

## Article

# Multichannel Acoustic Beamforming for Speaker Localization and DOA-Based Tracking

Jose Antonio Lopez-Olvera <sup>1</sup>, Hector Manuel Perez-Meana <sup>1</sup>, Elizabeth Garcia-Rios <sup>2</sup>,  
Enrique Escamilla-Hernandez <sup>1,\*</sup>, Jose Manuel Carrichi-Chavez <sup>1</sup> and Jesus Betancourt-Martinez <sup>1</sup>

<sup>1</sup> Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), Sección de Estudios de Posgrado e Investigación (SEPI), Unidad Culhuacán, Instituto Politécnico Nacional (IPN), Mexico City 04440, Mexico; jlopezo1902@alumno.ipn.mx (J.A.L.-O.); hmperez@ipn.mx (H.M.P.-M.); jcarrichi2000@alumno.ipn.mx (J.M.C.-C.); jbetancourt1800@alumno.ipn.mx (J.B.-M.)

<sup>2</sup> Tecnológico Nacional de México (TecNM), Instituto Tecnológico Superior del Occidente del Estado de Hidalgo (ITSOEH), Mixquiahuala 42700, Hidalgo, Mexico; egarcia@itsoeh.edu.mx

\* Correspondence: eescamillah@ipn.mx

## Abstract

Real-time speaker localization and speech enhancement remain challenging for embedded acoustic systems operating in noisy and reverberant environments. This paper presents a multichannel beamforming framework based on Generalized Cross-Correlation with Phase Transform (GCC-PHAT) for Direction of Arrival (DOA) estimation and Delay-and-Sum (DAS) beamforming using a compact four-element circular MEMS microphone array. The proposed approach incorporates DOA smoothing and verification to improve localization stability before beamforming. Experimental evaluation in a controlled indoor environment demonstrated localization errors below 1°, while the beamforming stage achieved Signal-to-Noise Ratio (SNR) values predominantly between 25 dB and 40 dB, with peaks approaching 45 dB, and Root Mean Square Error (RMSE) values below 0.05 for most processing windows with an average processing time per frame of 3.473 ms and a memory consumption of 67.11 KB. Comparative results show that the proposed system provides competitive localization accuracy with a computationally simple processing pipeline, making it suitable for real-time embedded applications such as intelligent voice interfaces, videoconferencing systems, and service robotics.

**Keywords:** beamforming; DOA; GCC-PHAT; microphone arrays; speaker localization

## 1. Introduction

In recent years, human-machine interaction systems have evolved toward more natural environments, with voice becoming the primary communication channel. Applications such as virtual assistants, videoconferencing systems, social robotics, and accessibility technologies demand solutions capable of detecting, localizing, and tracking multiple speakers in real time, even in acoustically complex environments.

Multichannel beamforming is one of the most effective techniques for addressing this challenge, as it adaptively combines signals captured by a microphone array to enhance desired sources while attenuating interference and background noise.

The performance of beamforming relies heavily on the accurate estimation of the Direction of Arrival (DOA), which determines the spatial direction of incident sound sources. By combining DOA estimation with beamforming techniques, it is possible not only to improve signal intelligibility and signal-to-noise ratio, but also to provide



Academic Editor: Young-Ho Seo

Received: 5 June 2026

Revised: 22 July 2026

Accepted: 27 July 2026

Published: 1 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

continuous tracking of multiple speakers. This functionality supports the development of new applications in natural communication scenarios and intelligent assistance systems.

Traditional voice-interaction solutions often rely on single microphones, which are easier to deploy and maintain, but are limited in their ability to separate speakers from background noise and other interfering sources.

Despite these advances, several limitations remain. High-accuracy localization methods such as SRP-PHAT or subspace-based algorithms generally require increased computational resources, making their deployment on embedded systems difficult. Likewise, many studies focus on single-source localization or evaluate their systems only under controlled laboratory conditions [1–3].

To address this limitation, this paper proposes a multichannel beamforming system using the GCC-PHAT and Delay-and-Sum (DAS) algorithms for DOA estimation. The system is designed to enhance the signal intelligibility, reduce interference and enable the dynamic tracking of multiple speakers. Evaluation results show that the proposed system performs better than other systems previously reported in the literature.

## 2. Proposed System

Figure 1 illustrates the main workflow of the proposed methodology. First, an iteration is performed between each pair of microphones using the GCC-PHAT algorithm which allows us to select the highest energy pair. Next, using the angle obtained by the GCC-PHAT algorithm, the DAS algorithm is applied where each microphone gets its corresponding delay in order to get the beamformed signal.

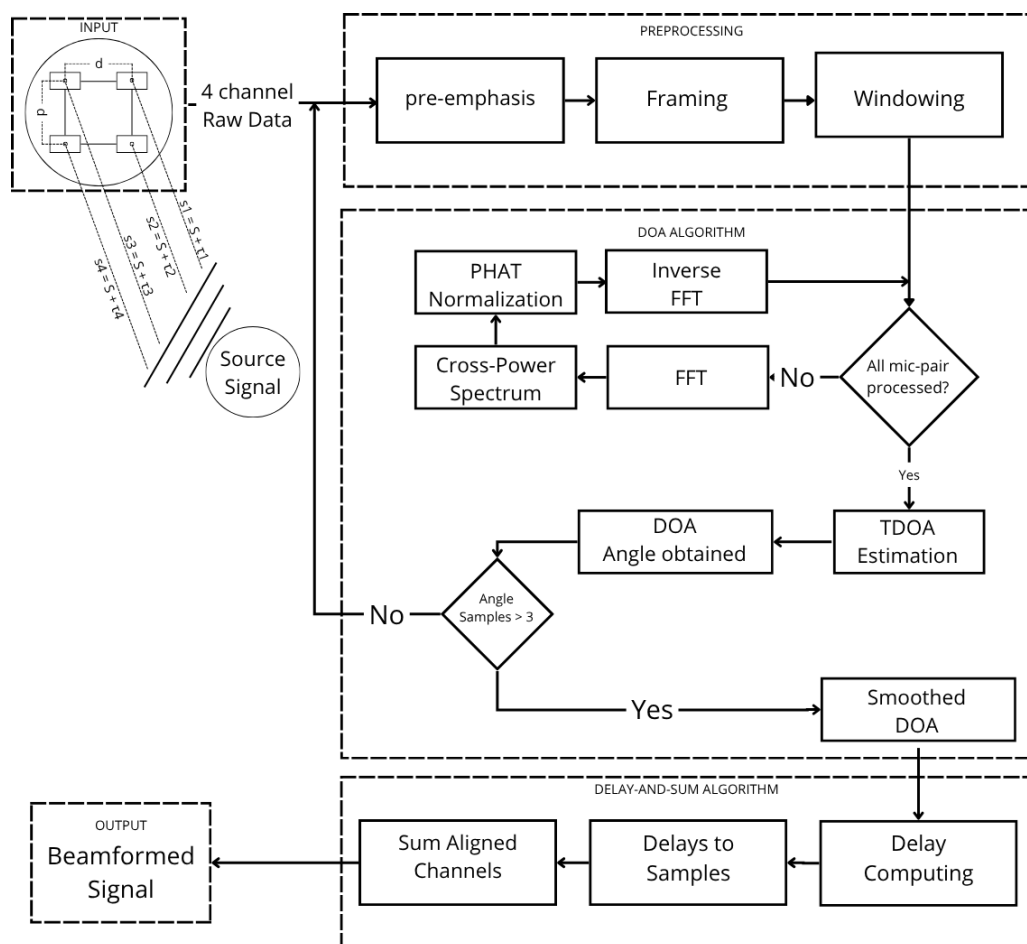


Figure 1. Flow chart of the proposed methodology.

The preprocessing stage is applied independently to each channel from the microphone array before performing the DOA estimation module. The GCC algorithm works with two input signals; therefore, six different mic-pairs were selected from the microphone array to perform the GCC-PHAT algorithm.

Once all microphone pairs were processed, the algorithm generates a  $1 \times 6$  vector containing the energy obtained from each pair. The pair that exhibits the highest energy value was selected as the most reliable DOA estimation. This method prioritizes the strongest spatial correlation, thereby improving the robustness under real-time operating conditions.

Once the DOA estimation is smoothed, the beamforming performance is significantly improved due to the reduction of abrupt angular variations between consecutive estimations. This smoothing process minimizes step-like changes in the estimated direction, allowing the beamformer to maintain a more stable steering response toward the acoustic source. As a result, the spatial filtering process becomes more robust and consistent, improving source enhancement and reducing fluctuations in the beamforming output.

The proposed system is composed of three main modules: the microphone array acquisition stage, the direction-of-arrival estimation based on the GCC-PHAT algorithm, and the beamforming stage for spatial filtering and source enhancement. The following subsections describe the operation and characteristics of each module in detail.

### 2.1. Microphone Arrays

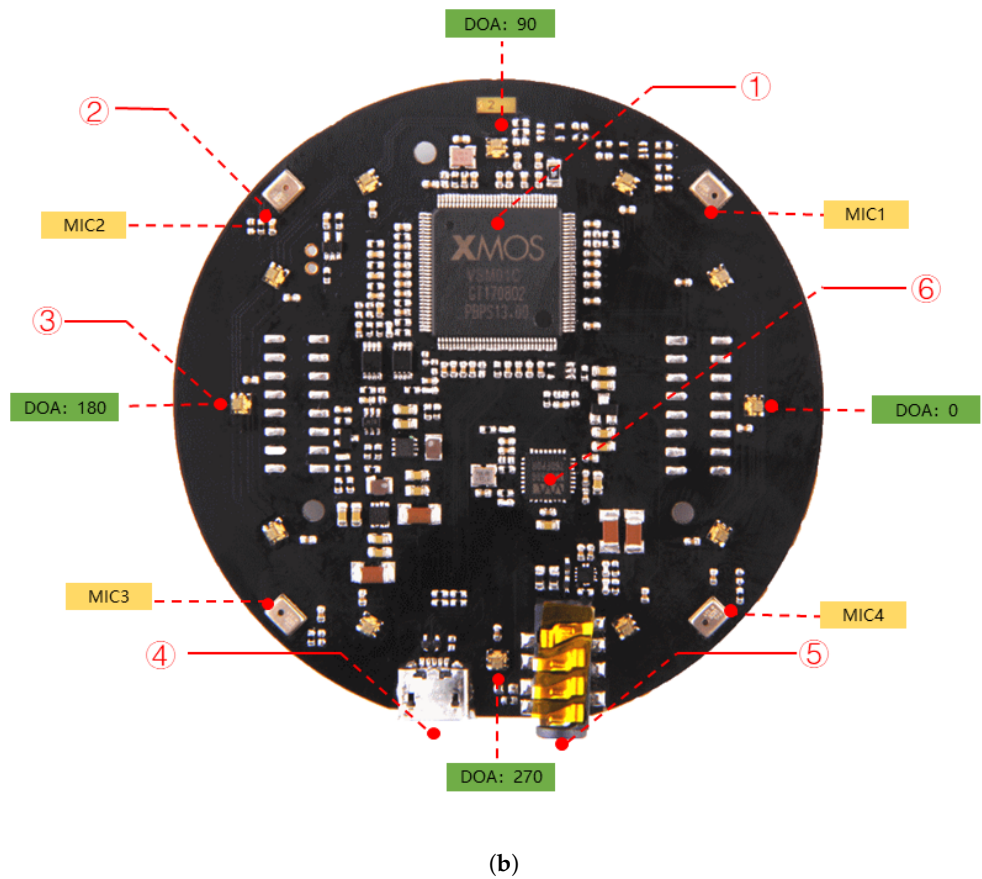
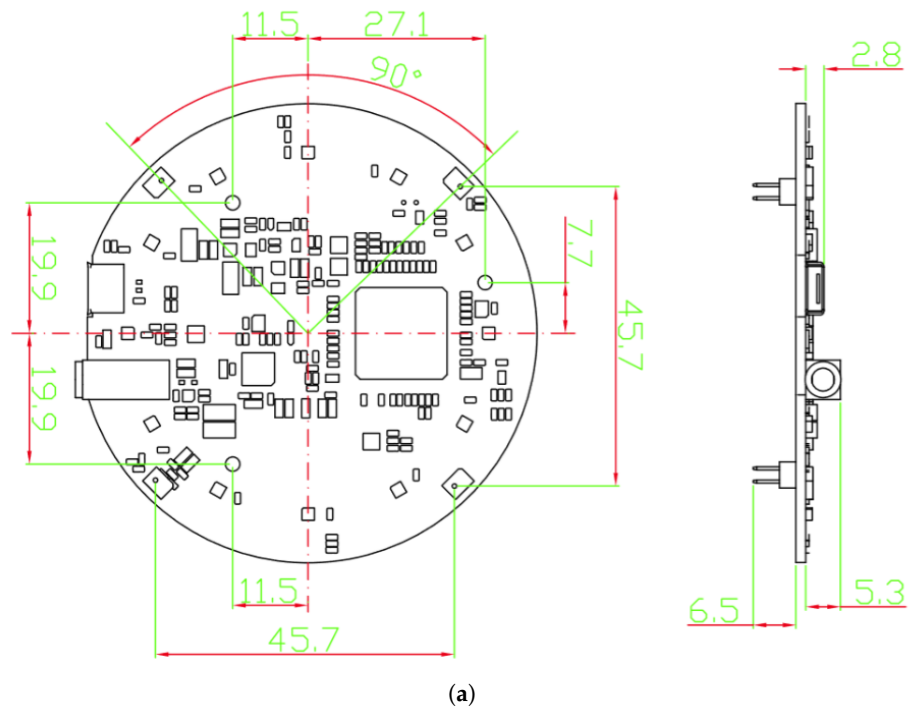
Microphone arrays can be configured using different geometries, such as linear, circular, and spherical arrangements, depending on the requirements of the target application. The proposed methodology is designed to operate with any even-numbered microphone array (e.g., 2, 4, 6, 8, or 10 microphones) with uniform inter-element spacing, making it scalable and adaptable to different hardware configurations. For the experimental validation presented in this work, the Sreed ReSpeaker Mic Array V2.0 [4] was selected as the evaluation platform. This device integrates four MP34DT01-M MEMS microphones with a signal-to-noise ratio (SNR) of 61 dB.

The microphones are arranged in a uniform circular array (UCA) with an inter-element spacing of 45.7 mm, as illustrated in Figure 2. This geometry provides 360° acoustic coverage, enabling sound acquisition from all directions and making it well suited for real-time direction-of-arrival (DOA) estimation and beamforming applications. Furthermore, the uniform spacing between adjacent microphones facilitates the accurate estimation of inter-microphone time differences (ITDs), which are exploited by the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) algorithm to estimate the direction of the incoming acoustic source.

In Figure 2, the numbered components correspond to the main hardware elements of the ReSpeaker microphone array. The function of each labeled element is described below:

1. **XMOS XVF-3000:** It integrates advanced digital signal processing (DSP) algorithms, including Acoustic Echo Cancellation (AEC), beamforming, dereverberation, noise suppression, and automatic gain control.
2. **Digital Microphones:** The MP34DT01-M is an ultra-compact, low-power, omnidirectional digital MEMS microphone that uses a capacitive sensing element combined with an integrated circuit interface for signal acquisition.
3. **RGB LED:** A three-color RGB LED used to indicate the operating status of the microphone array.
4. **USB Port:** Provides power supply to the microphone array and enables communication and control between the device and the host computer.
5. **3.5 mm Headphone Jack:** Provides an audio output interface where external active speakers or headphones can be connected.

6. **WM8960 Audio Codec:** A low-power stereo audio codec that integrates Class-D speaker drivers capable of delivering up to 1 W per channel into  $8\ \Omega$  loads.



**Figure 2.** Schematic of the ReSpeaker [4]. (a) Schematic overview. (b) Hardware distribution.

## 2.2. GCC-PHAT

### 2.2.1. GCC

The Generalized Cross-Correlation (GCC) [5,6] is a signal processing technique used to estimate the time delay between two signals received by a pair of microphones. These signals can mathematically be modeled as Equations (1a) and (1b).

$$x(t) = s_1(t) + n_1(t) \quad (1a)$$

$$y(t) = s_1(t + \tau) + n_2(t) \quad (1b)$$

where  $s_1$  is the captured signal for both microphones,  $n_1$  &  $n_2$  are the captured noise for each microphone and  $\tau$  is the delay between the first signal and the second signal which is assumed to be uncorrelated with  $n_1$  &  $n_2$ .

The cross-correlation function for Equations (1a) and (1b) is represented on Equation (2).

$$R_{12} = E[x(t)y(t)] = \int_{-\infty}^{\infty} x(t)y(t)dt \quad (2)$$

The cross correlation measures the projection between both signals, where the position of the maximum allows the estimation of  $\tau$  which gives the similarity instead of overlap. In order to get a proper measure of similarity it is necessary to apply a proper inner product which in complex functions requires the conjugated on one of the two signals. In order to get the time domain into frequency domain, it is necessary to use the convolution theorem, which states that a convolution in time becomes a multiplication in frequency and vice versa [7] as shown in Equations (3a) and (3b).

$$\mathcal{F}\{(f * g)(t)\} = F(\omega) \cdot G(\omega) \quad (3a)$$

$$\mathcal{F}^{-1}\{(fg) \cdot (t)\} = F(\omega) * G(\omega) \quad (3b)$$

Applying Equation (3a) to the cross-correlation definition in Equation (2) reveals that the cross-correlation is not itself a convolution. It becomes a convolution only when one of the signals is inverted in time, since time-reversal in the time domain corresponds to taking the complex conjugate of its Fourier transform as shown in Equation (4).

$$\mathcal{F}\{x(-t)\} = \int_{-\infty}^{\infty} x(-t)e^{-2j\pi ft}dt = \int_{-\infty}^{\infty} x(u)e^{+2j\pi ft}du = X^*(f) \quad (4)$$

Therefore, the cross-correlation can be expressed as a convolution with a time reversed signal applying the convolution theorem it yields Equation (5) [8] transforming it into the cross-power spectral density  $G_{xy}$ .

$$G_{xy}(f) = X(f) \cdot Y^*(f) \quad (5)$$

Then applying the Inverse Fourier Transform (IFT) to the cross spectral density function given by Equation (5), we get the cross power spectral density as shown in Equation (6). In the assumption that the observed signals are mutually independent from noise sources which, in real environments, this condition is rarely satisfied due to ambient noise, structural vibrations, sound propagation etc. Thus, it is necessary to use a reshape of Equation (6) before transforming it back to time domain by using a weighting function  $\psi(f)$ , which is applied to selectively amplify frequency bands where the signal-to-noise ratio (SNR) is high and attenuating those dominated by noise which lead to the Generalized Cross-Correlation function in Equation (7).

$$R_{xy}(\tau) = \mathcal{F}^{-1}\{G_{xy}(f)\} = \int_{-\infty}^{\infty} G_{xy}(f) e^{j2\pi f\tau} df \quad (6)$$

$$R_{xy}(\tau) = \mathcal{F}^{-1}\{G_{xy}(f)\} = \int_{-\infty}^{\infty} \psi(f) G_{xyL}(f) e^{j2\pi f\tau} df \quad (7)$$

### 2.2.2. PHAT

The Phase Transform is a weighting function used with GCC to improve the accuracy of time difference of arrival (TDOA) between the signals captured by a pair of microphones [9,10]. Unlike the GCC, this method focuses only on the phase information as shown in Equation (8), which carries the essential timing cues and sharpens the correlation peak making the estimation more robust in reverberant environments and resulting in the GCC-PHAT expression shown in Equation (9).

$$\psi(f) = \frac{1}{|X(f)Y^*(f)|} \quad (8)$$

$$GCC_{PHAT} = \int_{-\infty}^{\infty} \frac{G_{xy}(f)}{|G_{xy}(f)|} e^{j2\pi f\tau} df. \quad (9)$$

Although Equation (9) is valid for a single pair of microphones, as mentioned before, the microphone array that is being used has 4 microphones which give us 6 different combinations in order to get an angle. To resolve this problem Y. Zhao [6] and R. Wang [11] proposed that it is necessary to select the microphone pair with the highest peak of GCC-PHAT function.

This proposal can be done by making each microphone in the array represented by a coordinate vector, where each microphone location is defined by its spatial coordinates in the array reference system. The position vector of microphone  $i$  is defined as shown in Equation (10), where  $x_i$  and  $y_i$  represent the horizontal and vertical coordinates of microphone  $i$ , respectively.

$$\mathbf{r}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (10)$$

The direction of arrival of the acoustic wave can be represented by a unit vector that describes the propagation direction of the incoming signal. This vector depends on the unknown azimuth angle  $\theta$ , as defined in Equation (11).

$$\mathbf{u}(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \quad (11)$$

For a pair of microphones  $i$  and  $j$ , the theoretical Time Difference of Arrival (TDOA) is obtained from the projection of the distance vector between both microphones onto the direction of propagation. This relationship is expressed in Equation (12), where  $c$  represents the speed of sound.

$$\tau_{ij} = \frac{(\mathbf{r}_j - \mathbf{r}_i)^T \mathbf{u}(\theta)}{c} \quad (12)$$

By substituting the microphone position vector from Equation (10) and the direction vector from Equation (11), the TDOA relationship can be expanded into its scalar form. The resulting expression is presented in Equation (13).

$$\tau_{ij} = \frac{(x_j - x_i) \cos(\theta) + (y_j - y_i) \sin(\theta)}{c} \quad (13)$$

To simplify the formulation, the components of the direction vector are defined as two independent variables. The horizontal and vertical components are represented by  $u_x$  and  $u_y$ , respectively, as shown in Equations (14) and (15).

$$u_x = \cos(\theta) \quad (14)$$

$$u_y = \sin(\theta) \quad (15)$$

Using these definitions, Equation (13) can be rewritten as a linear relationship between the measured TDOA and the unknown direction vector components, resulting in Equation (16).

$$c\tau_{ij} = (x_j - x_i)u_x + (y_j - y_i)u_y \quad (16)$$

For the four-microphone array used in this work, six microphone pairs can be generated. Therefore, the TDOA measurements obtained from GCC-PHAT are grouped into the vector shown in Equation (17).

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_{12} \\ \tau_{13} \\ \tau_{14} \\ \tau_{23} \\ \tau_{24} \\ \tau_{34} \end{bmatrix} \quad (17)$$

By combining the six values obtained from Equation (16), the DOA estimation problem can be formulated as a matrix system. The resulting system is expressed in Equation (18).

$$\mathbf{A} \begin{bmatrix} u_x \\ u_y \end{bmatrix} = \mathbf{b} \quad (18)$$

Here, matrix  $\mathbf{A}$  contains the geometric information of the microphone array. This matrix is constructed from the spatial differences between microphone positions, as defined in Equation (19).

$$\mathbf{A} = \begin{bmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \\ x_4 - x_1 & y_4 - y_1 \\ x_3 - x_2 & y_3 - y_2 \\ x_4 - x_2 & y_4 - y_2 \\ x_4 - x_3 & y_4 - y_3 \end{bmatrix} \quad (19)$$

The vector  $\mathbf{b}$  contains the measured TDOA values obtained from the GCC-PHAT algorithm. By multiplying the measured delays by the speed of sound, the temporal differences are converted into equivalent propagation distances, as expressed in Equation (20).

$$\mathbf{b} = c \begin{bmatrix} \tau_{12} \\ \tau_{13} \\ \tau_{14} \\ \tau_{23} \\ \tau_{24} \\ \tau_{34} \end{bmatrix} \quad (20)$$

Since the system defined in Equation (18) contains six equations and only two unknown variables, it represents an overdetermined system. Therefore, the direction vector is estimated using the least-squares solution, which minimizes the error between the measured TDOAs and the values predicted by the array geometry. The least-squares estimation is given by Equation (21).

$$\hat{\mathbf{u}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (21)$$

The estimated direction vector obtained from Equation (21) is expressed as shown in Equation (22).

$$\hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_x \\ \hat{u}_y \end{bmatrix} \quad (22)$$

Finally, the DOA angle is calculated from the estimated direction vector using the two-argument arctangent function. This operation considers the signs of both vector components and provides the correct angular quadrant, as defined in Equation (23).

$$\theta_{DOA} = \text{atan2}(\hat{u}_y, \hat{u}_x) \quad (23)$$

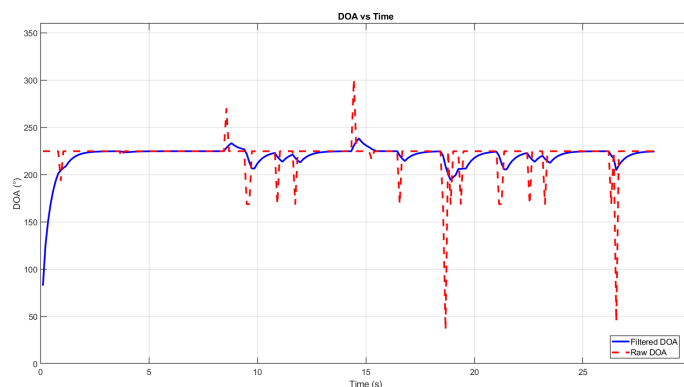
For representation in degrees, Equation (23) is converted into the degree-based form shown in Equation (24).

$$\theta_{DOA} = \text{atan2d}(\hat{u}_y, \hat{u}_x) \quad (24)$$

Since the output angle can be negative depending on the quadrant returned by the arctangent function, a normalization step is applied to restrict the DOA estimate to the range from  $0^\circ$  to  $360^\circ$ , as shown in Equation (25).

$$\theta_{DOA} = \begin{cases} \theta_{DOA} + 360^\circ, & \theta_{DOA} < 0, \\ \theta_{DOA}, & \theta_{DOA} \geq 0 \end{cases} \quad (25)$$

Because of the data acquisition characteristics, the estimated DOA is updated  $1/8$  s which can introduce noise if a transient source momentarily exhibits higher energy than the true source as can be shown in Figure 3 where the raw DOA (red dashed line) exhibits abrupt variations caused by noise and estimation uncertainty, whereas the filtered DOA (blue solid line) provides a smoother and more stable trajectory, improving the robustness of speaker localization.



**Figure 3.** Raw and smoothed direction-of-arrival (DOA) estimates as a function of time.

To mitigate this issue, it is necessary to compute an average between a certain number of acquired DOA. Although an arithmetic mean could be used, it is only valid when angles

are confined to the range  $0^\circ$  to  $180^\circ$ . In this work, the DOA spans the full range from  $0^\circ$  to  $360^\circ$ ; therefore, instead of using an arithmetic mean, a vector-based representation of the angles is employed, as shown in Equation (26).

In order to recover the captured angle, it is necessary to apply the  $\arctan^2$  function so it can show values from  $-\pi \leq \theta \leq \pi$ . The  $\arctan^2$  function can be defined by parts as shown in Equation (27).

$$\mathbf{v}_i = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \Rightarrow \bar{\mathbf{v}} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \quad (26)$$

$$\arctan^2 = \text{atan2}(y, x) = \begin{cases} \arctan\left(\frac{y}{x}\right) & \text{if } x > 0 \\ \arctan\left(\frac{y}{x}\right) + \pi & \text{if } x < 0, y \geq 0 \\ \arctan\left(\frac{y}{x}\right) - \pi & \text{if } x < 0, y < 0 \\ \frac{\pi}{2} & \text{if } x = 0, y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0, y < 0 \end{cases} \quad (27)$$

where

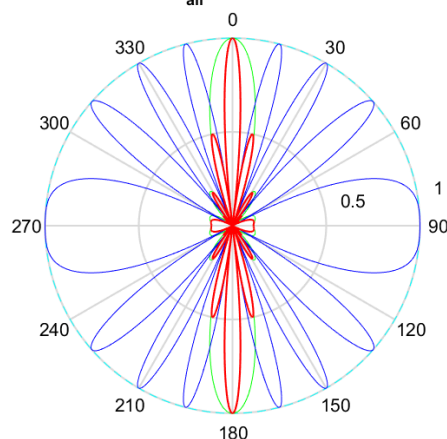
$$\bar{x} = \frac{1}{N} \sum_{i=1}^N \cos(\theta_i) \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N \sin(\theta_i) \quad (28)$$

### 2.3. Beamforming

Beamforming is a signal processing technique used in microphone or antenna arrays to spatially filter and enhance N-signals arriving from a specific direction by applying the appropriate delays weights to the individual sensor outputs. It constructs a spatial filter which can be analytically controlled, thereby improving the sound source localization accuracy.

The ideal implementation of a beamforming technique must address some robustness considerations like the environmental noise, interference from undesired sources, or sensor mismatches that can significantly degrade the beamforming performance [12,13]. To reproduce the specific beam pattern generated by [4], the Transducer Array Calculation (TAC) GUI [14] developed by B. Kohout was employed. This MATLAB (Version-R2024b)-based GUI offers an environment for defining and configuring transducer arrays e.g., the number of elements, spatial arrangement, etc. Figure 4 illustrates the resulting beam pattern generated by [4] where it demonstrated how mixing all the individual microphones can generate a focused main lobe along the vertical axis accompanied by smaller secondary lobes.

Directivity pattern  $C_{\text{all}}$  vertical (phi = azimuth =  $0^\circ$ )



**Figure 4.** Ideal beam pattern for [4]. Cyan line: Monopole; Green line: Rectangular Patch; Blue: XY-Array; Red: Total.

## 2.4. Delay-And-Sum

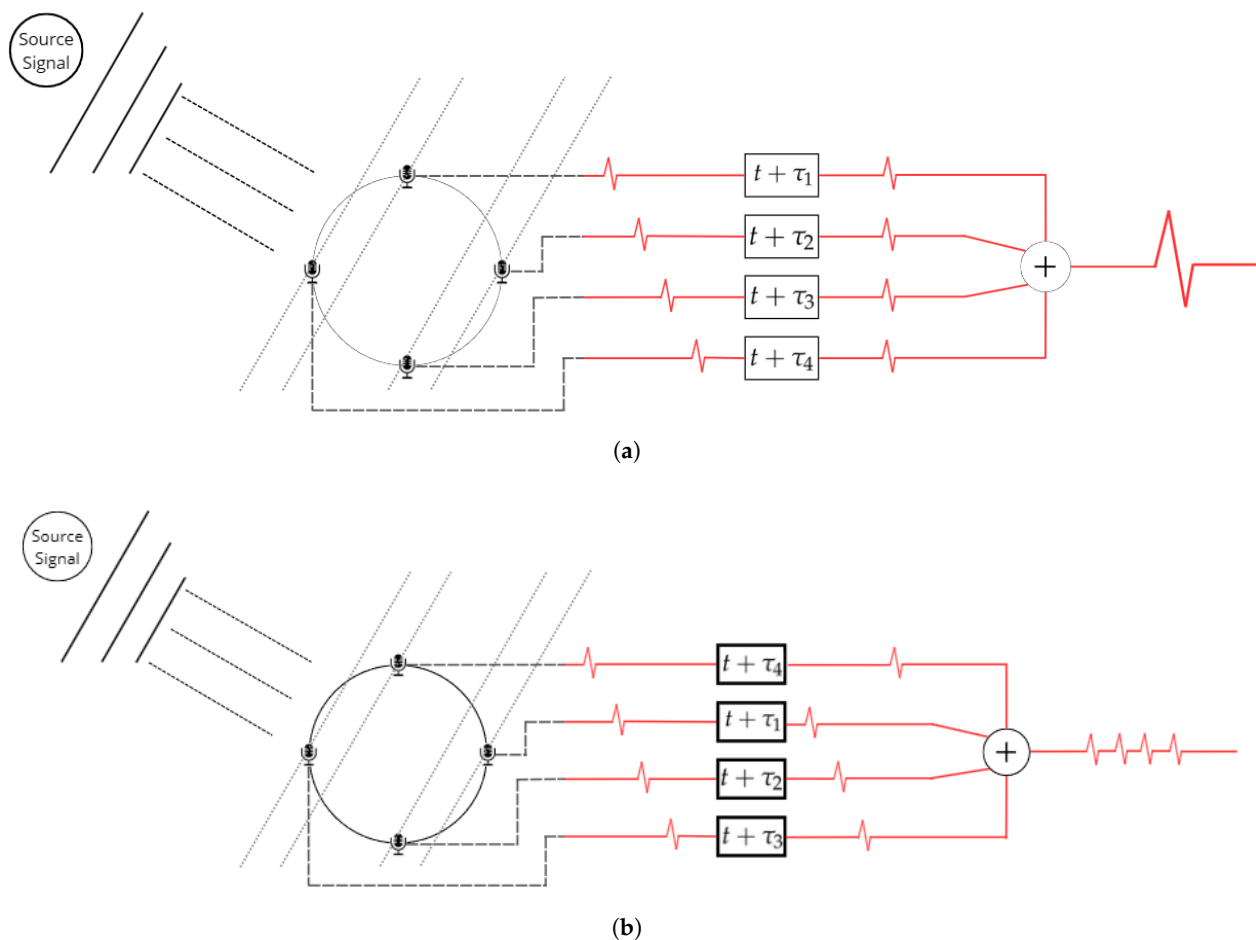
As described by Perrot et al. (2021) [15], the Delay-And-Sum (DAS) algorithm is one of the most classical beamforming techniques used in microphone array signal processing. Its purpose is to steer the array's sensitivity toward a desired direction, improving the SNR from the target source and discriminating the noise and interference from other directions.

Each microphone in the array receives the signal with a different time delay depending on the source position. The alignment of each signal can be computed as Equation (29) where  $\tau_i$  is the delay for the  $i$ -th microphone,  $p_i$  is the position of the  $i$ -th microphone,  $u$  is the unit vector pointing to the steering direction and  $c$  is the sound speed in the air ( $c \simeq 343$  m/s).

Once each microphone signal is time-shifted to compensate for its natural delay, they are summed together in order to reinforce the target Equation (30), if the signal comes from another direction, the signals do not align and are attenuated when summed as shown in Figure 5.

$$\tau_i = \frac{p_i \cdot u}{c} \quad (29)$$

$$y(t) = \sum_{i=1}^M x_i(t + \tau_i) \quad (30)$$



**Figure 5.** DAS algorithm time-shifting. (a) DAS working with accurate delays. (b) DAS working with non-accurate delays.

In the practical implementation of the system, the Delay-And-Sum (DAS) algorithm is used to improve the quality of the signal captured by the microphone array and emphasize

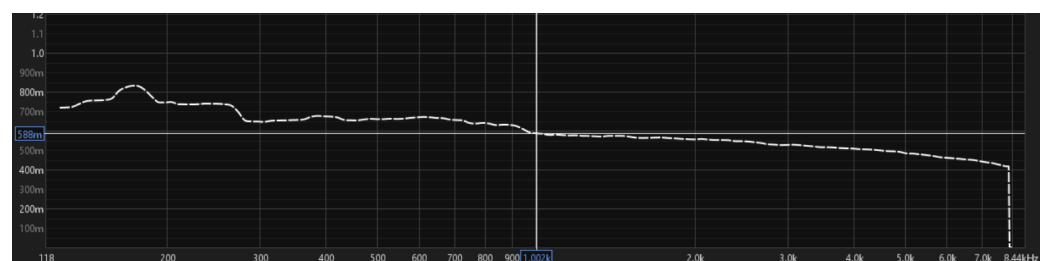
the detected speaker's voice. To achieve this, once the microphone signals have been preprocessed and the direction of arrival (DOA) has been estimated using the GCC-PHAT algorithm, the corresponding delays for each channel are calculated based on the estimated angle of the sound source. Subsequently, each signal is time-shifted to compensate for the differences in arrival time between the microphones. After this alignment stage, the corrected signals are summed sample by sample to generate a single output signal. Because the components from the estimated direction are synchronized, they are constructively reinforced during summation, while signals from other directions remain misaligned and are partially attenuated. As a result, a beamformed signal is obtained with a higher signal-to-noise ratio and a better representation of the target voice, facilitating subsequent speech analysis and processing stages.

### 3. Results

To evaluate the performance of the proposed system under different acoustic conditions, a comprehensive characterization of the test environments was conducted prior to the experimental evaluation. Since the acoustic properties of an enclosure directly affect both speech acquisition and signal processing, two key room parameters were measured for each environment: the reverberation time (RT60) [16] and the background noise level, expressed according to the Noise Criteria (NC) curves [17]. These parameters provide a quantitative description of the acoustic conditions under which the direction-of-arrival estimation and beamforming algorithms operate and allow the performance of the system to be analyzed under controlled and repeatable conditions.

The experimental evaluation was conducted inside a room with dimensions of  $8.53 \text{ m} \times 5.14 \text{ m} \times 2.65 \text{ m}$  (length  $\times$  width  $\times$  height), corresponding to a total volume of  $116.18 \text{ m}^3$ . The size of the enclosure provides an indoor acoustic environment for evaluating the proposed system under reproducible conditions.

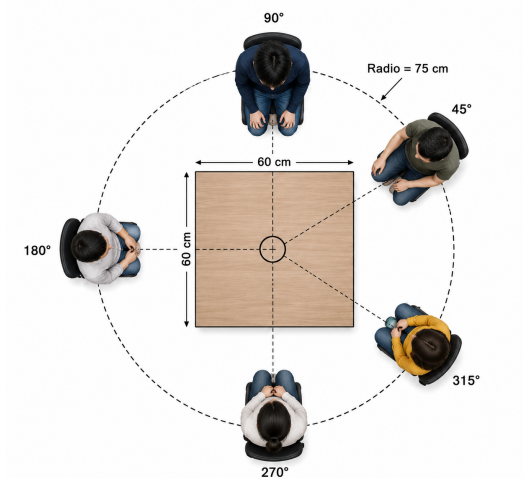
The reverberation time (RT60), defined as the time required for the sound pressure level to decay by 60 dB after the sound source is interrupted, was measured following a standardized procedure. First, broadband pink noise was reproduced through a KRK ROKIT 5 studio monitor acting as the acoustic source to generate a diffuse sound field within the enclosure. An ECM-8000 omnidirectional measurement microphone was calibrated to register a sound pressure level of approximately 80 dB at the measurement position. The room response was subsequently recorded and analyzed using Room EQ Wizard (REW) software (Version 5.31.3), which estimated the RT60 from the sound decay curve after the excitation signal was interrupted, as shown in Figure 6.



**Figure 6.** RT60 for the environment room. The solid line corresponds to the measured reverberation decay, whereas the dashed line represents the extrapolated decay employed by REW for the RT60 estimation.

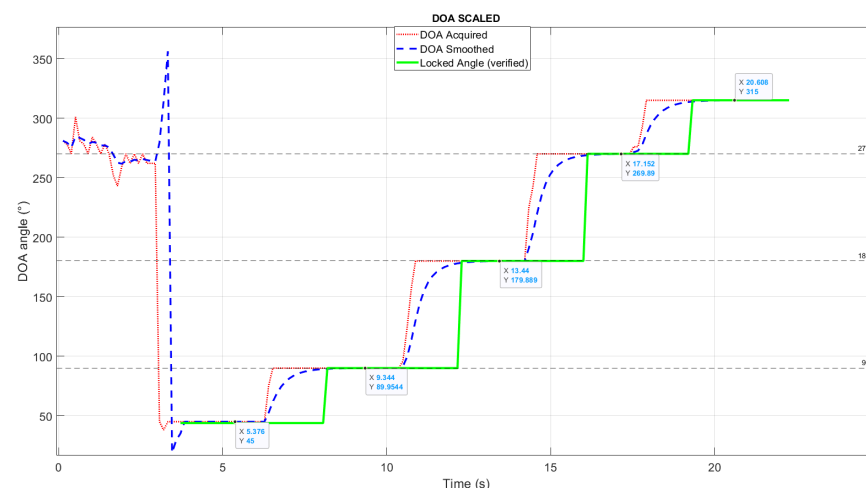
The first experimental scenario was designed to evaluate the accuracy of the proposed direction-of-arrival (DOA) estimation system at five predefined source locations:  $45^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ , and  $315^\circ$  at a distance of 75 cm from the center of the array as shown in Figure 7. These angles were intentionally selected to cover different regions of the microphone array,

including both principal axes and diagonal directions. Such a distribution enables the evaluation of the algorithm's angular consistency, verifies that the localization performance is not biased toward a specific direction, and provides a representative accuracy of the system.



**Figure 7.** Speaker positions for the first experimental scenario.

As shown in Figure 8, no active speaker is present during the first 4 s of the experiment. During this interval, the microphone array captures only ambient acoustic signals, resulting in arbitrary DOA estimates due to the absence of a dominant sound source. After 4 s, the speaker begins speaking from the predefined azimuth positions (45°, 90°, 180°, 270°, and 315°), allowing the proposed algorithm to perform continuous 360° angular scanning and estimate the direction of arrival in real time.



**Figure 8.** Comparison of the acquired DOA, smoothed DOA, and verified DOA used for beamforming.

The red dotted line represents the raw DOA estimated for each processing window using the GCC-PHAT algorithm. Although these instantaneous estimates rapidly respond to changes in the speaker position, they also exhibit abrupt fluctuations caused by background noise, reverberation, and estimation uncertainty. To improve temporal stability, an exponential smoothing stage is applied, producing the blue dashed line, which represents the filtered DOA. This smoothing process significantly reduces short-term variations while maintaining the ability to track changes in the speaker's position.

Finally, the green solid line corresponds to the verified or locked DOA used by the subsequent beamforming stage. Once the filtered DOA remains stable within a predefined angular tolerance for a sufficient period, the system validates the estimate and updates the

locked angle. As observed in the figure, the proposed approach successfully converges to each target direction while preventing unnecessary fluctuations, providing a robust and stable DOA estimate for real-time speaker localization and beamforming applications.

Table 1 compares the azimuth estimation accuracy of different sound source localization methods at a distance of 50 cm. The proposed system achieves very small angular errors, typically below  $1^\circ$ , which demonstrates a highly accurate Direction of Arrival (DOA) estimation performance. Compared with the work of Catur and H.M. Saputra [1], the proposed method shows significantly better stability and precision. Their GCC-PHAT implementation using a three-microphone equilateral triangle array reported errors up to approximately  $10^\circ$ , while the proposed system maintains errors close to zero for most tested angles. This improvement is mainly related to the use of a four-microphone MEMS array, better microphone synchronization, and more robust real-time processing with overlapping analysis windows.

**Table 1.** Single Source Localization at 50 cm distance.

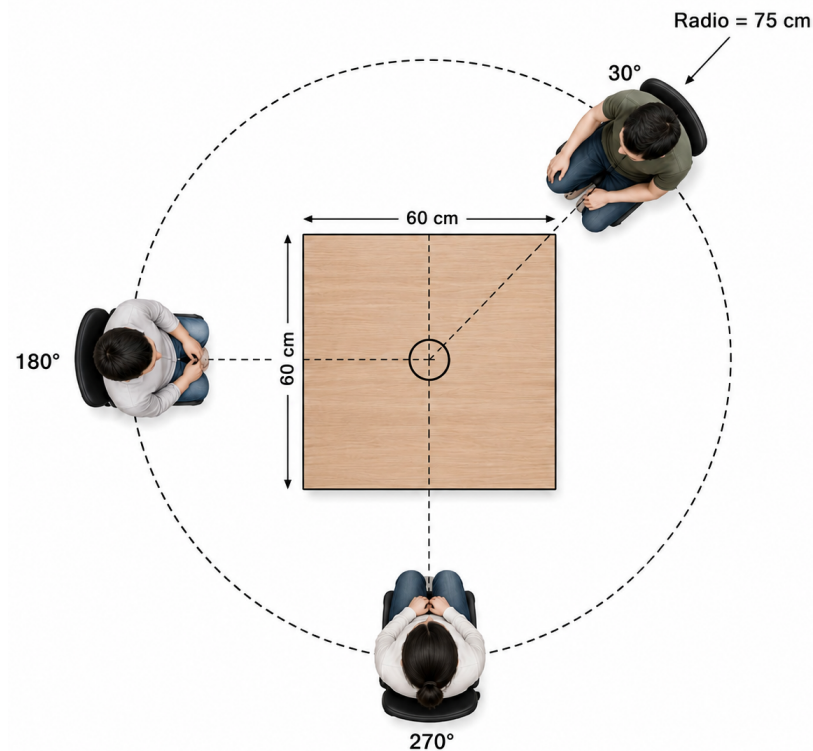
| Expected Azimuth | Catur and H. M. Saputra [1] | Y. Jiang [2] | Hunny [3]                        | OURS                             |
|------------------|-----------------------------|--------------|----------------------------------|----------------------------------|
| $0^\circ$        | $351.05^\circ$              | $1.90^\circ$ | $0.16^\circ$                     | <u><math>0.12^\circ</math></u>   |
| $45^\circ$       | -                           | $47.9^\circ$ | $45.15^\circ$                    | <u><math>45.00^\circ</math></u>  |
| $90^\circ$       | $96.04^\circ$               | -            | <u><math>89.90^\circ</math></u>  | $89.52^\circ$                    |
| $120^\circ$      | -                           | -            | <u><math>120.12^\circ</math></u> | $119.81^\circ$                   |
| $150^\circ$      | -                           | -            | $150.13^\circ$                   | <u><math>150.06^\circ</math></u> |
| $270^\circ$      | $262.67^\circ$              | -            | -                                | <u><math>270.07^\circ</math></u> |
| $300^\circ$      | $288.65^\circ$              | -            | -                                | <u><math>299.16^\circ</math></u> |
| $330^\circ$      | $321.42^\circ$              | -            | -                                | <u><math>329.63^\circ</math></u> |

Compared with the methodology proposed by Y. Jiang et al. [2], the proposed system achieves comparable performance despite using a computationally simpler architecture. Jiang's work combines DUET source separation with SRP-PHAT beamforming, which performs a spatial power search over multiple microphone pairs simultaneously. SRP-PHAT is generally more robust in reverberant environments because it integrates information from all microphone pairs and reduces localization ambiguity. However, this approach also introduces significantly higher computational complexity. The proposed system achieves similar angular precision using a direct GCC-PHAT-based TDOA estimation, demonstrating that careful implementation and optimized real-time processing can provide near-SRP-PHAT accuracy with lower computational cost.

The comparison with the work of Hunny [3] requires additional consideration because their methodology differs substantially from conventional GCC-PHAT localization systems. Their approach combines TDOA, Interaural Level Difference (ILD), and Head-Related Transfer Function (HRTF) techniques together with a hemispherical reflector structure. Although the reported errors are extremely small, the experiments appear to be performed under highly controlled conditions and depend strongly on the physical reflector geometry. The proposed system, on the other hand, operates using a compact MEMS microphone array in realistic real-time acquisition conditions without relying on external acoustic reflectors or idealized laboratory configurations, making it more practical and adaptable for embedded DOA applications.

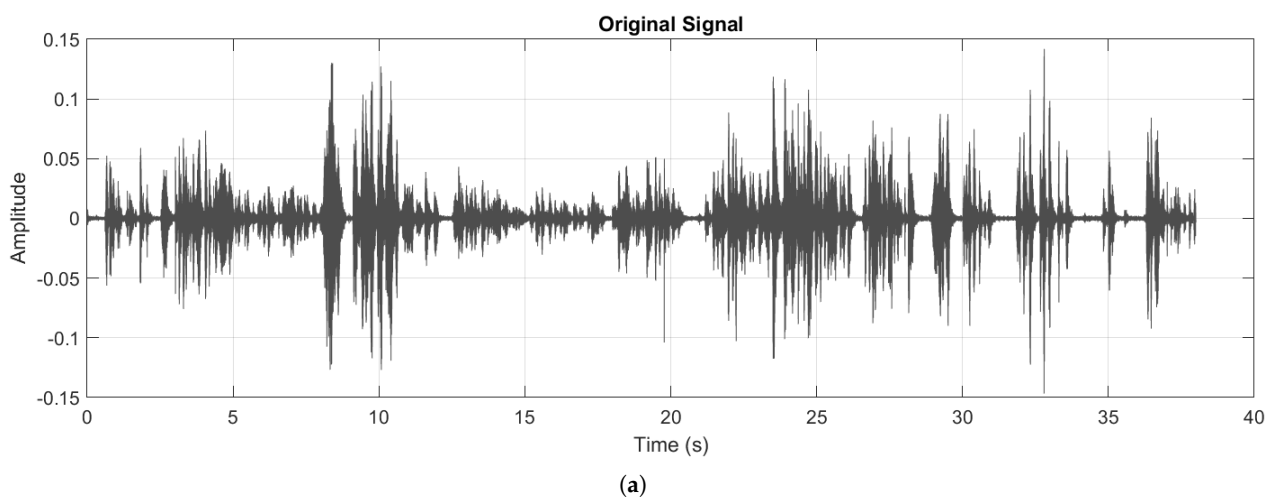
The second experimental scenario was designed to evaluate the accuracy of the proposed DOA+DAS system in enhancing the speech signal received by the microphone array. This experiment assessed the system's performance using three speakers positioned at azimuth angles of  $30^\circ$ ,  $180^\circ$ , and  $270^\circ$ , each located approximately 1 m from the microphone array as shown in Figure 9. This experiment lasted 38 s and the speakers were allowed

to move freely within their respective positions to emulate the natural body movements that typically occur during a normal conversation. This setup provided a more realistic evaluation of the proposed system under practical speaking conditions while maintaining a controlled acoustic environment.

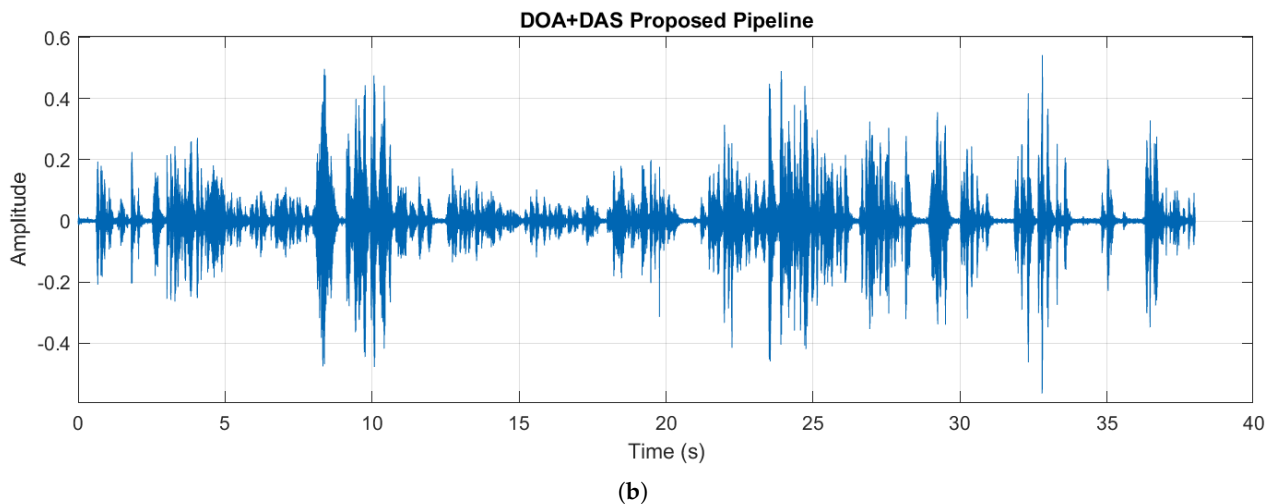


**Figure 9.** Speaker positions for the second experimental scenario.

Figure 10 shows a comparison between the original audio signal and the output of the proposed DOA+DAS pipeline. The original signal exhibits a maximum amplitude of approximately 0.15, whereas the beamformed output reaches a maximum amplitude of approximately 0.45. This increase occurs because the Delay-and-Sum (DAS) beamforming algorithm coherently combines the aligned signals received by the microphone array. Since the desired speech components are added constructively after time alignment, the resulting signal exhibits a higher amplitude while also improving the signal-to-noise ratio (SNR) for the target speaker.



**Figure 10.** Cont.



**Figure 10.** Audio signals for experimental scenario No. 2. (a) Original Audio Signal. (b) DOA+DAS Proposed Pipeline.

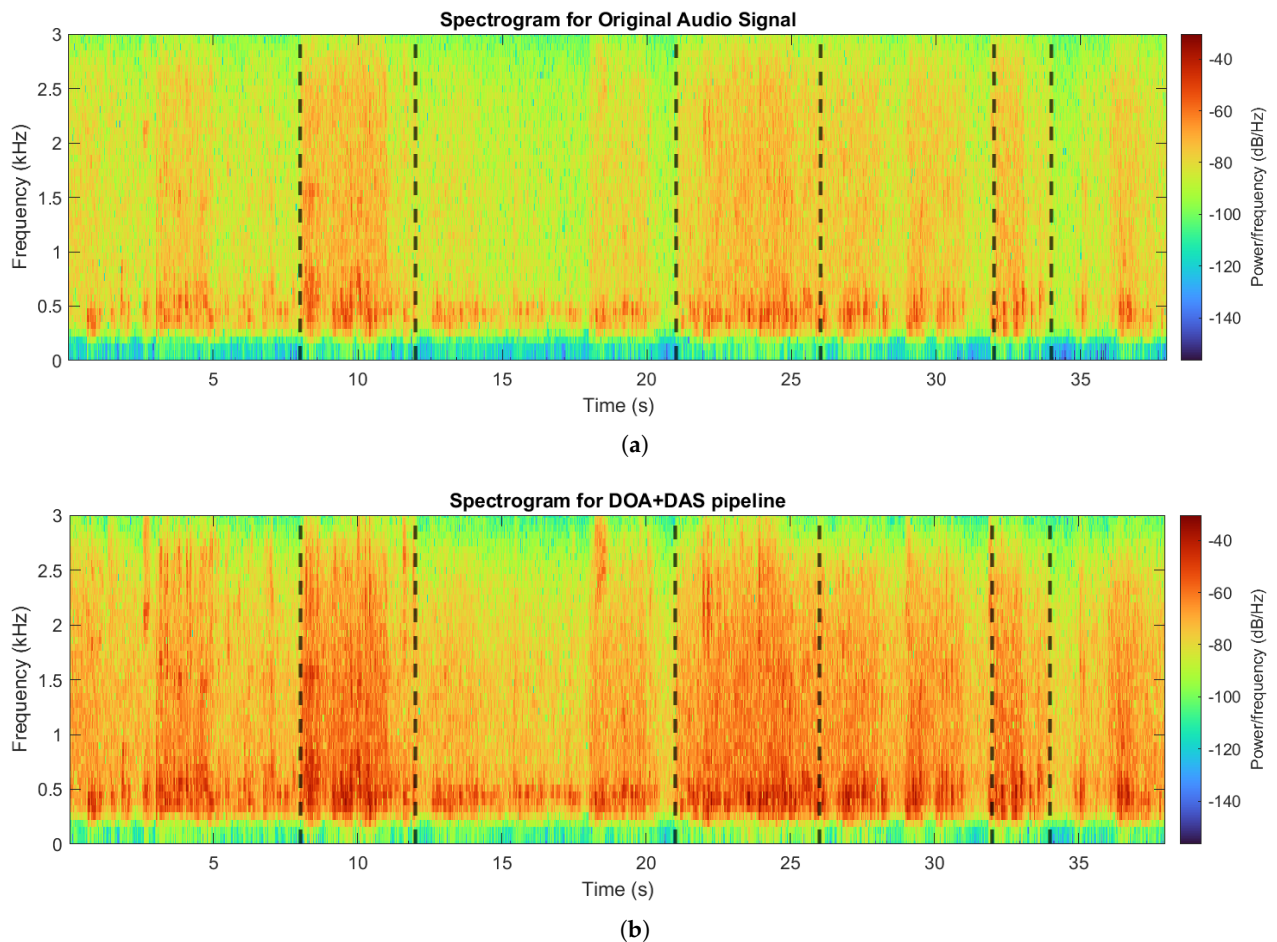
The spectrograms presented in Figure 11 compare the original audio signal with the signal processed by the proposed DOA+DAS pipeline. Both spectrograms exhibit similar temporal structures, indicating that the beamforming stage preserves the spectral content of the speech without introducing noticeable distortions or artifacts. The dominant speech energy remains concentrated below approximately 0.6 kHz, while harmonic components extend up to nearly 3 kHz, which corresponds to the frequency range expected for human speech.

Compared with the original signal, the spectrogram of the DOA+DAS output exhibits a greater concentration of energy in the speech regions. This can be observed by the larger areas represented in orange and red colors, particularly during the active speech intervals around 8–12 s, 22–27 s, and 30–37 s. The color scale also indicates that the maximum spectral power increases from approximately  $-50$  dB/Hz in the original signal to approximately  $-40$  dB/Hz after beamforming, representing an increase of nearly 10 dB in the strongest speech components.

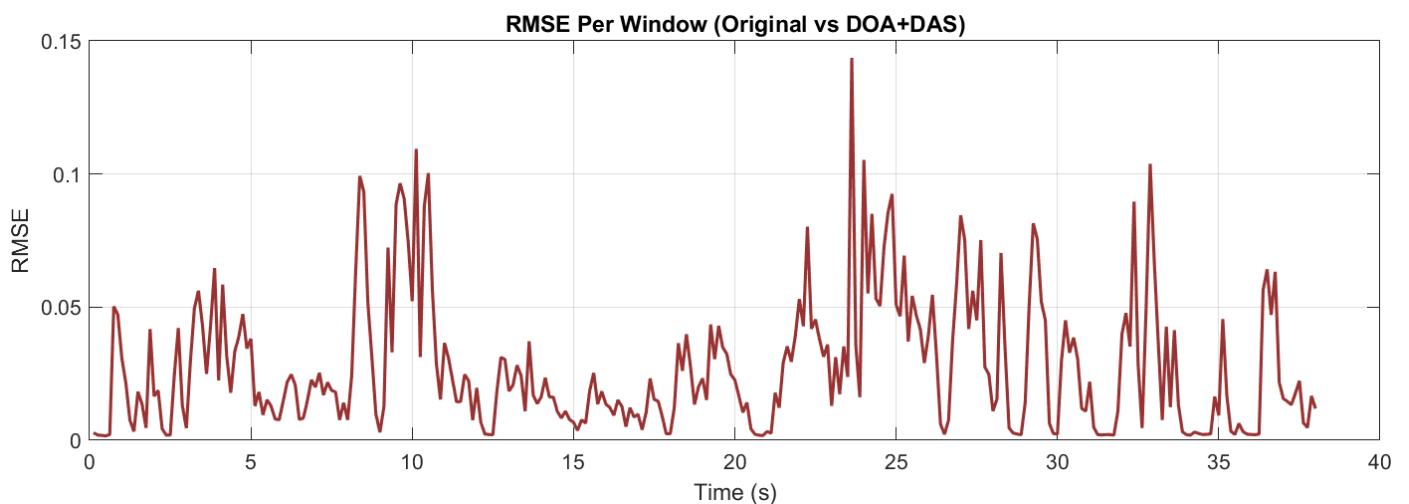
Figures 12 and 13 present the Root Mean Square Error (RMSE) and the Signal-to-Noise Ratio (SNR) computed on a window-by-window basis for the comparison between the original signal and the signal processed using the proposed DOA+DAS Pipeline.

The RMSE results presented in Figure 12 demonstrate the effectiveness of the proposed method. For most of the recording, the RMSE remains below 0.05, indicating a high degree of similarity between the original signal and the beamformed output. Such low error values demonstrate that the DOA+DAS algorithm preserves the temporal and amplitude characteristics of the desired signal while performing noise reduction. This behavior suggests that the beamforming process introduces only minimal distortion into the reconstructed waveform.

Several isolated RMSE peaks are observed during the recording, with the largest value approaching approximately 0.14. These increases occur only over short time intervals and are likely associated with the same acoustic conditions responsible for the temporary reductions in SNR, such as abrupt changes in the propagation environment, increased background noise, or transient errors in the DOA estimation. Despite these isolated events, the RMSE rapidly returns to low values, indicating that the overall signal reconstruction remains accurate and stable throughout the recording.



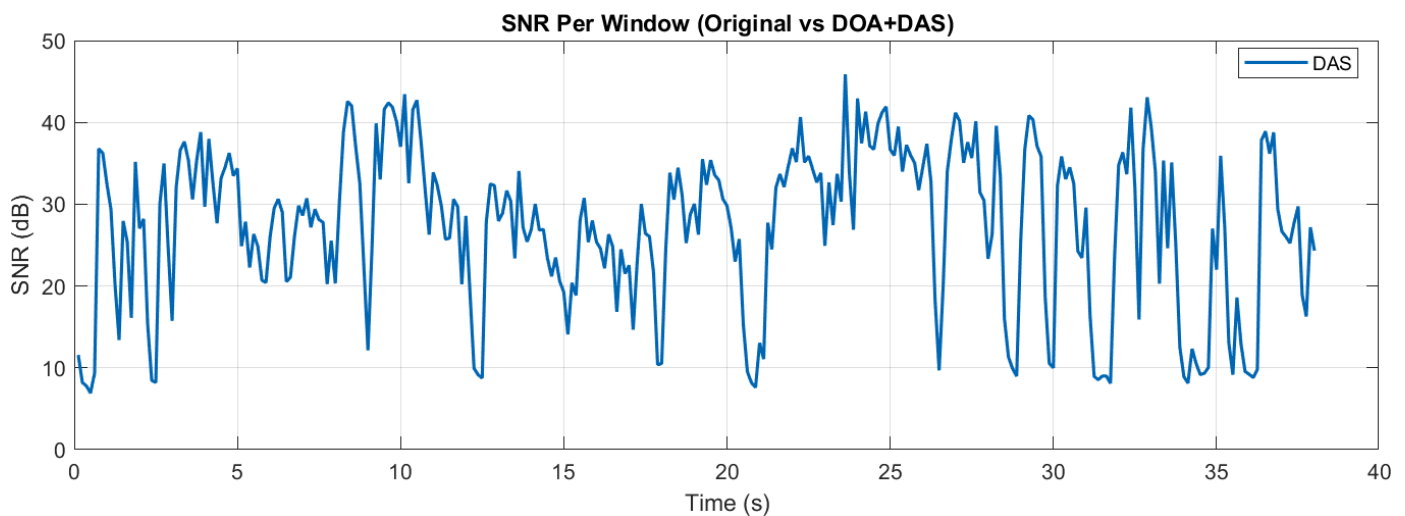
**Figure 11.** Spectrograms for experimental scenario No. 2. (a) Spectrogram of Original Audio Signal. (b) Spectrogram of DOA+DAS Proposed Pipeline.



**Figure 12.** RMSE for experimental scenario No. 2.

The SNR results shown in Figure 13 indicate that the proposed DOA+DAS approach maintains a relatively high signal quality throughout most of the recording. The majority of the analysis windows exhibit SNR values ranging from approximately 25 dB to 40 dB, with several intervals reaching values close to 45 dB. These results indicate that the desired acoustic signal remains significantly stronger than the background noise after beamforming, demonstrating the effectiveness of the proposed processing method. The consistently high

SNR values suggest that the beamformer successfully reinforces the target source while attenuating unwanted acoustic interference.



**Figure 13.** SNR for experimental scenario No. 2.

Although the overall SNR performance is favorable, several short-duration decreases can be observed where the SNR drops to values between 8 dB and 12 dB. These temporary reductions are expected in real acoustic environments and may result from changes in the position of the sound source, variations in the surrounding noise level, reverberation effects, or small inaccuracies in the estimated direction of arrival. Importantly, these decreases are brief and are followed by a rapid recovery of the SNR, indicating that the beamforming algorithm remains stable and quickly adapts to changing acoustic conditions without prolonged degradation in performance.

To evaluate the feasibility of real-time implementation on edge devices, the computational cost of the proposed DOA estimation framework was analyzed. The execution time of the main processing stages, including GCC-PHAT calculation, angular search, and delay-and-sum beamforming, was measured for each audio frame. The algorithm processes frames of 2048 samples at a sampling frequency of 16 kHz, corresponding to an audio frame duration of 125 ms.

The average execution time of the complete processing pipeline was 3.473 ms, where GCC-PHAT, DOA estimation, and delay-and-sum beamforming required 0.615 ms, 0.357 ms, and 0.176 ms, respectively. The obtained processing time corresponds to a real-time factor (RTF) of 0.0278, indicating that the algorithm requires only approximately 2.78% of the available frame time for computation. Additionally, the estimated memory consumption of the main processing buffers was 67.11 KB, demonstrating the suitability of the proposed method for resource-constrained embedded platforms. A summary of these data is presented in Table 2.

**Table 2.** Runtime and memory consumption analysis.

| Parameter                         | Measurement |
|-----------------------------------|-------------|
| Average processing time per frame | 3.473 ms    |
| GCC-PHAT processing time          | 0.615 ms    |
| DOA estimation processing time    | 0.357 ms    |
| Beamforming processing time       | 0.176 ms    |
| Audio frame duration              | 125 ms      |
| Real-Time Factor (RTF)            | 0.0278      |
| Memory consumption                | 67.11 KB    |

## 4. Conclusions

This paper presented a multichannel acoustic beamforming system for real-time speaker localization and signal enhancement based on the combination of Generalized Cross-Correlation with Phase Transform (GCC-PHAT) and Delay-and-Sum (DAS) beamforming. Unlike conventional localization approaches that rely on computationally demanding techniques such as SRP-PHAT or high-density microphone arrays, the proposed methodology estimates the Direction of Arrival (DOA) by selecting the microphone pair with the strongest GCC-PHAT response, followed by vector-based angular smoothing and a DOA verification stage to improve temporal stability before steering the beamformer. This design significantly reduces computational complexity while maintaining high localization accuracy, making it suitable for implementation on embedded and real-time systems.

The proposed system was experimentally validated using a four-element circular MEMS microphone array in a controlled indoor environment characterized by measured reverberation time (RT60) and background noise conditions. Two experimental scenarios were designed to evaluate both localization accuracy and beamforming performance. The first scenario demonstrated that the proposed localization algorithm accurately estimates the speaker direction over the entire 360° field of view, achieving angular errors generally below 1°, outperforming or matching more computationally intensive approaches reported in the literature while requiring fewer processing resources. The incorporation of DOA smoothing and angle verification effectively reduced abrupt estimation fluctuations, providing stable tracking suitable for continuous real-time operation.

The second experimental scenario evaluated the effectiveness of the complete DOA+DAS processing pipeline using multiple speakers under realistic conversational conditions. Time-domain and spectrogram analyses demonstrated that the beamforming stage successfully reinforced the desired speech signal while attenuating interference from other directions. The processed signal exhibited a noticeable increase in amplitude due to the coherent summation of aligned microphone signals, and the corresponding spectrograms showed higher energy concentration within the dominant speech frequency bands without introducing significant spectral distortion. These observations confirm that the proposed beamforming strategy preserves the essential temporal and spectral characteristics of the target speech while improving its intelligibility.

Overall, the experimental results demonstrate that accurate DOA estimation and effective beamforming can be achieved using a compact four-microphone array without the computational burden associated with exhaustive spatial-search algorithms. The proposed methodology provides an attractive compromise between localization accuracy, signal enhancement and implementation simplicity, making it well suited for embedded applications including intelligent voice assistants, teleconferencing systems, service robotics, smart environments, hearing assistance devices, and human–machine interaction platforms where low latency and real-time performance are essential.

**Author Contributions:** J.A.L.-O. is the principal author of this paper, responsible for the conception, development, and implementation of the methodology, as well as conducting the experiments and analyzing the results. H.M.P.-M., E.G.-R., E.E.-H., J.M.C.-C., J.B.-M. supervised the proposal, design of experiments, result analysis, and the writing process of the manuscript. All authors confirm that neither the manuscript nor any parts of its content are currently under consideration for publication with or published in another journal. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Catur, H.A.H.B.B.; Saputra, H.M. Azimuth Estimation Based on Generalized Cross Correlation Phase Transform (GCC-PHAT) Using Equilateral Triangle Microphone Array. In *Proceedings of the 2019 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*; IEEE: Piscataway, NJ, USA, 2019; pp. 89–93. [CrossRef]
2. Jiang, Y.; Hang, R.; Liu, B.; Yang, S.; Zhang, X.; Xu, Y. Development of a High-Precision Multi-Source Localization System Based on DUET-SRP-PHAT. In *Proceedings of the 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. [CrossRef]
3. Hunny. Microphone Array Based Localization of Sound Source. In *Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–3. [CrossRef]
4. Seed Technology Co., Ltd. ReSpeaker USB Mic Array. 2021. Available online: <https://www.seedstudio.com/ReSpeaker-USB-Mic-Array-p-4247.html> (accessed on 3 March 2026).
5. Jia, L.; Zhang, G.; Liu, Y.; Bai, Z.; Geng, Y.; Wu, Y.; Zhang, J.; Zhang, W. Sonar buoy active detection and localization for underwater targets using high-level sound sources and MEMS hydrophone. *Measurement* **2025**, *241*, 115740. [CrossRef]
6. Zhao, Y.; Chen, X.; Wang, B. Real-time sound source localization using hybrid framework. *Appl. Acoust.* **2013**, *74*, 1367–1373. [CrossRef]
7. Wei, D.; Ran, Q.; Li, Y. New convolution theorem for the linear canonical transform and its translation invariance property. *Optik* **2012**, *123*, 1478–1481. [CrossRef]
8. Diaz-Guerra, D.; Beltran, J.R. Source cancellation in cross-correlation functions for broadband multisource DOA estimation. *Signal Process.* **2020**, *170*, 107442. [CrossRef]
9. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [CrossRef]
10. van der Merwe, R.; Doucet, A.; de Freitas, N.; Wan, E. The Unscented Particle Filter. In *Proceedings of the Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2000; Volume 13.
11. Wang, R.; Chen, Z.; Yin, F. Speaker Tracking Based on Distributed Particle Filter and Interacting Multiple Model in Distributed Microphone Networks. *Appl. Acoust.* **2021**, *174*, 107741. [CrossRef]
12. Huang, G.; Chen, J.; Benesty, J. On the Design of Robust Steerable Frequency-Invariant Beampatterns with Concentric Circular Microphone Arrays. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2018; pp. 506–510. [CrossRef]
13. Luo, X.; Huang, G.; Chen, J.; Benesty, J. On the Design of Robust Differential Beamformers with Uniform Circular Microphone Arrays. In *Proceedings of the 31st European Signal Processing Conference (EUSIPCO)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [CrossRef]
14. Kohout, B. Transducer Array Calculation (TAC) GUI. MATLAB Central File Exchange. 2012. Available online: <https://www.mathworks.com/matlabcentral/fileexchange/35657-transducer-array-calculation-tac-gui> (accessed on 3 March 2026).
15. Perrot, V.; Polichetti, M.; Varray, F.; Garcia, D. So you think you can DAS? A viewpoint on delay-and-sum beamforming. *Ultrasonics* **2021**, *111*, 106309. [CrossRef] [PubMed]
16. Prodec Group. Conference Hall Acoustic Design: Professional Standards & Implementation Guide. 2026. Available online: <https://www.prodecgroup.com/architectural-acoustic-standards/conference-hall-acoustic-design/> (accessed on 3 March 2026).
17. The Engineering ToolBox. Noise Criterion (NC) Levels: Definitions, Standards & Calculator. 2004. Available online: [https://www.engineeringtoolbox.com/nc-noise-criterion-d\\_725.html](https://www.engineeringtoolbox.com/nc-noise-criterion-d_725.html) (accessed on 3 March 2026).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.