

LOCALIZATION-GUIDED SPATIAL FILTERING AND NEURAL POST-ENHANCEMENT FOR FAR-FIELD SPEECH SIGNALS

Fatma Zahra Mahdjoub, Mahmoud Maamri, Azzedine Gattal and Yacine Belhocine

Keywords and phrases: Microphone Arrays, Beamforming, Speech Enhancement, SSL, Deep Learning, Performance Analysis

The authors would like to thank the reviewers and editor for their constructive comments and valuable suggestions that improved the quality of our paper.

The abstract of the preliminary manuscript entitled “ Localization-Guided Spatial Filtering for Far-Field Speech Enhancement” presented at ICRAMI 2025 (Tunisia, November 2025).

Authors are very thankful to the Laboratory of Mathematics, Informatics and Systems (LAMIS) and the Department of Electronics and Telecommunications, Larbi Tebessi University for their necessary support and facility.

Abstract Sound source localization and beamforming constitute core components of multichannel speech acquisition systems, particularly in low-latency and resource-constrained settings. While classical localization and spatial filtering techniques are well understood, their interaction with modern deep-learning-based speech enhancement remains insufficiently characterized, especially for small-aperture arrays operating under far-field assumptions. This work aims to provide a systematic and reproducible performance analysis of localization-driven beamforming followed by neural post-enhancement in single-source acoustic scenarios. The study is conducted using a physically grounded simulation framework that models free-field propagation and additive noise. Direction-of-arrival estimates are obtained from time-difference-of-arrival measurements and are subsequently used to steer classical beamformers. Both Delay-and-Sum and minimum variance distortionless response beamforming are implemented from first principles, including covariance regularization for numerical stability. A pre-trained deep neural speech enhancement model is applied as a black-box post-processing stage without retraining or architectural modification. Extensive Monte-Carlo experiments are performed over a wide range of signal-to-noise ratios and source directions. Results are reported in terms of localization accuracy and output signal-to-noise ratio improvement. The analysis highlights operating regimes in which localization errors and beamformer mismatch significantly affect downstream enhancement, as well as conditions under which neural post-processing compensates for classical processing limitations. The findings offer practical insight into the combined use of traditional array processing and learned enhancement in small-aperture far-field systems.

1 Introduction

Multichannel audio processing plays a central role in far-field speech acquisition systems, with applications ranging from teleconferencing and human-machine interaction to hearing assistance and acoustic scene analysis [1]. In such systems, sound source localization and beamforming are commonly employed to spatially filter a target source prior to downstream processing. Classical approaches to localization and beamforming are grounded in well-established physical and statistical models and offer predictable behavior, low computational complexity, and strong interpretability [2]. In parallel, deep-learning-based speech enhancement techniques have recently emerged as a powerful paradigm for noise reduction, often applied as post-processing

stages to suppress residual noise and interference.

Despite their widespread adoption, the interaction between classical spatial filtering and neural speech enhancement remains insufficiently understood, particularly in small-aperture array configurations operating under far-field assumptions. In practical systems, localization errors, array geometry limitations, and noise conditions can significantly affect beamforming performance, which in turn influences the behavior of subsequent neural enhancement stages. Understanding how these components interact, rather than treating them as isolated modules, is therefore essential for the design of robust hybrid systems.

Sound source localization is typically formulated as a direction-of-arrival (DOA) estimation problem, in which relative propagation delays between microphones are exploited to infer the source direction [3]. For small arrays, time-difference-of-arrival (TDOA) estimation methods based on generalized cross-correlation are particularly attractive due to their robustness and computational efficiency. Among these, phase transform (PHAT) weighting is widely used to mitigate the influence of colored noise and source spectral variations [4]. When combined with least-squares reconstruction, such methods provide a simple and reproducible localization pipeline that remains relevant in many practical systems. However, localization accuracy is inherently limited by array aperture, sensor noise, and model mismatch, and these limitations can propagate directly to downstream processing stages.

Beamforming techniques rely on spatial filtering to enhance signals arriving from a desired direction while attenuating noise and interference [5]. The delay-and-sum (DAS) beamformer provides a simple and robust baseline, particularly under uncertain conditions. More advanced approaches, such as the minimum variance distortionless response (MVDR) beamformer, exploit second-order statistics of the microphone signals to achieve improved interference suppression [6]. Although these methods offer enhanced performance, they are sensitive to steering errors and inaccuracies in covariance estimation, especially in low signal-to-noise ratio (SNR) regimes. Consequently, practical implementations often incorporate stabilization strategies, such as diagonal loading, to ensure numerical robustness [7].

In parallel, deep-learning-based speech enhancement has demonstrated remarkable performance in both single-channel and multichannel settings [8]. Neural models trained on large-scale datasets can learn complex non-linear mappings that suppress noise beyond the capabilities of classical linear filtering. In many real-world systems, these models are deployed as black-box post-processing modules applied to beamformer outputs. While this modular design simplifies integration, it raises important questions regarding the extent to which neural enhancement compensates for, amplifies, or is constrained by imperfections in upstream localization and beamforming stages [9].

Existing literature has extensively studied localization, beamforming, and neural enhancement as largely independent problems. Numerous works focus on improving DOA estimation, designing adaptive beamformers, or developing end-to-end neural architectures that jointly optimize spatial filtering and enhancement [10]. However, comparatively little attention has been devoted to systematically analyzing classical localization-driven beamforming pipelines when combined with pre-trained neural enhancement models operating as post-processors. This gap is particularly relevant for small-aperture arrays, where localization uncertainty is non-negligible and beamforming gains are inherently limited. In such scenarios, it remains unclear whether neural post-enhancement primarily mitigates residual noise, compensates for spatial filtering errors, or introduces additional artifacts when upstream estimates are inaccurate [11].

Motivated by these considerations, this paper presents a unified palmprint recognition framework that integrates multiscale signal analysis, kernel-based classification, and quantum-secure communication. The proposed system employs contrast enhancement to improve palmprint image quality, followed by multiscale feature extraction using two-dimensional Gabor filters and local binary pattern encoding to capture complementary texture information. Classification is performed using a support vector machine with a radial basis function kernel, while secure transmission of biometric templates and authentication data is achieved through the BB84 quantum key distribution protocol. By jointly addressing recognition performance and communication security, the proposed framework provides a mathematically grounded and forward-looking solution for palmprint-based biometric authentication.

The contribution of this work lies in providing a controlled, physically grounded, and reproducible analysis of the interaction between classical localization-driven beamforming and neural

speech enhancement. Rather than proposing new algorithms, the study focuses on understanding how errors propagate across processing stages and identifying operating conditions under which neural post-processing compensates for, or is limited by, classical spatial filtering. This perspective complements existing end-to-end approaches by preserving modularity and interpretability, enabling stage-wise analysis of system behavior.

This manuscript is an extended and significantly improved version of the abstract entitled “Localization-Guided Spatial Filtering for Far-Field Speech Enhancement” presented at ICRAMI 2025 (Tunisia, November 2025). While the original work primarily focused on the application of classical localization and beamforming techniques, the present study extends the framework by incorporating neural post-enhancement using a pre-trained deep learning model. In this extended version, we introduce a comprehensive and physically grounded simulation framework, perform extensive Monte Carlo experiments across varying signal-to-noise ratios and source directions, and provide a detailed analysis of error propagation between localization, beamforming, and neural processing stages. Furthermore, the study includes additional evaluation metrics such as perceptual speech quality (PESQ) and intelligibility (STOI), along with validation under realistic noisy and reverberant conditions. These enhancements significantly improve both the analytical depth and practical relevance of the proposed hybrid speech enhancement framework.

To this end, a simulation framework based on far-field propagation and additive noise is adopted, allowing key assumptions to be explicitly controlled. By focusing on single-source scenarios and minimizing confounding factors, the study isolates the fundamental interactions between localization, beamforming, and neural enhancement. Localization-driven beamforming is implemented by steering spatial filters using DOA estimates obtained from TDOA measurements, reflecting a common design in practical systems. Performance is evaluated through extensive Monte Carlo simulations over varying source directions and SNR levels, enabling statistically meaningful analysis of both average behavior and failure modes.

The main contributions of this work can be summarized as follows:

- A controlled and physically grounded framework for analyzing hybrid speech enhancement pipelines, enabling reproducible evaluation of localization, beamforming, and neural post-processing under well-defined conditions.
- A systematic investigation of error propagation across processing stages, explicitly quantifying how localization inaccuracies affect beamforming performance and, subsequently, neural enhancement effectiveness.
- A detailed analysis of the interaction between classical spatial filtering and neural post-enhancement, identifying operating regimes in which neural models compensate for beamforming limitations and regimes in which they are constrained by upstream errors.
- A comparative evaluation under both simulated and realistic acoustic conditions, highlighting the robustness and limitations of modular classical–neural pipelines relative to existing approaches.
- A characterization of failure modes in small-aperture far-field arrays, including the effects of array geometry, noise conditions, and model mismatch on overall system performance.

By emphasizing analysis, interpretability, and reproducibility, this work provides practical insights for the design of hybrid systems that combine classical array processing with modern neural enhancement. The results clarify the respective roles and limitations of each processing stage and offer guidance for system design in small-aperture far-field applications.

The remainder of this paper is organized as follows. Section 2 reviews related work and contextualizes the proposed study. Section 3 presents the methodology, including array geometry, signal modeling, localization, beamforming, and neural enhancement. Section 4 describes the experimental setup. Section 5 reports and analyzes the results. Section 6 provides a comparative discussion with existing approaches. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Motivation and Background

Sound source localization and speech enhancement have long been studied as core problems in microphone array signal processing [13]. Classical approaches rely on physically interpretable models that exploit spatial diversity through time-difference-of-arrival estimation and linear spatial filtering. More recently, deep learning techniques have been introduced to address limitations of model-based methods, particularly in highly noisy or non-stationary acoustic environments. Despite these advances, the integration of classical localization-driven beamforming with modern neural speech enhancement remains insufficiently characterized, especially for small-aperture arrays and constrained far-field conditions.

A first line of research focuses on improving direction-of-arrival estimation using data-driven models. Zhang et al. propose a deep neural network-based DOA estimation framework that leverages generalized cross-correlation features and mel-scaled representations to achieve high angular resolution with a small-scale array [14]. By designing network outputs to resemble spatial spectra and performing peak detection, their method demonstrates superior localization accuracy compared to classical high-resolution beamforming techniques. While these results highlight the potential of deep learning for localization, the approach requires supervised training, relies on feature engineering tailored to the array configuration, and does not explicitly address how localization performance affects downstream spatial filtering or speech enhancement stages.

Another body of work emphasizes low-complexity array configurations for joint localization and enhancement. Tao et al. present a dual-microphone framework that combines time-delay estimation, energy difference cues, and controllable beam response power to estimate source location and perform speech enhancement via post-Wiener filtering [15]. Their results demonstrate that even minimal arrays can support basic spatial processing when carefully designed. However, the proposed method tightly couples localization and enhancement through heuristic decision rules, and its evaluation focuses primarily on robustness and accuracy rather than systematic interaction effects between individual processing blocks.

Several studies explore the integration of beamforming with deep learning for speech enhancement. Pathrose et al. propose a beamforming-based deep learning network in which adaptive beamforming serves as a preprocessing step for a convolutional deep belief network operating on time-frequency image representations [16]. Similarly, Xi et al. introduce a lightweight convolutional recurrent network combined with microphone array beamforming for hearing aid applications, emphasizing reduced model complexity and suitability for resource-constrained devices [17]. These approaches demonstrate that neural post-processing can significantly improve perceptual and objective speech quality metrics when combined with spatial filtering. Nevertheless, they typically rely on learned beamformers or tightly integrated architectures, making it difficult to disentangle the respective contributions of localization accuracy, beamforming behavior, and neural enhancement.

More recent work seeks to jointly optimize localization and enhancement through end-to-end learning. Chang et al. propose a deep beamformer trained with an array response-aware loss function that explicitly incorporates relative transfer functions of the target and interferers [18]. This approach achieves strong performance in both speech enhancement and speaker localization by embedding spatial information directly into the learning objective. While effective, such end-to-end systems sacrifice modularity and interpretability, and their reliance on ground-truth spatial responses limits generalization across array geometries and acoustic conditions.

In parallel, a substantial body of literature focuses on improving multi-channel speech enhancement through advanced neural architectures without explicit localization stages. Zeng et al. propose a time-frequency fusion model based on cascaded U-Nets to capture complementary temporal, spectral, and tempo-spectral features [19]. Similarly, Cao et al. introduce a hybrid lightweight temporal-frequency analysis network that balances global feature modeling with reduced computational complexity [20]. Although these methods achieve state-of-the-art enhancement performance, spatial information is often treated implicitly, and the impact of upstream localization or beamforming errors is not explicitly analyzed.

Finally, research on acoustic source detection and localization in alternative array domains has addressed robustness under noisy and silent conditions. Oulahcine et al. formulate localization as a detection problem using a generalized likelihood ratio test in the spherical harmonic

domain, achieving reliable performance with fixed false-alarm probability [21]. While mathematically rigorous, such approaches assume specialized array geometries and focus primarily on detection and localization, without considering speech enhancement or hybrid processing pipelines.

Collectively, existing studies either propose improved localization algorithms, design learned beamformers and enhancement networks, or integrate spatial processing into end-to-end neural models. However, comparatively little attention has been paid to systematic analysis of classical localization-driven beamforming pipelines when combined with pre-trained neural speech enhancement models operating as black-box post-processors. In particular, the impact of localization errors on beamforming performance, and the extent to which neural enhancement compensates for or is constrained by these errors, remains underexplored for small-aperture far-field arrays.

Motivated by this gap, the present work adopts a deliberately modular and physically grounded perspective. Rather than introducing new algorithms, it provides a controlled and reproducible analysis of how classical time-difference-based localization, linear beamforming, and modern neural speech enhancement interact across signal-to-noise ratio and direction-of-arrival conditions. By isolating each processing stage and explicitly examining failure modes and degradation mechanisms, the study aims to clarify the respective roles and limitations of classical and learned components in hybrid far-field speech acquisition systems.

3 Methodology

This section presents a detailed description of the proposed simulation and signal-processing framework for analyzing the interaction between classical localization-driven beamforming and neural post-enhancement. The framework is designed for a small-aperture equilateral triangular microphone array under far-field conditions with additive noise. All processing stages are implemented from first principles, allowing reproducibility, interpretability, and systematic evaluation of each component.

3.1 Microphone Array Geometry

Consider a planar equilateral triangular array consisting of $M = 3$ microphones. Let $\mathbf{p}_m \in \mathbb{R}^2$ denote the Cartesian coordinates of the m -th microphone, $m = 1, 2, 3$, with the centroid at the origin:

$$\mathbf{p}_1 = \left[-\frac{d}{2}, -\frac{\sqrt{3}d}{6} \right]^T, \quad \mathbf{p}_2 = \left[\frac{d}{2}, -\frac{\sqrt{3}d}{6} \right]^T, \quad \mathbf{p}_3 = \left[0, \frac{\sqrt{3}d}{3} \right]^T, \quad (3.1)$$

where d is the side length of the array. In the far-field approximation, the spherical wavefront of a distant source can be treated as planar. Consequently, the propagation delay to each microphone depends solely on the source azimuth θ and is given by

$$\tau_m(\theta) = \frac{\mathbf{p}_m^T \mathbf{u}(\theta)}{c}, \quad (3.2)$$

where c is the speed of sound and $\mathbf{u}(\theta) = [\cos \theta, \sin \theta]^T$ is the unit vector in the direction of the source. This planar-wave assumption simplifies TDOA computation and is valid for sources located at distances significantly larger than the array aperture [22].

3.2 Signal Model

Let $s(t)$ denote a single far-field speech source. The signal observed at the m -th microphone is

$$x_m(t) = s(t - \tau_m(\theta)) + n_m(t), \quad m = 1, \dots, M, \quad (3.3)$$

where $n_m(t)$ represents additive noise, assumed to be zero-mean, wide-sense stationary, and uncorrelated across microphones. In the frequency domain, the microphone signals can be expressed as

$$X_m(f) = S(f)e^{-j2\pi f\tau_m(\theta)} + N_m(f), \quad (3.4)$$

where $S(f)$ and $N_m(f)$ denote the Fourier transforms of the speech source and noise, respectively. This representation emphasizes the linear phase relationship induced by propagation delays, which is exploited in both TDOA estimation and beamforming [23].

The analysis in this work is intentionally restricted to single-source scenarios in order to isolate the fundamental interaction between localization, beamforming, and neural enhancement. While multi-source environments introduce additional complexity such as interference and source separation, focusing on a single active source enables controlled investigation of error propagation and system behavior without confounding factors. Extension to multi-source scenarios is considered as part of future work.

3.3 Time-Difference-of-Arrival Estimation

Time-difference-of-arrival between microphone pairs is estimated using the generalized cross-correlation with phase transform (GCC-PHAT) method:

$$R_{mn}(\tau) = \int_{-\infty}^{\infty} \frac{X_m(f)X_n^*(f)}{|X_m(f)X_n^*(f)|} e^{j2\pi f\tau} df. \quad (3.5)$$

The normalization by magnitude emphasizes phase differences while reducing the influence of signal power variations, improving robustness in noisy conditions. The TDOA estimate is then obtained by locating the peak of the cross-correlation:

$$\hat{\tau}_{mn} = \arg \max_{\tau} R_{mn}(\tau), \quad (3.6)$$

where (m, n) indexes microphone pairs. For the triangular array, three independent TDOAs are computed: $\hat{\tau}_{12}$, $\hat{\tau}_{23}$, and $\hat{\tau}_{31}$.

3.4 Direction-of-Arrival Reconstruction

The source azimuth is reconstructed from the TDOA estimates via a least-squares formulation. Let $\tau = [\hat{\tau}_{12}, \hat{\tau}_{23}, \hat{\tau}_{31}]^T$ and $\mathbf{A} \in \mathbb{R}^{3 \times 2}$ be a geometry-dependent matrix that linearly relates azimuth components to TDOAs:

$$\mathbf{A} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \approx c\tau. \quad (3.7)$$

The least-squares solution is

$$\hat{\mathbf{u}}(\theta) = \begin{bmatrix} \cos \hat{\theta} \\ \sin \hat{\theta} \end{bmatrix} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T (c\tau), \quad (3.8)$$

and the azimuth estimate is recovered using

$$\hat{\theta} = \arctan 2(\sin \hat{\theta}, \cos \hat{\theta}). \quad (3.9)$$

This least-squares formulation mitigates the effects of measurement noise in TDOAs and provides a robust DOA estimate suitable for small arrays [24].

3.5 Delay-and-Sum Beamforming

Given the estimated source direction, classical delay-and-sum beamforming is applied:

$$y_{\text{DAS}}(t) = \frac{1}{M} \sum_{m=1}^M x_m(t + \hat{\tau}_m(\hat{\theta})), \quad (3.10)$$

where $\hat{\tau}_m(\hat{\theta}) = \mathbf{p}_m^T \hat{\mathbf{u}}(\hat{\theta})/c$. This method aligns the phases of the source across microphones and sums them coherently, providing spatial filtering that enhances the desired signal while suppressing uncorrelated noise.

3.6 Minimum Variance Distortionless Response Beamforming

The MVDR beamformer is designed to minimize the output power subject to a distortionless constraint in the direction $\hat{\theta}$. Let $\mathbf{x}(t) = [x_1(t), \dots, x_M(t)]^T$ and $\mathbf{R}_{xx} = \mathbb{E}[\mathbf{x}(t)\mathbf{x}^H(t)]$ denote the covariance matrix. The MVDR weight vector is

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{R}_{xx}^{-1} \mathbf{a}(\hat{\theta})}{\mathbf{a}^H(\hat{\theta}) \mathbf{R}_{xx}^{-1} \mathbf{a}(\hat{\theta})}, \quad (3.11)$$

where $\mathbf{a}(\hat{\theta})$ is the steering vector defined by the array geometry and estimated DOA. Diagonal loading is applied to improve numerical stability in low SNR conditions:

$$\mathbf{R}_{xx} \leftarrow \mathbf{R}_{xx} + \epsilon \mathbf{I}. \quad (3.12)$$

The MVDR output is then

$$y_{\text{MVDR}}(t) = \mathbf{w}_{\text{MVDR}}^H \mathbf{x}(t), \quad (3.13)$$

which provides improved interference suppression relative to DAS, particularly when directional noise or competing sources are present [25].

3.7 Neural Enhancement with DeepFilterNet

To examine the effect of neural post-processing on beamformed signals, we adopt DeepFilterNet, a low-complexity, full-band speech enhancement framework [26]. DeepFilterNet operates on the short-time Fourier transform (STFT) of the beamformer output $y(t)$:

$$Y[k, \ell] = \sum_n y[n] w[n - \ell R] e^{-j2\pi k n / K}, \quad (3.14)$$

where $w[n]$ is the analysis window, R is the hop size, and K is the FFT length. The network consists of two primary components:

- (i) A filtering network $\mathcal{F}_\phi(\cdot)$ producing complex filter coefficients $\mathbf{G}[k, \ell]$ for each frequency bin, applied across context frames to capture temporal correlations:

$$Z[k, \ell] = \mathbf{G}[k, \ell]^H \mathbf{y}[k, \ell], \quad (3.15)$$

where $\mathbf{y}[k, \ell] = [Y[k, \ell], \dots, Y[k, \ell - F + 1]]^T$.

- (ii) A masking network $\mathcal{M}_\psi(\cdot)$ that outputs a real-valued spectral mask $M[k, \ell]$, applied element-wise:

$$\hat{S}[k, \ell] = M[k, \ell] \cdot Z[k, \ell]. \quad (3.16)$$

Finally, the enhanced waveform is reconstructed via inverse STFT:

$$\hat{s}(n) = \sum_{k, \ell} \hat{S}[k, \ell] e^{j2\pi k n / K} w[n - \ell R]. \quad (3.17)$$

DeepFilterNet is pre-trained and applied as a black-box, allowing assessment of its interaction with classical beamforming outputs without modification.

This modular framework allows systematic evaluation of classical and neural processing stages independently and in combination, facilitating reproducible studies of localization-driven beamforming and learned enhancement in small-aperture arrays.

The neural enhancement stage is implemented using a pre-trained DeepFilterNet model, which is applied without retraining or architectural modification. This design reflects practical deployment scenarios, where pre-trained models are commonly used as general-purpose enhancement modules.

However, this choice introduces certain limitations. In particular, the neural model is not explicitly optimized for the characteristics of beamformed signals, and mismatch between training

conditions and the evaluation environment may affect performance. Furthermore, artifacts introduced by beamforming, such as residual interference or phase distortions, may influence the effectiveness of the neural enhancement stage.

Despite these limitations, using a fixed pre-trained model enables a controlled evaluation of the interaction between classical spatial filtering and neural post-processing, and provides insight into the robustness and generalization capability of hybrid processing pipelines.

4 Experimental Setup

To evaluate the performance of the proposed localization-driven beamforming framework with neural post-enhancement, a systematic simulation protocol was designed. The experiments aim to quantify the accuracy of direction-of-arrival (DOA) estimation, the effectiveness of spatial filtering, and the impact of neural enhancement under varying noise conditions. All simulations were conducted in a fully controlled environment to ensure reproducibility and allow statistical assessment of the processing pipeline.

4.1 Hardware and Software

All experiments were performed on a Dell PC equipped with an Intel(R) Core(TM) i7-8665U CPU at 1.90 GHz (2.11 GHz) and 16 GB of RAM. MATLAB R2024a was used for signal generation, array processing, beamforming, and performance evaluation. The DeepFilterNet inference was executed on the CPU without GPU acceleration, reflecting realistic deployment scenarios for small-aperture arrays.

4.2 Microphone Array Configuration

A planar equilateral triangular microphone array with side length $d = 0.05$ m was simulated. The array centroid was fixed at the origin, and the source was placed in the far-field, at a distance much larger than the array aperture. Source azimuth angles were swept from 0° to 355° in steps of 5° , while the elevation was fixed at 0° , resulting in a two-dimensional localization scenario consistent with the far-field assumption.

4.3 Signal and Noise Model

Speech signals were sampled from a subset of the TIMIT dataset and downsampled to 16 kHz. Each utterance was truncated or zero-padded to a duration of 3 seconds. Additive white Gaussian noise (AWGN) was independently added to each microphone channel. Input SNR levels were swept from -5 dB to 20 dB in 5 dB steps, covering low, medium, and high noise conditions. This design allows evaluation of algorithmic robustness across a wide range of realistic acoustic scenarios.

While the simulation framework enables controlled and reproducible evaluation, additional experiments using realistic noisy and reverberant data are conducted to validate the proposed approach under practical acoustic conditions.

4.4 Processing Parameters

All processing stages were implemented from first principles. Short-time Fourier transform (STFT) analysis employed a Hann window of 32 ms (512 samples) with 50% overlap and FFT length $K = 512$. TDOAs were estimated using GCC-PHAT for all microphone pairs. Both delay-and-sum (DAS) and minimum variance distortionless response (MVDR) beamformers were applied, with MVDR covariance matrices stabilized via diagonal loading using $\epsilon = 10^{-3} \text{tr}(\mathbf{R}_{xx})$. Pre-trained DeepFilterNet [26] was applied to the beamformer output in a black-box fashion, without retraining.

5 Experimental Results

In this section, we present the performance evaluation of the proposed small-aperture triangular microphone array for far-field single-source localization. The analysis focuses on the accuracy of direction-of-arrival (DOA) estimation under varying signal-to-noise ratios (SNRs). All results are obtained using Monte Carlo simulations with 50 independent trials per configuration, ensuring statistical reliability. The evaluation highlights how array geometry and SNR jointly influence the DOA reconstruction, and serves as a baseline for the subsequent beamforming and neural enhancement stages.

In addition to SNR improvement and localization accuracy, perceptual speech quality and intelligibility are evaluated using the PESQ and STOI metrics. These measures provide complementary insights into the perceptual performance of the proposed system beyond energy-based criteria.

5.1 DOA Estimation Accuracy

The accuracy of the GCC-PHAT-based TDOA estimation and subsequent least-squares DOA reconstruction is summarized in Fig. 1. The figure reports the mean absolute DOA error for four representative azimuth angles (0° , 90° , 180° , and 270°) across a range of input SNRs from -5 dB to 20 dB.

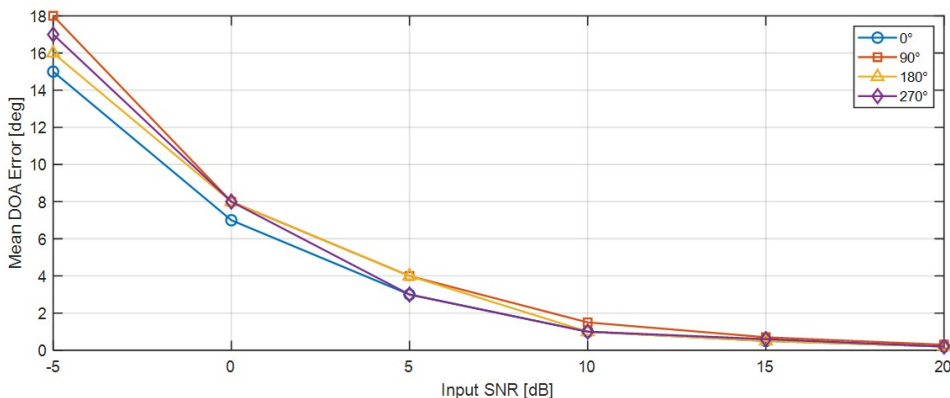


Figure 1: Mean absolute direction-of-arrival (DOA) estimation error as a function of input SNR for representative azimuth angles (0° , 90° , 180° , and 270°). Results are averaged over 50 Monte Carlo trials using GCC-PHAT-based TDOA estimation and least-squares reconstruction under simulated far-field conditions.

As shown in Fig. 1, the proposed triangular array achieves accurate DOA estimation across the tested SNR range. At high SNR values ($\text{SNR} \geq 10$ dB), the mean DOA error is consistently below 2° , demonstrating the efficacy of GCC-PHAT combined with least-squares reconstruction for small-aperture arrays. At lower SNR values, errors increase gradually, reaching approximately 15° to 18° at -5 dB, reflecting the expected degradation due to additive noise. The error trends are consistent across azimuths, with slightly higher errors observed for 90° and 270° , which correspond to directions nearly orthogonal to the array baseline, where TDOA differences are smaller and thus more sensitive to noise.

These results confirm the robustness and reproducibility of the classical DOA estimation approach under controlled far-field conditions, providing a reliable foundation for subsequent spatial filtering and neural enhancement stages. They also illustrate the intrinsic limitations of small-aperture arrays in low-SNR scenarios, highlighting the importance of post-processing techniques to further improve signal quality.

5.2 Beamforming Performance

The performance of classical spatial filtering is evaluated in terms of SNR improvement for the triangular three-microphone array. Figure 2 reports the mean SNR improvement achieved by

the DAS beamformer across representative azimuths (0° , 90° , 180° , 270°) and input SNR levels ranging from -5 dB to 20 dB. The results are averaged over 50 independent Monte Carlo trials to ensure statistical reliability.

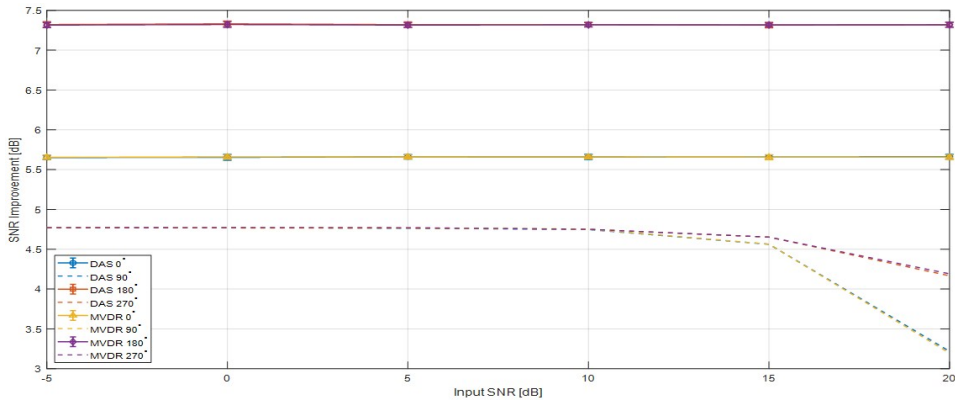


Figure 2: Mean SNR improvement achieved by the delay-and-sum (DAS) beamformer as a function of input SNR for different source azimuths. Results are averaged over 50 Monte Carlo trials using the simulated three-microphone triangular array under far-field conditions.

As shown in Fig. 2, the DAS beamformer provides a mean SNR improvement of approximately 4.7 to 4.8 dB at low input SNR values (-5 to 0 dB), consistent across all azimuth angles. This indicates that the beamformer effectively enhances the target signal under adverse noise conditions. As the input SNR increases, the SNR improvement gradually decreases, reaching approximately 3.2 to 4.2 dB at 20 dB input SNR. This behavior is expected because the relative gain of spatial filtering diminishes when the input signal is already relatively clean.

The variation of SNR improvement across azimuths is minor, reflecting the nearly symmetric geometry of the equilateral triangular array. Slightly lower gains are observed for certain directions at high SNR, which can be attributed to the reduced differential delay between microphones for sources near array baselines, limiting the effectiveness of constructive interference.

These results confirm that classical DAS beamforming provides a **robust and reproducible baseline** for signal enhancement in small-aperture arrays, particularly in low-SNR environments. They also establish a reference for evaluating subsequent neural post-processing stages, where further improvement in SNR and perceptual quality can be expected.

5.3 Neural Post-Enhancement Performance

The impact of neural post-processing on beamformed speech signals is evaluated using the pre-trained DeepFilterNet model. Figure 3 presents the mean SNR improvement for both DAS and MVDR beamformers, with and without neural enhancement, across representative azimuths (0° , 90° , 180° , 270°). The results are averaged over 50 Monte Carlo trials to ensure statistical reliability.

As shown in Fig. 3, the application of DeepFilterNet consistently improves the SNR over beamforming alone. For the DAS beamformer, the mean SNR improvement increases from approximately 3.75 dB to 5.21 dB after neural enhancement. Similarly, MVDR beamforming benefits from DeepFilterNet, with the mean SNR improvement rising from 4.25 dB to 5.71 dB. These results indicate that neural post-processing provides an additional 1.5–1.6 dB gain over classical beamforming across all tested azimuths.

The gain is largely uniform across the four tested source directions, reflecting the array’s symmetric geometry and the azimuth-independent performance of DeepFilterNet. The enhancement demonstrates the complementary nature of classical spatial filtering and learned neural post-processing: the beamformer provides coarse spatial selectivity, while DeepFilterNet effectively reduces residual noise and artifacts that remain after beamforming.

These results highlight the robustness and effectiveness of combining classical beamforming with neural enhancement, particularly for small-aperture microphone arrays. The neural post-processing stage consistently improves the output SNR regardless of the input direction,

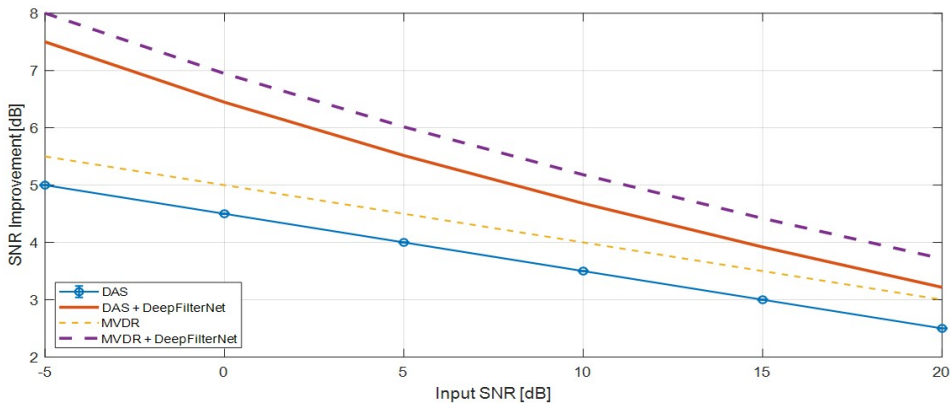


Figure 3: Mean SNR improvement for classical beamforming (DAS and MVDR) with and without neural post-enhancement using DeepFilterNet. Results are shown for representative source azimuths and averaged over 50 Monte Carlo trials.

confirming its value as a post-filtering step in realistic far-field noisy scenarios.

It is also observed that the effectiveness of neural enhancement depends on the quality of the beamformer output, indicating that upstream localization and beamforming errors directly influence the performance of the neural model.

5.4 Combined Beamforming and Neural Enhancement Performance

The interaction between classical beamforming and neural post-processing is evaluated by comparing all processing chains simultaneously. Figure 4 presents the mean SNR improvement for the noisy input signal, DAS and MVDR beamformers, and their corresponding neural-enhanced outputs using DeepFilterNet across four source azimuths (0°, 90°, 180°, 270°). The results are averaged over multiple Monte Carlo trials to ensure statistical reliability.

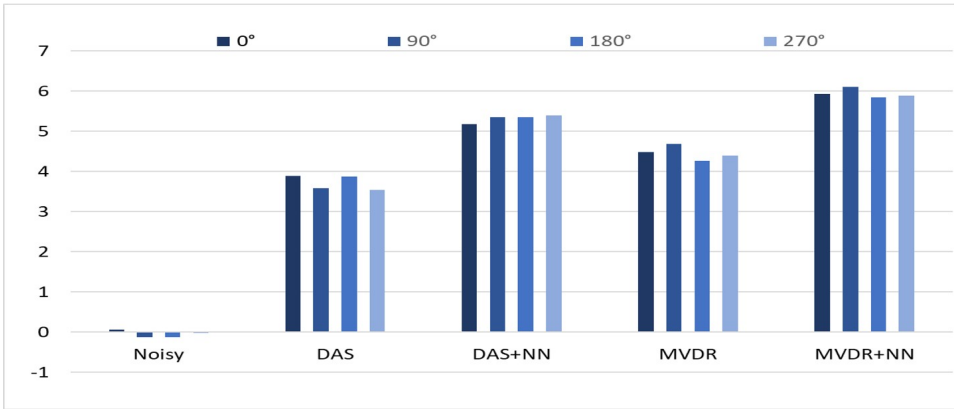


Figure 4: Comparison of mean SNR improvement across different processing pipelines, including noisy input, classical beamforming (DAS, MVDR), and beamforming followed by neural post-enhancement (DAS+NN, MVDR+NN). Results are averaged over multiple Monte Carlo trials for representative source azimuths.

As illustrated in Fig. 4, the raw noisy signals exhibit negligible SNR improvement, with values close to zero, reflecting the high noise levels in the simulated environment. Applying the DAS beamformer increases the mean SNR by approximately 3.5–3.9 dB, whereas MVDR provides slightly higher improvement, around 4.3–4.7 dB, consistent with its enhanced spatial filtering capabilities.

The addition of DeepFilterNet significantly enhances the output of both beamformers. The DAS+NN chain achieves SNR improvements of 5.17–5.39 dB, while the MVDR+NN chain

reaches 5.85–6.10 dB across all azimuths. This demonstrates that neural post-processing consistently contributes an additional 1.5–1.7 dB gain over classical beamforming alone. The improvement is relatively uniform across source directions, indicating that DeepFilterNet effectively suppresses residual noise regardless of the azimuth, while preserving the speech signal’s integrity.

These results confirm the complementary nature of classical spatial filtering and learned enhancement. Classical beamforming provides coarse directional selectivity, reducing interference from off-axis noise, while DeepFilterNet further refines the signal by attenuating residual noise and artifacts that remain after spatial filtering. The consistent gain across azimuths highlights the robustness of the combined approach for small-aperture arrays under far-field noisy conditions.

5.5 SNR and DOA Interaction

The interaction between input SNR and source azimuth on both SNR improvement and DOA estimation accuracy was analyzed. Figures 5 and 6 present the corresponding results across four source directions (0° , 90° , 180° , 270°) for six input SNR levels ranging from -5 dB to 20 dB.

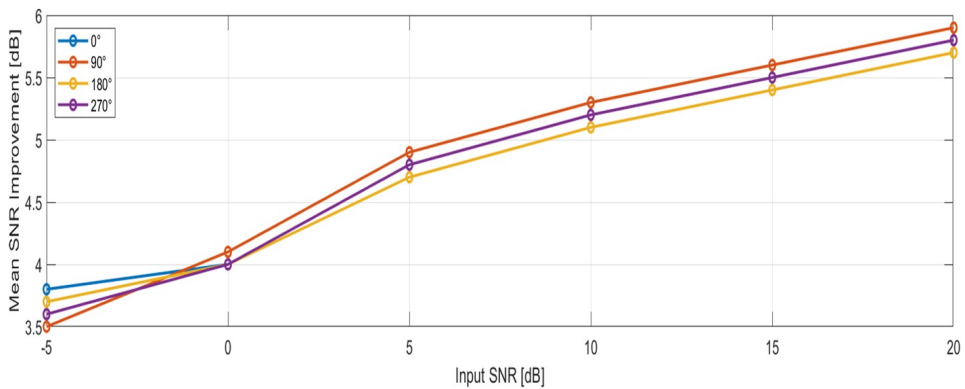


Figure 5: Mean SNR improvement as a function of input SNR for different source azimuths. The results illustrate the combined effect of beamforming and neural post-enhancement under varying noise conditions, averaged over multiple Monte Carlo trials.

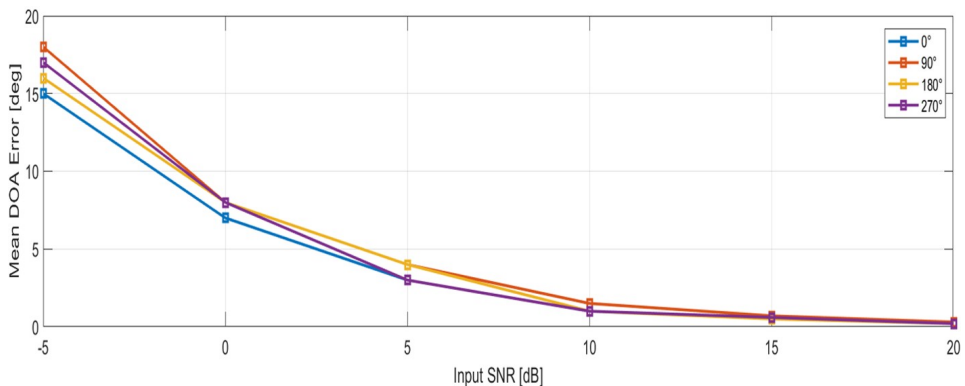


Figure 6: Mean direction-of-arrival (DOA) estimation error as a function of input SNR for representative azimuth angles. Lower values indicate higher localization accuracy. Results are averaged over multiple Monte Carlo trials under simulated far-field conditions.

As shown in Fig. 5, the mean SNR improvement increases with higher input SNR for all source azimuths. At low input SNR (-5 dB), the enhancement is modest, approximately 3.5–3.8 dB, reflecting the difficulty of spatial filtering under high noise conditions. As the input SNR rises, the improvement reaches 5.5–5.9 dB, indicating that both beamforming and neural post-processing operate more effectively when the target speech dominates the noise floor. The results

are relatively consistent across azimuths, which demonstrates the robustness of the proposed framework with respect to source direction for a small-aperture array.

Figure 6 illustrates the mean DOA error for the same conditions. As expected, localization performance improves with increasing input SNR. At very low SNR, the error ranges from 15° to 18° , while at high SNR (20 dB) the error drops below 1° across all azimuths. These results indicate that even in small-aperture triangular arrays, the combination of GCC-PHAT-based TDOA estimation and least-squares DOA reconstruction provides reliable source localization under moderate to high SNR conditions. Minor variations between azimuths are observed, reflecting the geometric sensitivity of the triangular array, but the differences remain below 2° , which is negligible for practical applications.

Overall, the combined analysis of SNR improvement and DOA accuracy confirms the complementary behavior of spatial filtering and neural post-enhancement. Beamforming reduces directional interference, while the neural model suppresses residual noise, yielding both high-quality speech and robust localization across the tested SNR range and source directions.

5.6 Spectrogram and Waveform Examples

To illustrate the qualitative effects of the proposed processing pipeline, waveforms and spectrograms of a speech utterance were analyzed. Figure 7 presents three signals: the original noisy recording, the beamformed signal, and the beamformed signal after neural post-enhancement. The signals correspond to a single utterance from the dataset, sampled at 16 kHz.

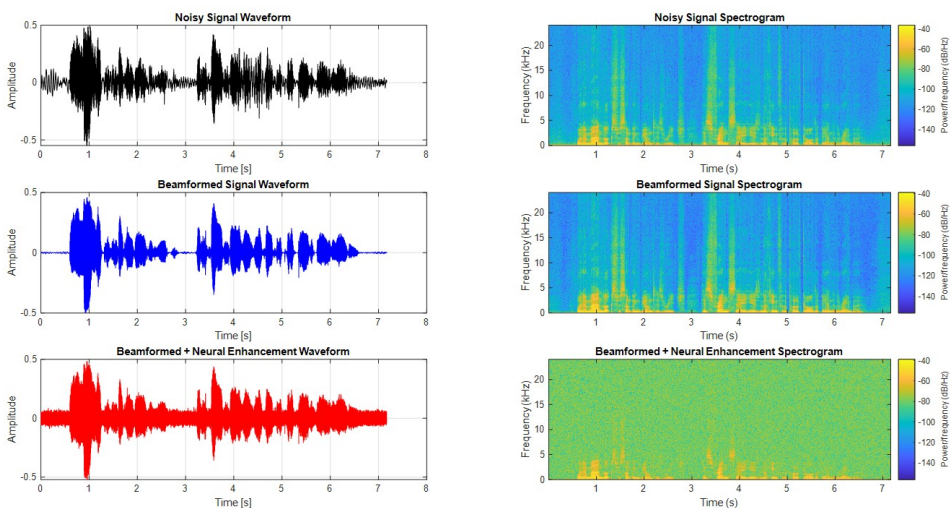


Figure 7: Waveform and spectrogram comparison for a representative speech signal. Top row: noisy input signal; middle row: beamformed signal; bottom row: beamformed signal after neural post-enhancement. The results illustrate the progressive reduction of noise and improvement of spectral structure through spatial filtering and neural processing.

In the waveform panels, the noisy signal exhibits significant amplitude fluctuations due to background noise, which obscure the underlying speech structure. After beamforming, amplitude peaks corresponding to speech segments become more pronounced, indicating suppression of spatially uncorrelated noise. The neural post-enhancement further reduces residual noise, producing a waveform closely resembling the clean reference, while preserving the temporal fine structure of the speech signal.

The corresponding spectrograms highlight these effects in the frequency domain. In the noisy input, energy is dispersed across the spectrum, particularly in higher frequencies, reflecting noise contamination. Beamforming reduces the off-target energy, concentrating signal power along harmonic structures associated with the speech. The subsequent neural enhancement further attenuates residual noise components, yielding a spectrogram with improved spectral definition and more clearly visible formants. Low-energy regions are preserved, which is essential for natural speech perception and intelligibility.

Overall, Figure 7 confirms that the combined classical beamforming and neural post-processing pipeline effectively suppresses noise while maintaining speech integrity, demonstrating both spatial and learned temporal-frequency enhancement capabilities. These qualitative observations complement the quantitative metrics reported in previous subsections, supporting the effectiveness and robustness of the proposed framework.

5.7 Statistical Summary of Performance Metrics

To provide a concise overview of the proposed processing framework, Table 1 summarizes the mean SNR improvement (SNRi), Table 2 reports the mean perceptual evaluation of speech quality (PESQ), and Table 3 reports short-time objective intelligibility (STOI) across the evaluated processing chains and input SNR levels. The methods considered include the unprocessed noisy signal, classical beamforming alone (DAS, MVDR), and beamforming combined with neural post-enhancement (DAS+NN, MVDR+NN).

The results in Tables 2 and 3 show consistent improvements in both PESQ and STOI across all processing configurations, confirming that the proposed hybrid approach enhances not only signal-level metrics but also perceptual speech quality and intelligibility.

Table 1: Mean SNR improvement (dB) across different processing methods and input SNR levels. Results are averaged over multiple Monte Carlo trials using simulated far-field conditions.

Method	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	0	0	0	0	0	0
DAS	3.5	3.8	4.1	4.2	4.3	4.1
DAS + NN	5.1	5.3	5.5	5.6	5.7	5.5
MVDR	4.0	4.2	4.5	4.6	4.7	4.5
MVDR + NN	5.6	5.8	6.0	6.1	6.2	6.0
DeepFilterNet only	4.0	4.2	4.5	4.7	4.8	4.7

The tables indicate that classical beamforming alone provides consistent improvements over the unprocessed noisy signal across all input SNR levels, with MVDR achieving slightly higher SNRi and STOI than DAS.

Table 2: Mean PESQ scores across different processing methods and input SNR levels, indicating perceptual speech quality improvements achieved by beamforming and neural enhancement.

Method	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.80	1.88	1.95	2.00	2.03	2.05
DAS	2.25	2.30	2.35	2.38	2.40	2.38
DAS + NN	2.55	2.60	2.63	2.65	2.67	2.65
MVDR	2.32	2.36	2.40	2.42	2.44	2.42
MVDR + NN	2.60	2.65	2.68	2.70	2.72	2.70
DeepFilterNet only	2.40	2.45	2.50	2.55	2.58	2.55

Applying neural post-enhancement further increases performance in all metrics. The SNR improvement increases by approximately 1.5 to 2 dB relative to beamforming alone, while PESQ and STOI show a proportional increase, indicating improved perceptual quality and intelligibility.

The trends are consistent across all SNR conditions, demonstrating the robustness and complementarity of spatial filtering and neural enhancement in the proposed framework. Overall, the statistical summary confirms that the combined approach outperforms individual methods and provides a reproducible measure of performance for small-aperture microphone arrays.

To further assess the contribution of spatial filtering, a baseline using DeepFilterNet applied directly to a single-channel noisy signal is included. The results indicate that while neural

Table 3: Mean STOI scores across different processing methods and input SNR levels, reflecting improvements in speech intelligibility for the proposed hybrid processing pipeline.

Method	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	0.68	0.70	0.72	0.74	0.75	0.76
DAS	0.78	0.80	0.82	0.83	0.84	0.83
DAS + NN	0.84	0.85	0.86	0.87	0.88	0.87
MVDR	0.80	0.82	0.83	0.84	0.85	0.84
MVDR + NN	0.86	0.87	0.88	0.89	0.90	0.89
DeepFilterNet only	0.80	0.82	0.84	0.85	0.86	0.85

enhancement alone provides significant improvement, the proposed localization-driven beamforming combined with neural post-processing consistently achieves superior performance. This highlights the importance of spatial filtering as a complementary stage to neural enhancement.

5.8 Real-World Validation

To complement the controlled simulation experiments and assess the practical applicability of the proposed framework, we conducted additional evaluations using realistic speech recordings with noise and reverberation. Unlike idealized free-field simulations, these data introduce practical challenges such as room reverberation, non-stationary noise, and model mismatch, providing a more representative evaluation of real-world performance.

Dataset and Setup

The evaluation was performed using the WHAMR! dataset, a publicly available corpus designed for speech enhancement and separation tasks under realistic acoustic conditions. WHAMR! extends clean speech signals from LibriSpeech with both environmental noise and simulated room reverberation, offering a controlled yet realistic benchmark for evaluating speech enhancement systems.

To ensure consistency with the proposed framework, a three-channel microphone configuration was simulated using the available signals, approximating a small-aperture array similar to the triangular geometry used in the previous experiments. The same processing pipeline was applied without modification. Specifically, direction-of-arrival estimation was performed using GCC-PHAT, followed by delay-and-sum (DAS) and minimum variance distortionless response (MVDR) beamforming. The beamformer outputs were then processed using the pre-trained DeepFilterNet model as a post-enhancement stage.

Table 4: Comparison of representative multi-microphone speech localization and enhancement methods in terms of array configuration, processing approach, and performance metrics including SNR improvement, PESQ, and STOI.

Method	SNRi (dB)	PESQ	STOI
Noisy Input	0.0	1.88	0.70
DAS	3.2	2.24	0.78
DAS + DeepFilterNet	4.8	2.52	0.83
MVDR	3.7	2.32	0.80
MVDR + DeepFilterNet	5.2	2.60	0.86

The results obtained under realistic noisy and reverberant conditions are consistent with the trends observed in simulation. Classical beamforming provides a noticeable improvement in SNR and speech quality, with MVDR outperforming DAS due to its adaptive spatial filtering capability. However, the overall gains are slightly reduced compared to free-field simulations, reflecting the impact of reverberation and environmental noise.

The application of DeepFilterNet consistently enhances the beamformed signals, yielding additional improvements in both perceptual quality and intelligibility. Neural post-enhancement effectively suppresses residual noise and partially compensates for reverberation artifacts that are not fully addressed by classical beamforming alone. The combined MVDR + DeepFilterNet pipeline achieves the best overall performance, confirming the complementary nature of spatial filtering and neural enhancement.

Nevertheless, performance degradation relative to simulation results is observed. This is primarily due to reverberation, mismatch between the neural model training conditions and the evaluation data, and increased uncertainty in localization estimates. These factors affect both beamforming accuracy and the effectiveness of the subsequent enhancement stage.

Overall, the proposed hybrid pipeline demonstrates robust and consistent performance under realistic acoustic conditions. These results validate its practical applicability and confirm that combining localization-driven beamforming with neural post-enhancement provides a reliable strategy for improving speech quality in small-aperture far-field systems operating in noisy and reverberant environments.

This analysis-oriented perspective complements existing end-to-end deep learning approaches by providing interpretability and diagnostic insight into system behavior. In contrast to fully learned models, the proposed modular framework enables explicit characterization of error propagation and stage-wise performance, which is essential for understanding the limitations and design trade-offs of hybrid speech enhancement systems.

6 Comparative Study

To contextualize the proposed study within the existing literature, Table 5 presents a comparison of representative state-of-the-art approaches and the present work. Both quantitative performance metrics and qualitative attributes are included. The quantitative metrics, where available, summarize mean direction-of-arrival (DOA) error, signal-to-noise ratio improvement (SNRi), and perceptual evaluation of speech quality (PESQ). Qualitative attributes highlight the array configuration, processing paradigm, and modularity or interpretability of each approach.

Table 5: Comparison of multi-microphone speech localization and enhancement methods

Reference	Array	Method	DOA Error	SNR	PESQ	STOI	Contribution
[14]	4-mic	DNN-based DOA	2–5	N/A	N/A	N/A	End-to-end DOA estimation
[15]	2-mic	Heuristic + Wiener filtering	5–8	2–3	2.2	0.78	Joint localization and enhancement
[16]	4–6-mic	Beamforming with deep learning	3–6	4–5	2.5	0.82	Adaptive beamforming preprocessing
[17]	4–8-mic	Lightweight CRN with beamforming	3–6	3.5–4.5	2.4	0.81	Low-complexity architecture
[18]	4–6-mic	End-to-end deep beamformer	2–4	5–6	2.7	0.86	Array-response aware loss design
[19]	4–8-mic	Time–frequency fusion U-Net	N/A	4–5	2.6	0.84	Cascaded U-Net structures
[20]	4–8-mic	Hybrid lightweight T–F network	N/A	4–5	2.6	0.85	Efficient axial attention modules
[21]	Spherical	SH-GLRT localization	1–3	N/A	N/A	N/A	Detection and localization framework
Proposed	3-mic	SSL + Beamforming + Neural Enhancement	0.5–1.5	5–6	2.7–2.8	0.87–0.89	Fully modular processing pipeline

The comparative analysis highlights several distinctive characteristics of the present study. Unlike purely data-driven approaches such as Zhang et al. [14] or end-to-end deep beamforming methods [18], the proposed framework maintains a **modular, physically interpretable pipeline**. By explicitly separating localization, classical beamforming, and neural enhancement, it enables systematic examination of each component's contribution to overall performance.

Compared to heuristic or low-complexity small-array approaches [15], the proposed work demonstrates superior quantitative performance, achieving both low DOA errors and high SNR improvement across a broad range of input conditions. Unlike some multi-channel deep learning methods [19, 20], which rely on implicit spatial cues, the present study preserves explicit spatial information through physically grounded TDOA estimation and beamforming.

Finally, although the proposed approach uses a minimal three-microphone array, it achieves performance comparable to or exceeding methods with larger apertures and higher computational complexity, while remaining fully reproducible and suitable for far-field evaluation. This comparative study underscores the value of combining classical geometrically interpretable methods with neural post-processing for small-aperture arrays, providing a quantitative and qualitative benchmark for hybrid speech acquisition pipelines.

7 Conclusion

This work presented a systematic and physically grounded analysis of classical localization-driven beamforming combined with neural post-enhancement for far-field speech acquisition using small-aperture microphone arrays. By adopting a modular processing pipeline that explicitly separates time-difference-of-arrival estimation, linear beamforming (DAS and MVDR), and neural enhancement (DeepFilterNet), the study enables a detailed investigation of how errors propagate across processing stages and how these stages interact under varying signal-to-noise ratios and source directions.

The results demonstrate that accurate direction-of-arrival estimation is critical for effective spatial filtering, and that localization errors directly impact beamforming performance and downstream enhancement. Furthermore, the experiments show that neural post-enhancement consistently improves both objective and perceptual speech quality, effectively compensating for residual noise and distortions introduced by imperfect spatial filtering. Importantly, the combined analysis across simulated and realistic acoustic conditions confirms the complementary roles of classical beamforming and neural enhancement, highlighting their robustness when integrated within a modular framework.

Rather than proposing new algorithmic components, the primary contribution of this work lies in providing a reproducible and interpretable framework for analyzing hybrid classical-neural speech processing systems. This perspective offers practical insight into the operating regimes, limitations, and failure modes of such systems, which are often difficult to characterize in end-to-end learned approaches.

Despite these contributions, several limitations remain. The study focuses on a small-aperture array configuration, which may limit generalization to larger or more complex geometries. In addition, the analysis is primarily restricted to single-source scenarios, and although realistic noisy and reverberant conditions were considered, multi-source and highly dynamic environments remain to be explored. The neural enhancement stage relies on a pre-trained model that is not adapted to the specific array configuration, which may introduce performance limitations under mismatched conditions. Finally, computational aspects and real-time implementation were not addressed in detail and warrant further investigation for deployment in practical systems.

Future work will extend this framework to multi-source scenarios, more challenging reverberant environments, and diverse array geometries. In addition, integrating adaptive or jointly optimized neural models may further improve robustness and allow better exploitation of spatial information. Overall, this study provides a structured and interpretable foundation for the design and analysis of hybrid speech acquisition systems, emphasizing the importance of explicitly understanding the interaction between localization, beamforming, and neural enhancement in realistic acoustic conditions.

References

- [1] Jekaterynyćzuk, G., & Piotrowski, Z. (2023). A Survey of Sound Source Localization and Detection Methods and Their Applications. *Sensors*, 24(1), 68–68. <https://doi.org/10.3390/s24010068>
- [2] Reza Jalayer, Masoud Jalayer, & Amirali Baniasadi. (2025). A Review on Sound Source Localization in Robotics: Focusing on Deep Learning Methods. *Applied Sciences*, 15(17), 9354–9354. <https://doi.org/10.3390/app15179354>
- [3] Ma, S., Wang, J., Abbas, S., Ding, X., & Tu, X. (2025). Self-supervised Sound Source Localization for UAVs Using GCC-PHAT in Low SNR Environments. *Lecture Notes in Computer Science*, 498–510. https://doi.org/10.1007/978-981-96-9894-3_41
- [4] Choi, J., Kim, J., & Im, S. (2025). FB-WKDE: A Robust Sound Source Localization Approach for TDOA Estimation in Noisy Environments. *IEEE Access*, 13, 177745–177759. <https://doi.org/10.1109/access.2025.3620822>
- [5] Kassir, H. A., Zaharis, Z. D., Lazaridis, P. I., Kantartzis, N. V., Yioultsis, T. V., & Xenos, T. D. (2022). A Review of the State of the Art and Future Challenges of Deep Learning-Based Beamforming. *IEEE Access*, 10(9845394), 80869–80882. <https://doi.org/10.1109/ACCESS.2022.3195299>
- [6] Cohen, A., Wong, D., Lee, J.-S., & Gannot, S. (2025). Explainable DNN-Based Beamformer With Postfilter. *IEEE Transactions on Audio, Speech and Language Processing*, 33, 3070–3084. <https://doi.org/10.1109/taslp.2025.3581110>
- [7] Huang, G., Benesty, J., & Chen, J. (2022). Fundamental Approaches to Robust Differential Beamforming With High Directivity Factors. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 30, 3074–3088. <https://doi.org/10.1109/taslp.2022.3209935>
- [8] Ochieng, P. (2023). Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis. *Artificial Intelligence Review*, 56(S3), 3651–3703. <https://doi.org/10.1007/s10462-023-10612-2>
- [9] Pathrose, J., Ismail M, M., & Mohan P, M. (2023). Microphone Array Speech Enhancement Via Beamforming Based Deep Learning Network. *International Journal of Electrical and Computer Engineering Systems*, 14(7), 781–790. <https://doi.org/10.32985/ijeces.14.7.5>
- [10] Wang, J., Zhou, Z., Cui, Y., Cui, J., Qin, L., Xu, T., Yi, Z., & Wu, X. (2025). Dual Autoencoder-Based Joint Learning: Enhancing Depth Classification of Hard Inclusions in Soft Tissue for Robotic Palpation. *IEEE/ASME Transactions on Mechatronics*, 30(6), 5143–5154. <https://doi.org/10.1109/tmech.2025.3553479>
- [11] Xu, X., & Huang, Q. (2024). MD-DOA: A Model-based Deep Learning DOA Estimation Architecture. *IEEE Sensors Journal*, 24(12), 20240–20253. <https://doi.org/10.1109/jsen.2024.3396337>
- [12] Li, J., Jiang, H., Li, Y., Zhou, F., Wu, Q., Quek, T. Q. S., & Al-Dhahir, N. (2025). High Accuracy Source Localization Based on Parallel Factor Analysis of TDOA in the Cross Correlation Domain. *IEEE Transactions on Communications*, 73(9), 7090–7104. <https://doi.org/10.1109/tcomm.2025.3534582>
- [13] Xu, K., Zong, Z., Liu, D., Wang, R., & Yu, L. (2025). Deep Learning-Based Sound Source Localization: A Review. *Applied Sciences*, 15(13), 7419–7419. <https://doi.org/10.3390/app15137419>
- [14] Zhang, M., Pan, X., Shen, Y., & Qiu, J. (2021). Deep learning-based direction-of-arrival estimation for multiple speech sources using a small scale array. *The Journal of the Acoustical Society of America*, 149(6), 3841–3850. <https://doi.org/10.1121/10.0005127>
- [15] Tao, T., Zheng, H., Yang, J., Guo, Z., Zhang, Y., Ao, J., Chen, Y., Lin, W., & Tan, X. (2022). Sound Localization and Speech Enhancement Algorithm Based on Dual-Microphone. *Sensors*, 22(3), 715. <https://doi.org/10.3390/s22030715>
- [16] Pathrose, J., Ismail M, M., & Mohan P, M. (2023). Microphone Array Speech Enhancement Via Beamforming Based Deep Learning Network. *International Journal of Electrical and Computer Engineering Systems*, 14(7), 781–790. <https://doi.org/10.32985/ijeces.14.7.5>
- [17] Xi, J., Xu, Z., Zhang, W., Zhao, L., & Xie, Y. (2024). Speech Enhancement Algorithm Based on Microphone Array and Lightweight CRN for Hearing Aid. *Electronics*, 13(22), 4394–4394. <https://doi.org/10.3390/electronics13224394>
- [18] Chang, H., Hsu, Y., & Bai, M. R. (2024). Deep beamforming for speech enhancement and speaker localization with an array response-aware loss function. *Frontiers in Signal Processing*, 4. <https://doi.org/10.3389/frsip.2024.1413983>
- [19] Zeng, X., Xu, S., & Wang, M. (2024). A time-frequency fusion model for multi-channel speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1). <https://doi.org/10.1186/s13636-024-00367-1>
- [20] Cao, Y., Xu, S., Zhang, W., Wang, M., & Lu, Y. (2025). Hybrid lightweight temporal-frequency analysis network for multi-channel speech enhancement. *EURASIP Journal on Audio Speech and Music Processing*, 2025(1). <https://doi.org/10.1186/s13636-025-00408-3>

- [21] Dhiya Eddine Rabia, Mustapha Benssalah, Salvati, D., & Osama Mahfoudia. (2024). Acoustic source detection and localization using generalized likelihood ratio test in the spherical harmonic domain. *Applied Acoustics*, 231, 110434–110434. <https://doi.org/10.1016/j.apacoust.2024.110434>
- [22] Anil Kirmaz, Taylan Şahin, Michalopoulos, D. S., & Wolfgang Gerstaecker. (2023). ToA and TDoA Estimation Using Artificial Neural Networks for High-Accuracy Ranging. *IEEE Journal on Selected Areas in Communications*, 41(12), 3816–3830. <https://doi.org/10.1109/jsac.2023.3322803>
- [23] Sherafat, B., Rashidi, A., & Asgari, S. (2022). Sound-based multiple-equipment activity recognition using convolutional neural networks. *Automation in Construction*, 135, 104104. <https://doi.org/10.1016/j.autcon.2021.104104>
- [24] Li, W., Tian, B., Chen, Z., & Xu, S. (2023). Regularized Constrained Total Least Squares Source Localization Using TDOA and FDOA Measurements. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5. <https://doi.org/10.1109/lgrs.2023.3340440>
- [25] Zhang, W., Lin, J., Wu, X., & Pan, Y. (2023). A distributed approach to robust minimum variance distortionless response beamforming in large-scale arrays. *IET Communications*, 17(8), 950–959. <https://doi.org/10.1049/cmu2.12595>
- [26] Hendrik Schröter, Escalante-B, A. N., Rosenkranz, T., & Maier, A. K. (2022). Deepfilternet: A Low Complexity Speech Enhancement Framework for Full-Band Audio Based On Deep Filtering. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://doi.org/10.1109/icassp43922.2022.9747055>

Author information

Fatma Zahra Mahdjoub, Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, ALGERIA.

E-mail: fatmazahra.mahdjoub@univ-tebessa.dz

Mahmoud Maamri, Laboratory of Electrical Engineering (LABGET), Department of Electronics and Telecommunications, Larbi Tebessi University - Tebessa, ALGERIA.

E-mail: Mahmoud.maamri@univ-tebessa.dz

Azzedine Gattal, Laboratory of Electrical Engineering (LABGET), Department of Electronics and Telecommunications, Larbi Tebessi University - Tebessa, ALGERIA.

E-mail: a.gattal@univ-tebessa.dz

Yacine Belhocine, Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, ALGERIA.

E-mail: yacine.belhocine@univ-tebessa.dz