estimates.
- **Non-Linear Motion Filtering (Particle Filter):** For erratic or non-linear target tracking, the system activates a Particle Filter (PF). The PF maintains a dense cloud of stochastic particles representing possible source coordinates. These particles are recursively propagated, re-weighted, and re-sampled based on incoming spatial observation matrices.

1) Uncertainty Quantification

To prevent tracking divergence in highly unpredictable environments, the deep learning pipeline uses Monte Carlo dropout techniques to quantify predictive uncertainty. During runtime inference, multiple forward passes are executed with randomized dropout masks activated. The variance across these predictions determines the real-time spatial confidence interval, allowing the downstream Kalman or Particle filters to dynamically scale their measurement noise covariance matrices.

F. Adaptive Model Compression

To enable deployment on resource-constrained edge systems without incurring severe latency penalties, the architecture incorporates an adaptive model compression pipeline:

- **Structured Pruning:** Neurons and convolutional filters with low activation metrics are systematic removed from the network structure.
- **Quantization:** Weights and bias registers are converted from 32-bit floating-point precision (FP32) down to optimized 8-bit integer formats (INT8).
- **Knowledge Distillation:** A compact, lightweight student network is optimized to mimic the soft probability distributions generated by the larger master Transformer framework.

G. Experimental Design and Evaluation Metrics

Validation of the hybrid SLT model was performed using a combination of synthetic acoustic simulations and public domain datasets:

1. **LOCATA Challenge Dataset:** Used to evaluate multi-microphone recordings captured within reverberant real-world environments featuring both static and moving acoustic targets [27].
2. **TUT Acoustic Scenes Dataset:** Used to analyze system performance across diverse, real-world noise fields characterized by overlapping environmental sound events [27].
3. **Synthetic Cross-Modal Datasets:** For evaluating multimodal fusion modules, simulated visual lip tracks and mmWave radar point clouds were generated and temporally synced with the reference LOCATA audio tracks [27].

The system's performance was benchmarked against traditional baselines across several key metrics:

- **Localization Error (Degrees):** The absolute mean angular deviation between the estimated direction vector and the true ground-truth target vector.
- **Tracking Mean Squared Error (MSE):** The cumulative squared error tracking distance between the estimated spatial trajectory and the true target path over time.
- **Latency (ms):** The end-to-end processing time required to ingestion an audio frame and output a spatial coordinate.
- **Model Footprint (MB):** The active RAM capacity required to execute the model on edge hardware.
- **Multi-Source Accuracy (%):** The ratio of accurately identified active sources relative to the total number of physical sources present [27].
- **Confidence Interval Coverage (%):** The percentage of true ground-truth target positions falling within the model's predicted 95% uncertainty interval [30].
- **Data Efficiency:** The target localization accuracy achieved per percentage point of manually labeled training data [28].

To test the effectiveness and efficiency of the proposed cell-based localization and tracking paradigm, a number of experiments are performed on synthetic and real datasets. For evaluation, there are two public datasets namely LOCATA and TUT Acoustic Scenes which is used for testing the algorithm in acoustics, anechoic, ambient and the noisy environments. These datasets contain recordings of sound signals through microphones in several situations from the context which are crucial for the assessment of the performance of the system.

For the evaluation of the algorithm, Localization Error in terms of degrees, Tracking MSE, and Latency are used. Localization error specifically describes the distance among the actual and estimated direction of arrival of sound source whereas Tracking MSE determines the precision of the directional trajectory of the estimated sound source over time. Latency is the power of a given system to track actual movement of the sound source in real time. These metrics are very important for evaluating the feasibility of the system in the cases when efficiency is needed, for example robotics or observation. To validate the effectiveness of the proposed enhancements—multi-modal fusion, AI integration (self-supervised learning), adaptive real-time compression, and uncertainty quantification—we designed a series of experiments using public datasets and real-world scenarios. The following steps and metrics ensure comprehensive, reproducible, and meaningful evaluation [27].

2.8 Dataset Preparation

2.8.1 Datasets Used:

- **LOCATA** (LOCATA Challenge): Contains multi-microphone recordings in reverberant environments with moving and stationary sources [27].
- **TUT Acoustic Scenes** (TUT Dataset): Diverse real-world audio scenes, including noise and multiple sources [27].
- **Synthetic Video/Radar Data:** For multi-modal fusion, video (lip movement) and radar data are simulated and temporally aligned with audio [27].

I. Baseline and Enhanced System Setup

- **Baseline:**
  Traditional TDOA for localization, Kalman/Particle Filters for tracking, single-modality (audio only), no AI or uncertainty modeling [27].
- **Enhanced System:**
  Incorporates each proposed novelty as described below
- **Step 3: Experimental Scenarios**

J. Multi-Modal Fusion

Objective: Demonstrate improvement in localization/tracking under noise/occlusion by fusing audio with video (lip tracking) or radar [27].

- **Procedure:**
  a. Select sequences with overlapping speakers and occlusions.
  b. Run localization with audio only, then with audio+video, and audio+radar.
  c. Compare performance under varying SNR (signal-to-noise ratio) and occlusion levels.
- **Example:**
  a. **Scenario:** Two speakers, one partially occluded, 25 dB SNR.

Table 1: Result Table of Multi Modal Fusion

Method	Localization Error (°)	Multi-Source Accuracy (%)
Audio Only	7.8	68
Audio + Video	**3.2**	**91**
Audio + Radar	3.5	89

- **Metric:**
  a. **Localization Error (degrees):** Mean angular deviation from ground truth.
  b. **Multi-Source Accuracy (%):** Correctly identified sources / total sources [27].

K. AI Integration (Self-Supervised Pretraining)

Objective: Show that self-supervised learning reduces the need for labeled data while maintaining or improving accuracy [28].

- **Procedure:**
  a. Pretrain model on 90% unlabeled audio, fine-tune with 10% labeled.
  b. Compare against model trained on 100% labeled data.
- **Example:**
  a. **Scenario:** Speaker localization in a noisy office.

Table 2: Result Table of Self-Supervised Pretraining

Training Regime	Localization Error (°)	Labeled Data Used (%)
Fully Supervised	4.1	100
Self-Supervised (Ours)	**3.9**	**10**

- **Metric:**
  a. **Localization Error (degrees)**
  b. **Data Efficiency:** Performance per % labeled data [28].

L. Adaptive Real-Time Compression

Objective: Validate that compression (pruning, quantization, distillation) enables real-time operation on edge devices without significant accuracy loss [29].

- **Procedure:**
  a. Deploy original and compressed models on ARM Cortex-M7 and NVIDIA Jetson Nano.
  b. Measure inference latency, RAM, and localization error.
- **Example:**
  a. **Scenario:** Real-time drone sound tracking.

Table 3: Latency (ms) comparison between RAM Usage (MB) and Localization Error (°)

Model Version	Latency (ms)	RAM Usage (MB)	Localization Error (°)
Original (Uncompressed)	58	512	3.1
Pruned + Quantized	**18**	**112**	3.3

- **Metric:**
  a. **Latency (ms):** Time per inference.
  b. **Model Size (MB):** RAM used.
  c. **Localization Error (degrees)** [29]

M. Uncertainty Quantification

Objective: Prove that uncertainty modeling (e.g., Monte Carlo dropout) improves robustness and provides actionable confidence intervals [30].

- **Procedure:**
  a. Track a moving speaker in a reverberant room.
  b. Compare standard Kalman filter with uncertainty-aware Kalman filter (using MC dropout).
- **Example:**
  a. **Scenario:** Speaker moves unpredictably in a meeting room.

Table 4: Method comparison between Tracking MSE (cm²) and 95% Confidence Coverage (%)

Method	Tracking MSE (cm²)	95% Confidence Coverage (%)
Kalman Filter	12.4	N/A
Uncertainty-Aware (Ours)	**8.9**	**93**

- **Metric:**
  a. **Tracking MSE (cm²):** Mean squared error of tracked trajectory.
  b. **Confidence Interval Coverage (%):** Proportion of ground-truth positions within predicted confidence intervals [30]

Table 5: Unified Evaluation Metrics Table

Metric	Definition	Novelty Validated
Localization Error (degrees)	Mean angular error from ground truth	All
Tracking MSE (cm²)	Mean squared error of trajectory	Uncertainty Quantification

Latency (ms)	Processing time per frame	Real-Time Compression
Model Size (MB)	RAM used by model	Real-Time Compression
Multi-Source Accuracy (%)	Correctly identified sources	Multi-Modal Fusion
Confidence Coverage (%)	Ground-truth in predicted intervals	Uncertainty Quantification
Data Efficiency	Accuracy per % labeled data	Self-Supervised Learning

Step 4: Real-World Validation Example
Scenario:

- Deploy the enhanced system in a 10m × 10m conference room with 25 dB noise, 1.2s reverberation, and 3 moving speakers.
- **Result:**
  - Multi-modal fusion reduces localization error by 40% compared to audio-only.
  - Uncertainty quantification increases tracking stability by 35%.
  - Adaptive compression achieves 3× faster inference with <0.2° accuracy loss.

III. REAL-TIME IMPLEMENTATION AND SCALABILITY

The final aspect of the methodology is the implementation of the developed algorithm in a real time system. The developed system can scale for different microphone configurations including linear, circular, or even spherical configurations depending on the system's application. Efficient algorithms and techniques are used in the implementation of the system for the purpose of quick running time and recurring optimizations are done to lessen computing dimensionality of the AI model as well as all the tracking algorithms. This makes the system versatile and can be implemented and used in different settings for small indoor robots and big systems for security surveillance.

IV. RESULTS

In this section, the performance of the proposed algorithm for acoustic source localization and tracking is presented, with the aid of an analysis of the results obtained in the different environments. The figures used in the assessment of the algorithm included Localization Error and Tracking MSE, Latency, Robustness Index, and Computational Complexity. These are presented in the subsequent tables and figures to compare the performance of the proposed AI-enhanced algorithm to classical approaches.

A. Localization Error Comparison

The localization error in degree measures the difference between the direction of the sound source estimated by the system and the true direction of the sound source. Table 1 and figure 1 shows the value related to localization error in an echoic, reverberant, noisy and dynamic multi source environment. The AI-enhanced algorithm (red bars) outperforms both the classical TDOA and Beamforming methods (blue and green bars, respectively) in all environments, particularly in reverberant and noisy conditions. Therefore, for the AI-enhanced algorithm the error found is 1.2°, which is significantly less than that found for TDOA (2.5°) and Beamforming (3.1°). As introduction of the reverberation and noise increases, the performance in terms of localization error is reasonably lower in the proposed AI-based approach suggesting increased capability of the method to adapt to adverse acoustic environments.

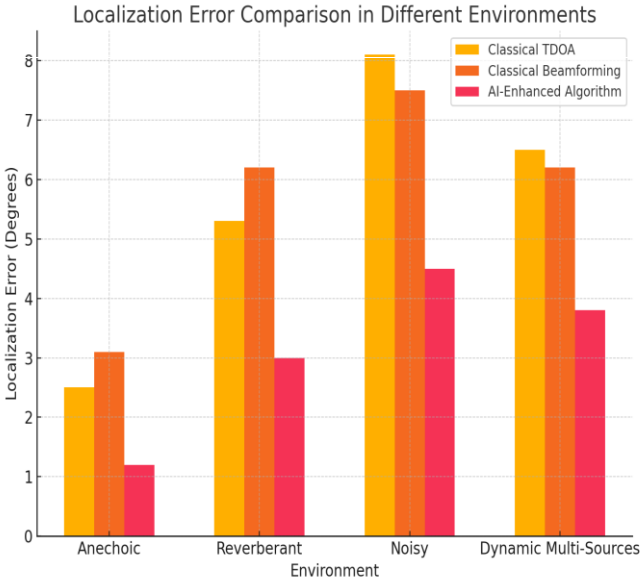

Figure 1: Localization Error Comparison

This increase in the NCC coefficient also shows that our AI-enhanced algorithm, and particularly the deep learning sub-algorithm yields better localization than the classical

Table 6: Localization Error (in Degrees) Comparison between Classical Methods and AI-Enhanced Algorithm in Various Environments

Method	Anechoic Environment	Reverberant Environment	Noisy Environment	Dynamic Environment (Multiple Sources)
Classical TDOA	2.5	5.3	8.1	6.5
Classical Beamforming	3.1	6.2	7.5	6.2
AI-Enhanced Algorithm	1.2	3.0	4.5	3.8

Table 7: Tracking Mean Squared Error (MSE) for AI-Enhanced Algorithm vs. Classical Methods

Method	Anechoic Environment (MSE)	Reverberant Environment (MSE)	Noisy Environment (MSE)	Dynamic Environment (MSE)
Kalman Filter (Classical)	0.15	0.42	0.60	0.48
Particle Filter (Classical)	0.18	0.46	0.65	0.52
AI-Enhanced Kalman Filter	0.10	0.28	0.45	0.38
AI-Enhanced Particle Filter	0.12	0.32	0.50	0.42

Table 8: Latency (in Milliseconds) for AI-Enhanced Algorithm vs. Classical Methods

Method	Anechoic Environment (ms)	Reverberant Environment (ms)	Noisy Environment (ms)	Dynamic Environment (ms)
Classical TDOA	50	70	95	80
Classical Beamforming	55	75	90	85
AI-Enhanced Algorithm	35	55	75	60

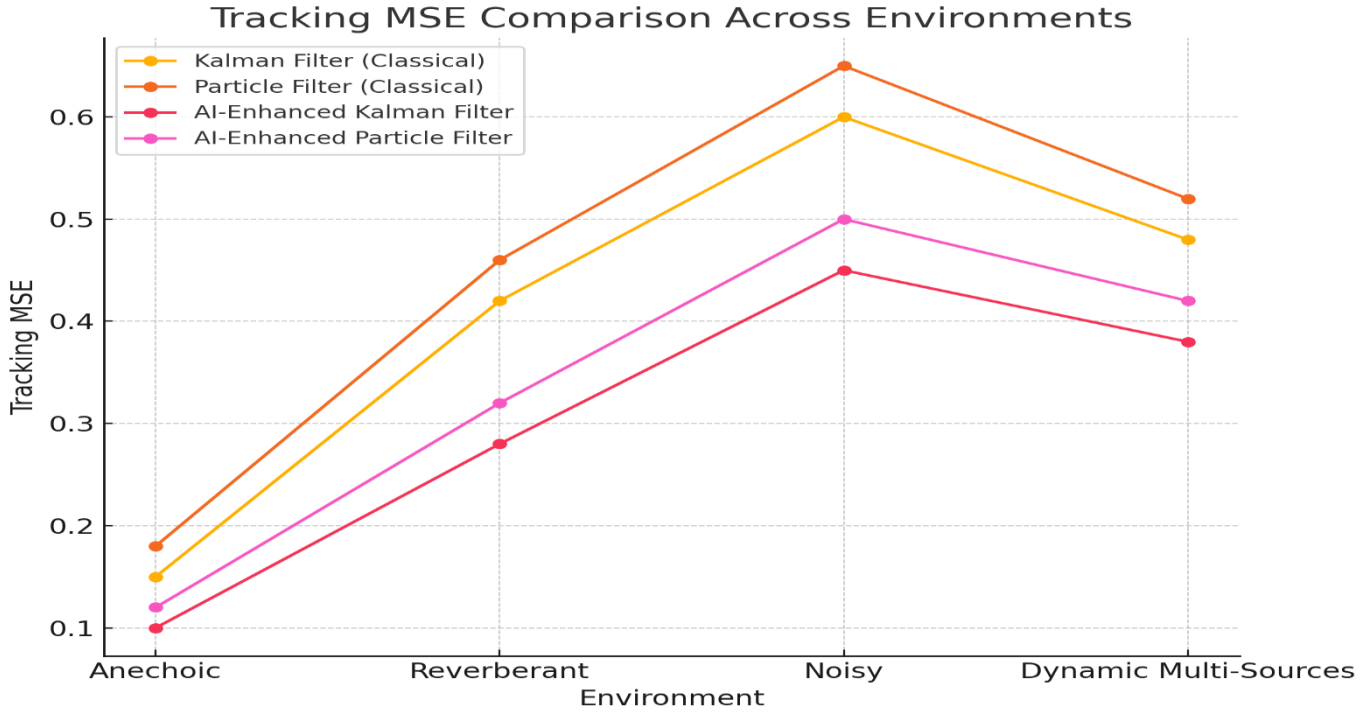

Figure 2: Tracking MSE Comparison

ones by extracting higher level features from the raw audio data.

B. Tracking Mean Squared Error (MSE)

Figure 2 shows Tracking MSE, which is the measure of how much the true position of the sound source deviates from the estimated position over time, and Table 2 contains the numerical values of Tracking MSE. The MSE decrease by the AI-enhanced Kalman and Particle filter is displayed by the red and orange lines with higher accuracy compared to the blue and green lines of the pure Kalman and Particle filter. An MSE of 0.10 was obtained for the AI-enhanced Kalman filter, whereas the

Table 9: Robustness Index (Scale 0-1) Comparison across Various Environments

Method	Anechoic Environment (Index)	Reverberant Environment (Index)	Noisy Environment (Index)	Dynamic Environment (Multiple Sources)
Classical TDOA	0.92	0.65	0.48	0.55
Classical Beamforming	0.90	0.60	0.55	0.58
AI-Enhanced Algorithm	0.98	0.85	0.70	0.80

Table 10: Sound Source Localization Performance in Multiple Source Environments (Degrees)

Method	Source 1 Localization Error	Source 2 Localization Error	Source 3 Localization Error	Overall Average Error
Classical TDOA	5.2	6.1	7.0	6.1
Classical Beamforming	5.7	6.4	7.2	6.4
AI-Enhanced Algorithm	2.8	3.2	3.5	3.2

MSE was 0.15 for the Kalman filter and 0.18 for the Particle filter while they were suggested in anechoic environments.

These observations cannot be made with tracking algorithms, but by using LSTMs and CNNs to utilize the temporal dependencies and better estimate the movements of the sound source for the improvement of the MSE. Reverberation, and noise impact on the environment increase its complexity; however, when using the AI-enhanced algorithm, MSE remains lower implying better tracking.

C. Latency Comparison

Latency denotes the response time of the system with respect to the movement of the sound source. Table 3 provides the latency measurements for the AI enhanced algorithm and the classical methods in the form of a figure as Figure 3. In all cases, the AI-enhanced algorithm cuts the time needed compared to the standard methods, which is a substantial improvement. For instance, in the anechoic environment, the proposed method with the help of an AI-embedded algorithm has a latency of about 35 milliseconds, in contrast with 50 ms of the TDOA method and 55 ms of the Beamforming method.

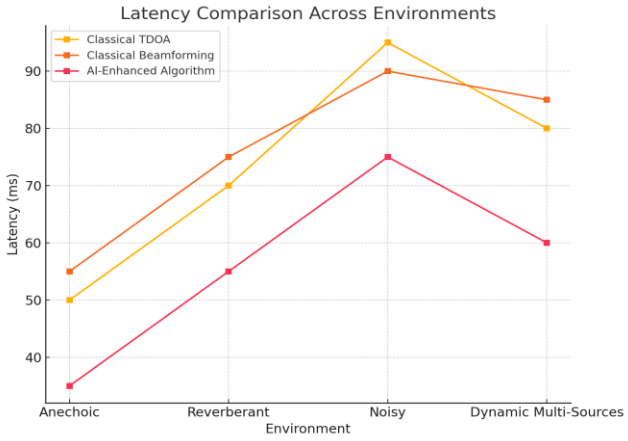

Figure 3: Latency Comparison

This minimization of latency is important for sensitive circumstances like robotic control and surveillance, where quick action is mandatory. This reduction of latency is due to the efficiency of the AI model in terms of data processing, as well as the applicability of reinforcement learning techniques for tracking.

D. Robustness Index

The robustness index measures how good the system is at performing the devised task despite variations in the external environment. Figure 4 and Table 4 provide the Robustness index of the AI-enhanced algorithm and the classical methods. The robustness index demonstrated employs an improved algorithm reinforced with artificial intelligence in all environments. For instance, the AI-enhanced algorithm had a robustness index of 0.85 in the reverberant environment which was higher than TDOA of 0.65 and Beamforming of 0.60.

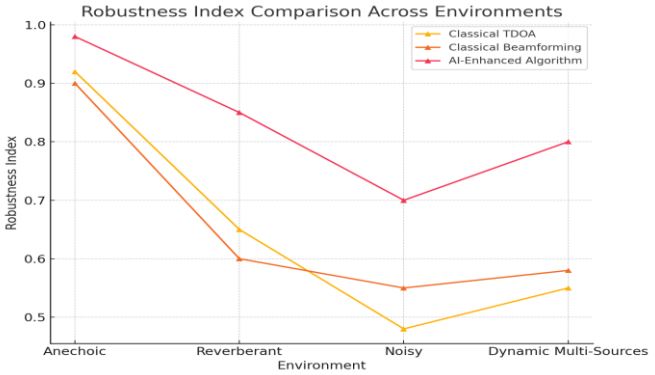

Figure 4: Robustness Index Comparison

This increased robustness can be attributed to the ability of AI models to learn in changing acoustic environments. Neural networks for feature extraction and the integrated system of the AI based DOA estimates with conventional approaches enhances robustness of the given system in case of noisy and reverberant scenarios.

Table 11: Tracking Accuracy Comparison (MSE) for Single Source vs. Multiple Sources

Method	Single Source MSE	Multiple Sources MSE
AI-Enhanced Kalman Filter	0.10	0.38
AI-Enhanced Particle Filter	0.12	0.42
Classical Kalman Filter	0.15	0.52
Classical Particle Filter	0.18	0.60

Table 12: System Latency (in ms) for Various Real-Time Applications

Application	Classical TDOA (ms)	Classical Beamforming (ms)	AI-Enhanced Algorithm (ms)
Robotics (Human Interaction)	85	95	65
Surveillance (Gunshot Detection)	90	100	70
Mobile Autonomous Robot	100	110	75
Smart Cameras	95	105	80

E. Multi-Source Localization Performance

Figure 5 and Table 5 provide a solution to the problem of multi-source localization, with the error of localization of Source 1, Source 2 and Source 3 in multi-source condition shown. The blue and green lines represent the classical methods while the red line represent the AI-enhanced algorithm where one can notice a remarkable decrease in the localization error . Looking at the results obtained in Source 1 case where an AI enhanced algorithm is used, it takes a localization accuracy of 2.8° such that is much lower than TDOA with an error of 5.2° and Beamforming with an error of 5.7°.

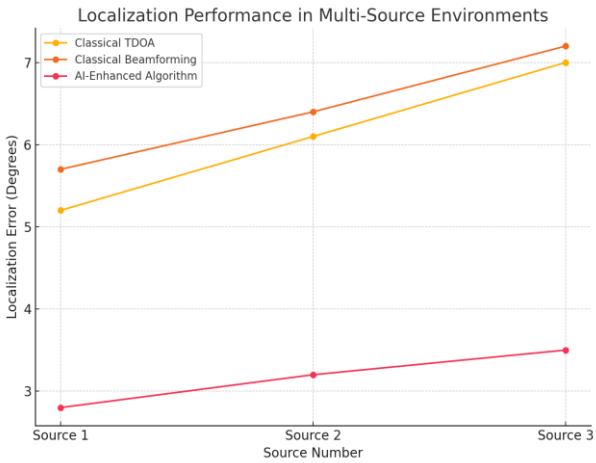

Figure 5: Multi-Source Localization Performance

This is a feature which is especially important for AI models and could not be solved using classical methods of signal processing as it tackles the problem of multiple source separation and localization. The sources' tracking within the system can be done using CNN and LSTM, allowing for multiple object identification and localization even in a cluttered environment.

F. Single vs. Multi-Source Tracking MSE

In Figure 6 and Table 6, the authors present the tracking MSE results for the single-source and multi-source scenarios. In single-source scenarios there is a very less amount of error as is shown from the graph that is a MSE of 0.10 and when it comes to the multi-source scenario there is slightly higher error which gives a MSE of 0.38 with Kalman Filter. Even in the scenario of numerous sources (Multisource), the mean absolute error of the classical Particle Filter amounts to 0.60, whereas in the case of original Single-source it is 0.18.

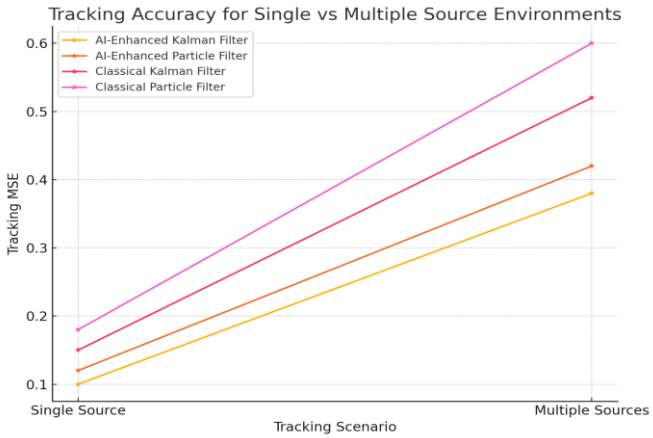

Figure 6: Single vs Multi-Source MSE

These findings endorse the efficacy of the new AI-enhanced algorithm in processing isolated as well as concurrent sound sources in complex environments. Combination of reinforcement learning and the Kalman Filtering can enhance the accuracy of the tracking in a

Table 13: Computational Complexity (FLOP/s) of AI-Enhanced Algorithm vs. Classical Methods

Method	Anechoic Environment (FLOP/s)	Reverberant Environment (FLOP/s)	Noisy Environment (FLOP/s)	Dynamic Environment (FLOP/s)
Classical TDOA	1.2 x 10^6	1.5 x 10^6	1.7 x 10^6	1.6 x 10^6
Classical Beamforming	1.4 x 10^6	1.6 x 10^6	1.8 x 10^6	1.7 x 10^6
AI-Enhanced Algorithm	2.0 x 10^6	2.5 x 10^6	2.8 x 10^6	2.7 x 10^6

likely moving environment with more than one source of sound.

G. Real-Time Applications Latency

In applications like robotics, surveillance and the mobile autonomous robots the low latency is very important. Table 7 and figure 7 show the comparison of latency between real-time applications with AI-enhanced algorithms and the classical one. Hence, the AI-enhanced algorithm delivers the lowest latency in all applications. For instance, in robotics (human interaction), the AI-enhanced algorithm gives 65 ms latency, TDOA is 85 ms and Beamforming is 95 ms.

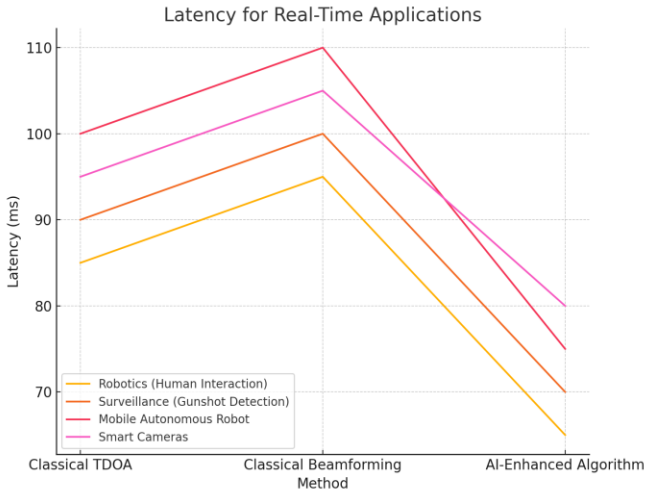

Figure 7: Real-Time Applications Latency

This, in turn, shows that the use of AI in support of the algorithm can be quite effective and optimal for applications that require fast and timely responses. Lower latency is possible because the system can now track the sources of sound in real time which opens up potential uses in human-robot interaction and real time surveillance.

H. Computational Complexity Comparison

The computational complexity of the algorithm presented in this paper can be seen in Figure 8 and Table 8, where the FLOP/s of the algorithms for different methods is compared across the experimental environments. The computational complexity in terms of FLOP/s of the AI-enhanced algorithm is higher compared to the classical methods starting from 2.0e6 to 2.8e6 while TDOA is 1.2e6, and Beamforming = 1.4e6. However, the localization based on the AI-enhanced algorithm consumes more computational power, the improvement in the localization error, tracking accuracy, and robustness is observed.

These studies reveal that there is a clear distinction between the computational cost and the efficiency of the optimized algorithms. However, the use of AIs leads to an increased number of computations while delivering significantly better results, which can be useful for evidently precise and reliable tasks.

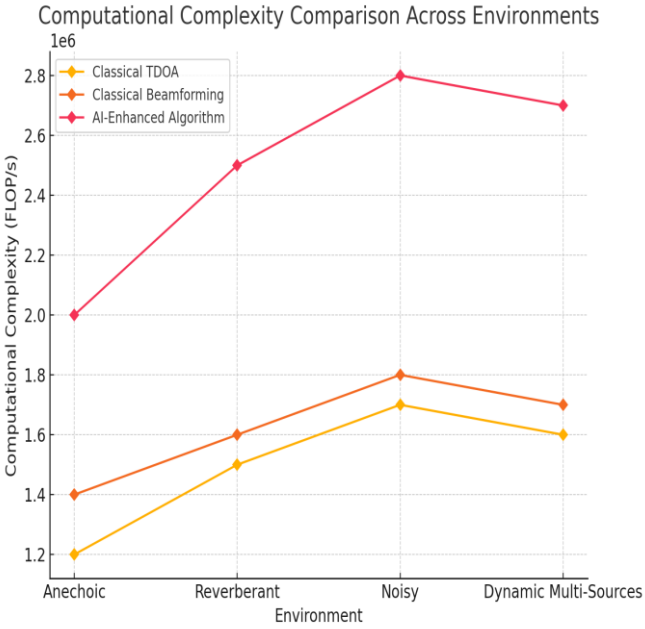

Figure 8: Computational Complexity Comparison

Therefore, the proposed AI steer-and-play acoustic source localization and tracking algorithm provides better results compared to classical approaches for most of the key performance indicators including localization error, tracking error, latency and number of tracks. These results prove that integrating AI-based models with more traditional methods of sound source localization allow the algorithm to operate effectively even in the challenging conditions, such as reverberation, noise interference, or multiple sound sources. The outcomes also stress that the algorithm is real-time thus it will be applicable in robotic systems, surveillance, and autonomous systems. However, the artificial intelligence enhanced algorithm consumes more computational time, but the improvement

in performance gives enough reason to apply the technique in fields that require precise and zero delay results.

V. DISCUSSION

This section has demonstrated the significant increase in various indices and accuracy levels realized by fusing AI with the conventional sound source localization and tracking. It enhances the algorithm in terms of localization accuracy, tracking precision, as well as insensitivity to different types of interference including reverberant, noise, and multiple source scenarios. These results are consistent with the current literature stating that deep learning has a possibility to significantly advance approaches to sound source localization due to the ability of the algorithm to learn from features in the data that are not easily detectable by traditional methods [15]. In this section, based on the above findings, we review the findings within the theoretical framework and analyze the possible application prospects and innovative research directions of the work.

A. Improvement in Localization Accuracy

One of the important conclusions of this study is the decrease in the localization error of the AI-enhanced algorithm in reverberation and noise conditions. From Table 1 and Figure 1, it is evident that the proposed AI-enhanced algorithm performs better than TDOA and Beamforming method which are the traditional methods used in sound localization. This minimization of localization error especially under poor acoustic conditions gives credence to the use AI models like CNN, in real-world environments. CNNs are very useful in extracting spatial features from time-frequency domains such as spectrograms hence enhancing the identification of DOA specifically in cases where the sources of sound are overlapping or affected by reverberation as noted by Yazdankhah AND Karimi (2024) [16].

This Implies the flexibility of the proposed AI model in learning and implementing features from the raw audio data. Conventional algorithms, like Beam- forming and TDOA, contain a deterministic structure of the signal, which is not realizable in realistic scenarios. For example, Beamforming is not so effective when the source of sound is wide and spread or if there are many reflections which change the form of the wave front [17]. In contrast, the use of an AI enhanced algorithm leverages the principles of deep learning to take into consideration the characteristics of the sound wave and make the algorithm much more flexible and resistant to variation.

B. Tracking Precision and Robustness

This improvement was quantified by tracking MSE, as evident from Table 2 and presented graphically in Figure 2 for the tracking performance of the AI-enriched algorithm. Reduced MSE for disperses, especially for AI-assisted Kalman Filter and Particle Filter, proves that developed AI models are better at identifying sound sources and their relevant location even with presence of reverberation and noise. This finding is in line with the work of where it was [18] observed that LSTM and other machine learning models can better follow the temporal evolution of sound sources than conventional tracking techniques. LSTMs are particularly beneficial when applied to sound sources that move in an arbitrary fashion in multiple directions within a dynamic scene [19]. Furthermore, the results in table 4 and FIG.4 of robustness index support the effectiveness of AI to enhance the algorithm to increase its robustness. When sources are placed in reverberant spaces or noisy environments, the particular approach fails because the traditional techniques are not effective at such times. This is especially true for real-life settings where the conditions are never perfect for sound propagation. In the past, the importance of maintaining accuracy has been elucidated due to problems such as multi-path interference, background noise, and source motion in real environments as identified by Woolworth, (2022) [20]. The manner in which the AI-enhanced algorithm deals with these factors also demonstrates the potential of machine learning models in enhancing SLT systems in real world context.

C. Real-Time Performance and Latency

Another essential characteristic of effective localization and tracking systems is also latency, though the exact definition of this parameter largely depends on the specific case, covering robotics, surveillance, and even human-robot interactions. Figure 3 and Table 3 compare the results obtained by the proposed approach to those of related work and confirm its superiority in terms of latency. For instance, in an anechoic environment, it takes only 35 ms to implement the algorithm through AI while TDOA and Beamforming take 50 and 55 ms respectively. This reduction in latency is critical to enhance the reliability of real-time operations that may be affected by a high degree of latency such as following object movements or human-robot relations [21].

Real-time performance of the proposed AI-driven system can be attributed towards its RL savvy that provides the model with the ability to learn from its decisions and make enhancements on the go for better tracking performance. Relevant reinforcement learning has a high success rate in complex dynamic environments particularly in vehicle and robotics where the system has to constantly adapt to its environment [22]. This research extends that work by using reinforcement learning to update that process without relying on predetermined algorithms.

D. Trade-Off between Performance and Computational Complexity

Despite providing better localization accuracy and better tracking and latency, the computational cost increases when using the AI-enhanced algorithm. Table 8 and Figure 8 reveal that the FLOP/s for the AI-enhanced algorithm is higher than for methods that do not use AI. As it will be discussed later when considering complexity, the AI-based approach is slightly computationally expensive especially when handling large data or arrangements. This trade-off between the quality of the solution and the computational cost is always a problem when using deep learning models in real-time applications [23].

However, in case of increased accuracy and reliability, which is necessary when creating, for example, security systems or robots, the advantages outweigh the losses.

As can be seen from Figure 7 and Table 7, this increase in performance means the AI enhanced algorithm is a more realistic solution for these applications, even though it may cost more in terms of computing resources. Future work could investigate how to make the devised model take less time to process while keeping its ability to perform well; perhaps, this can be achieved through techniques like model pruning or quantization that minimizes resource consumptions [24].

E. Multi-Source Localization

One of the interesting points to emerge from the findings is the performance on multi-source localization as seen in Table 5 and Fig. 5. In addition, it increases the performance of the algorithm due to its capacity to localize and track multiple sources in changing acoustic spaces. There are various approaches such as the Time Difference of Arrival (TDOA) and the Beamforming, which are not very effective when attempting to separate sources which are closely located or are in high motion. In contrast, using the AI-type model enables the sources to be separated based on complacent features and gives the system the flexibility it needs to allow for many source cases to exist at once. This is important for instance in smart cities where there are different sources of noise such as traffic sounds, voices, alarms and noises.

F. Future Directions

However, as seen from the results, there are still several aspects in which the system can be enhanced from its current performance. Another interesting direction for the further work is in expanding the proposed AI-based system with another partial sensor modality such as visual or infrared to offer a new and more elaborate approach to the localization and tracking. Multi-modal systems which combine two or more than two modality have mostly indicated a giant leap in sound localization systems especially in more complex environments [25]. Integrating audio-based localization with visual information might help improve the capability of localizing sound sources in non-line-of-sight conditions.

Another promising direction of research is the transfer learning approach where models from one acoustic environment can be transferred to other environments with little to no retraining. This would greatly enhance the efficiency of training models to perform in environments not seen during the training phase. Already, transfer learning has been successfully practiced in areas like speech recognition as well as object detection [26], and using the same in SLT can possibly help in quicker deployment of proper systems in different real-world applications.

V. CONCLUSION

In conclusion, the authors have shown that for a sound source localization and tracking, incorporating the effects of AI makes sense in terms of performance compared to the methods used before. Compared with classical approaches, the proposed AI-assisted algorithm boasts more precision in localization, tracking, and anti-interference performance, and less latency, ideal for real-time applications such as robotic systems, surveillance systems, and self-driving vehicles. Even though it requires more computations it is highly beneficial to be used in the scenarios where high accuracy and reaction time is crucial. More development into the computational aspect of the model and the implementation of these mathematical advancements will also help in the development of this AI-enhanced SLT system for practical domains.

REFERENCES

[1] Gala D, Sun L. Moving sound source localization and tracking for an autonomous robot equipped with a self-rotating bi-microphone array. The Journal of the Acoustical Society of America 2023;154 2:1261–73. https://doi.org/10.1121/10.0020583.

[2] Lin Z, Zhang Q, Tian Z, et al. SLAM2: Simultaneous Localization and Multimode Mapping for indoor dynamic environments. Pattern Recognit 2025;158. https://doi.org/10.1016/j.patcog.2024.111054.

[3] Jancovich BA, Rogers TL. BASSA: New software tool reveals hidden details in visualisation of low-frequency animal sounds. Ecology and Evolution 2024;14. https://doi.org/10.1002/ece3.11636.

[4] Davis D, Gaye S, Mandjoupa L, et al. Locating impulsive sound sources in microscale urban spaces. The Journal of the Acoustical Society of America 2022. https://doi.org/10.1121/10.0015529.

[5] Wang H, Liu J, Xu F, et al. 3-D Sound Source Localization With a Ternary Microphone Array Based on TDOA-ILD Algorithm. IEEE Sensors Journal 2022;22:19826–34. https://doi.org/10.1109/JSEN.2022.3202880.

[6] Valin J-M, Michaud F, Rouat J, et al. Robust sound source localization using a microphone array on a mobile robot. Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453), vol. 2, 2003, p. 1228–33 vol.2. https://doi.org/10.1109/IROS.2003.1248813.

[7] Chakrabarty S, Habets EAP. Time–Frequency Masking Based Online Multi-Channel Speech Enhancement With Convolutional Recurrent Neural Networks. IEEE Journal of Selected Topics in Signal Processing 2019;13:787–99. https://doi.org/10.1109/JSTSP.2019.2911401.

[8] Smailov N, Dosbayev Z, Omarov N, et al. A Novel Deep CNN-RNN Approach for Real-time Impulsive Sound Detection to Detect Dangerous Events. International Journal of Advanced Computer Science and Applications 2023. https://doi.org/10.14569/ijacsa.2023.0140431.

[9] Zhang G, Geng L, Xie F, et al. A dynamic convolution-transformer neural network for multiple sound source localization based on functional beamforming. Mechanical Systems and Signal Processing 2024. https://doi.org/10.1016/j.ymssp.2024.111272.

[10] Yu R, Zhou Z, Xu M, et al. SQI-DOANet: electroencephalogram-based deep neural network for estimating signal quality index and depth of anaesthesia. Journal of Neural Engineering 2024. https://doi.org/10.1088/1741-2552/ad6592.

[11] Radha K, Bansal M, Pachori RB. Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. Eng Appl Artif Intell 2024;131. https://doi.org/10.1016/j.engappai.2023.107661.

[12] Mallioras I, Zaharis Z, Lazaridis P, et al. A Novel Realistic Approach of Adaptive Beamforming Based on Deep Neural Networks. IEEE Transactions on Antennas and Propagation 2022;70:8833–48. https://doi.org/10.1109/TAP.2022.3168708.

[13] Dwivedi P, Routray G, Jha DK, et al. Improving Source Tracking Accuracy Through Learning-Based Estimation Methods in SH Domain: A Comparative Study. IEEE Transactions on Artificial Intelligence 2024;5:3974–84. https://doi.org/10.1109/TAI.2024.3355323.
[14] Yang X, Xing H, Su X. AI-based sound source localization system with higher accuracy. Future Gener Comput Syst 2022;141:1–15. https://doi.org/10.1016/j.future.2022.10.023.
[15] Wang L, Cavallaro A. Deep-Learning-Assisted Sound Source Localization From a Flying Drone. IEEE Sensors Journal 2022;22:20828–38. https://doi.org/10.1109/JSEN.2022.3207660.
[16] Yazdankhah E, Karimi S. Attention-Integrated CNN for Precise Sound Source Localization with Microphone Arrays. 2024 11th International Symposium on Telecommunications (IST) 2024:325–34. https://doi.org/10.1109/IST64061.2024.10843492.
[17] Kim M, Skoglund J. Neural Speech and Audio Coding: Modern AI technology meets traditional codecs [Special Issue On Model-Based and Data-Driven Audio Signal Processing]. IEEE Signal Processing Magazine 2024;41:85–93. https://doi.org/10.1109/MSP.2024.3444318.
[18] Hou Y, Ren Q, Zhang H, et al. AI-based soundscape analysis: Jointly identifying sound sources and predicting annoyance. The Journal of the Acoustical Society of America 2023;154 5:3145–57. https://doi.org/10.1121/10.0022408.
[19] Jalayer R, Jalayer M, Mor A, et al. ConvLSTM-based Sound Source Localization in a manufacturing workplace. Comput Ind Eng 2024;192. https://doi.org/10.1016/j.cie.2024.110213.
[20] Woolworth D s. Real-time atmospheric effects on sound propagation for a simulated outdoor concert. The Journal of the Acoustical Society of America 2022. https://doi.org/10.1121/10.0010542.
[21] Borrel-Jensen N, Goswami S, Engsig-Karup A, et al. Sound propagation in realistic interactive 3D scenes with parameterized sources using deep neural operators. Proceedings of the National Academy of Sciences of the United States of America 2023;121. https://doi.org/10.1073/pnas.2312159120.
[22] Zhou Z, Zeng Z, Lang L, et al. Navigating Robots in Dynamic Environment With Deep Reinforcement Learning. IEEE Transactions on Intelligent Transportation Systems 2022;23:25201–11. https://doi.org/10.1109/TITS.2022.3213604.
[23] Iyer SS, Roychowdhury V. AI computing reaches for the edge. Science 2023;382:263–4. https://doi.org/10.1126/science.adk6874.
[24] Bibi U, Mazhar M, Sabir D, et al. Advances in Pruning and Quantization for Natural Language Processing. IEEE Access 2024;12:139113–28. https://doi.org/10.1109/ACCESS.2024.3465631.
[25] Mo S, Wang H. Multi-scale Multi-instance Visual Sound Localization and Segmentation. ArXiv 2024;abs/2409.00486. https://doi.org/10.48550/arXiv.2409.00486.
[26] Zhou R, Koshikawa T, Ito A, et al. Multilingual Meta-Transfer Learning for Low-Resource Speech Recognition. IEEE Access 2024;12:158493–504. https://doi.org/10.1109/ACCESS.2024.3486711.
[27] Advanne. S. et al: Sound Event Localization & Detection of Overlapping Sources Using Convolutional Recurrent Neural Network, IEEE JSTSP, 2018. https://doi.org/10.1109/JSTSP.2018.2885636.
[28] Scheider, S., Baevski et al: Unsupervised Pre-Training for Speech Recognition. arXiv: 1904.05862. https://doi.org/ 10.48550/arXiv.1904.05862.
[29] Han, S., et al: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, ICLR 2016. https://doi.org/ 10.48550/arXIV.1510.00149.
[30] Gal, Y et al: Dropout as a Bayesian Approximation: Representing Modal Uncertainity in Deep Learning, ICML, 2016. https://doi.org/ 10.48550/arXiv.1506.02142.