Research Article

A NOVEL METHOD FOR MULTIPLE SOUND SOURCES LOCALIZATION WITH LOW COMPLEXITY

Nguyen Trung HIEU[1,*], *Kou YAMADA*[2]

[1]Faculty of Electronics Engineering, Posts and Telecommunications Institute of Technology, 122 Hoang Quoc Viet Street, HaNoi, Vietnam
[2]Graduate School of Science and Technology, Gunma University, Gunma, Japan

hieunt@ptit.edu.vn, yamada@gunma-u.ac.jp

[*]Corresponding author: Nguyen Trung Hieu; hieunt@ptit.edu.vn

DOI: 10.15598/aeee.v23i3.240708

Article history: Received Jul 22, 2024; Revised Jan 02, 2025; Accepted Jan 25, 2025; Published Dec xx, 202x.

Abstract. *Sound source localization is essential in many areas such as robotics interaction, teleconferencing, sound extraction and recognition, noise cancellation in vehicles, object location detection, assessment of noise pollution in living spaces, and search and rescue. Interaction in natural settings requires the detection of different sources of sounds from the environment. Accurately detecting and differentiating incoming sound directions always attracts attention and has been researched using various methods. However, most of these methods still require complex algorithms or large amounts of calculations, which are accompanied by the cost of hardware and system resources. In this paper, we present a novel method and metrics for estimating the direction of multiple sound sources based on a combination of beamforming, time difference of arrival (TDOA), and frequency sparsity. Our new proposals are well-suited for deployment on resource-limited devices, offering reduced processing complexity, short computation time, and real-time response.*

Keywords

Beamforming, TDOA, frequency sparsity, DOA estimation, sound source localization (SSL), microphone array, lightweight, real-time.

1. Introduction

Sound source direction detection is a primary importance in assisting sound acquisition, exchanging information between objects, and navigating other related information collection devices. Indeed, speaking and hearing are crucial senses for communication and socialization. Interestingly, humans can identify the sound source positions around them without relying on vision. Furthermore, with that information, we can even focus on the desired sound in a noisy environment. These abilities enable us to process sound more effectively. In modern times, interactions occur not only between humans but also between humans and machines. This is accompanied by a demand for quality and speed in practical sound processing techniques. For this reason, sound source localization has gained more and more attention nowadays due to its wide areas of applications, such as teleconference, sound extraction and recognition, robot audition, search and rescue operations, and aeroacoustics.

The DOA estimation is closely linked to the hardware usage, with almost every study using more than two microphones. These microphones are installed on various platforms linked with many characteristics and approaches. Some methods that simulate head shape provide algorithms such as the Inverse Head-Related Transfer Function (IHRTF) [1] or multidirectional reception microphone arrays [2, 3] to calculate intensity difference. Another significant type of hardware employs the microphone array with three main techniques: multi-signal classification (MUSIC) [4, 5, 6, 7, 8], correlation-based approach (of which the common is treatment TDOA) [9, 10, 11, 12], beamforming-based approaches [13, 14, 15, 16]. The microphone array model is com-

monly used and offers many advantages for detecting the sound source direction.

The MUSIC method (first proposed by Schmidt in 1986 [4]) takes advantage of the differentiation between signal and noise subspace. It falls under the category of the so-called 'high-resolution' approach due to the sharpness of its results. However, since MUSIC assumes that the source signal is narrowband, applying it to sound signals (which are broadband) requires performing multiple narrowband MUSIC across the frequency bins. Therefore, some extended versions of this method for acoustic applications, especially robotics auditions, have been developed, including SEVD-MUSIC [5], GEVD-MUSIC [6]. However, in addition to those performance advantages, they all require a significant amount of computational resources to operate in real-time. Nakamura proposed his own upgraded version called GSVD-MUSIC [7], which significantly reduced the computational cost to address this issue. Using the information-theoretic detection method, Lunati implemented the MUSIC+MAICE version on a system-on-a-programmable-chip (FPGA Virtex 4) [8] achieving a processing time of about 22ms for each SSL.

Another approach is to exploit the TDOA between each pair of the microphone in a microphone array. The most common method for estimating TDOA is GCC-PHAT (first proposed by Knapp [9]). Since then, many adaptations have been made to enhance the performance of GCC-PHAT for different purposes. However, it still faces issues, especially in the presence of broadband noise when PHAT is applied [10]. More improved versions have been introduced to solve this, involving the application of different weighting functions [11, 12] based on the online noise calculation. Furthermore, it is noteworthy that most of the cross-correlation approaches only focus on a single DOA estimation, which somehow reduces their practical appliance ability.

Beamforming-based method is perhaps the most widely used strategy for source localization. This technique relies on spatial filtering of the signals captured by a microphone array, enhancing the array's signal-capturing capability in the desired direction. The most basic beamforming technique, also known as delay-and-sum (DAS) beamforming, involves scanning over a set of desired DOAs. This process creates a steered-response power (SRP) map, which can be used to accentuate multiple sound sources. However, it requires one beamformer per proposed DOA, and also, beamforming performance depends on the resolution of the grid search. Consequently, it is classified as a hardware-heavy method [13]. Additionally, every beamforming map can be considered a "dirty map" because it is spoiled by the influence of the array geometry and the presence of side-lobes (a side effect when forming the beam) [14], which results in signals from non-focused directions. To overcome these challenges, researchers have recently proposed various variations of beamforming. Some notable examples include Iterative Capon MVDR [15] and Functional beamforming [16]. Functional Beamforming involves adjusting the total beamformer output power to enhance its robustness. It also utilizes a method known as eigenvalue decomposition of the cross-spectral matrix (CSM), which is also used in Orthogonal beamforming to separate the signal and noise subspace (alike MUSIC) [17].

Furthermore, to improve the resolution and obtain a "cleaner map", deconvolution approaches like CLEAN-SC [18] and DAMAS [19] have been developed. However, because of the "super-resolution" results they produce, these deconvolution methods require an extreme amount of computational load. Most importantly, all of these improved beamforming-based methods only focus on enhancing the results and still rely on the DOA scanning, which is considered an exhaustive search and forms the basis of the grid scan resolution and side-lobes problem [20].

Therefore, inversion methods have been gaining increasing interest over the years. Although these methods are not classified as beamforming since they do not form beams or perform scans, they are still categorized as beamforming-based because their idea is to find the combination of sources that can reconstruct pressure as closely as possible to the actual pressure measured at each measurement point (hence the name 'inverse'). The calculation depends on the source definition, such as plane waves for SONAH [21] or a cloud of monopoles for Bayesian approaches [22, 23]. However, these methods can be susceptible to noise, require a regularization procedure, and are generally underdetermined [24]. Iterative inverse methods have been presented to address this issue, such as the L1-Generalized beamforming [25]. Nevertheless, this method brings about the computational cost problem once again.

It is worth mentioning that the source and array's specific properties have been shown to affect DOA estimation [26, 27]. Therefore, in combination with the existing methods, they are often utilized to enhance the efficiency of the DOA estimation process. For example, sparsity in time-frequency domains is a valuable feature that can be exploited for sound source localization and blind source separation [28, 29, 30]. Additionally, it is noteworthy that sound, especially speech, often exhibits robust short-term correlations. The signal power is not evenly distributed across the entire frequency range, even though it is a wide-band signal; instead, it is concentrated at a set of equally spaced discrete frequency points, i.e., harmonics of the pitch frequency.

As can be seen, various strategies have been employed in the SSL field, each with its own set of strengths and weaknesses. Several challenges are still being addressed by researchers interested in overcoming them. For ex-

ample, when real-time computation is required, some methods still demand a significant amount of computational resources. The hardware size and the necessity of using large microphone arrays like beamforming and MUSIC also limit their practical applicability. Furthermore, implementing SSL in real-life situations remains a significant challenge, as it requires the ability to operate in flexible environments with noise, reverberation, and variations in the number and movement of sound sources.

Sound source localization is a challenging task that traditionally relies on complex and computationally intensive techniques. This complexity can lead to slower processing speeds, increased hardware requirements, and higher costs, making it a formidable challenge for many applications. Our method addresses these challenges by combining beamforming, time difference of arrival (TDOA), and frequency sparsity in a way that reduces computational complexity while maintaining high accuracy. In addition to its simplicity, our method can be executed in real-time on low-resource devices, such as the ARM STM32 microcontroller, which is widely available and easy to use.

Next, the main contribution of this paper can be summarized as follows:

- Hardware Requirements: The proposed method is designed to run on a low-resource device, such as the ARM STM32 microcontroller, which has limited computational power and memory. Other methods often require more powerful hardware, such as PCs, FPGAs, or larger arrays of microphones, making the proposed method more efficient in terms of hardware use.
- Processing Time: The paper reports an average processing time of 20ms, including 16ms for data sampling and 4ms for calculations. Other methods, such as those using SRP-PHAT or MUSIC, typically have longer processing times due to complex algorithms and hardware requirements. This makes the proposed method faster and more suitable for real-time applications.
- Computational Load: The method avoids complex matrix operations, Fourier transforms, or DOA scans, which are computationally expensive in other methods like beamforming or MUSIC. Instead, it uses a simplified scoring method (SCORE) that significantly reduces the computational burden.
- Memory Usage: The implementation occupies only 44KB of memory on the microcontroller, making it lightweight compared to other methods that need larger memory resources for complex calculations or larger microphone arrays.
- Algorithmic Simplicity: By using frequency sparsity and simpler data types (phase shifts rather than full signal processing), the method reduces the overall algorithmic complexity compared to traditional approaches like GCC-PHAT or beamforming, which involve heavy signal processing and grid searches.

This paper is divided into in 6 sections. Section 2 presents background knowledge of the techniques used to develop our proposal. Our novel approach, including the methodology and calculation system, is presented in section 3. Section 4 covers hardware implementation and performance evaluation. Results and discussions of related issues are presented in Section 5. Finally, Section 6 concludes the paper and recommends directions for future development.

2. Background

Beamforming is a signal processing technique used for directional signal transmission. It requires multiples antennas in close proximity (phased array), all broadcasting the same signal but at a slightly different time (phase shifting). This causes the signals to experience constructive/ destructive interference in particular directions. For example, Fig. 1 demonstrates how the

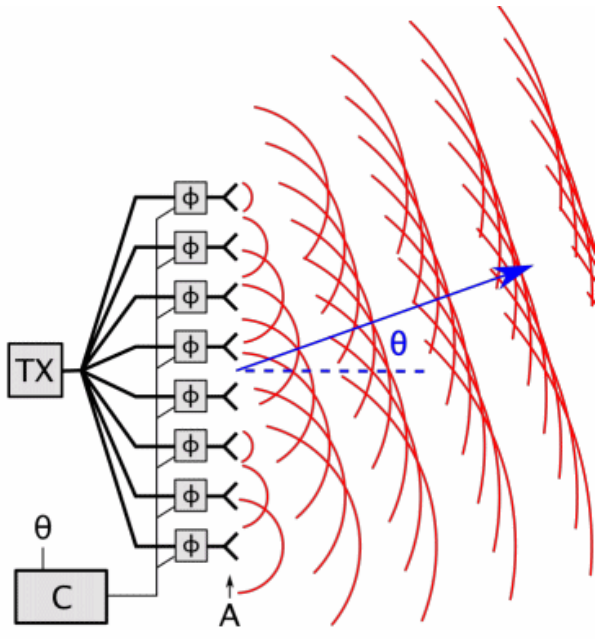


Fig. 1: Beamforming demonstration.

simplest form of beamforming, known as delay-and-sum beamforming (DAS) works. It contains an array of antenna elements (A) powered by a transmitter (TX). The feed current for each antenna passes through a phase shifter (ϕ) controlled by a computer (C). The individual red waves are omnidirectional or spherical

shapes but they combine and create an overlapping plane wave or a beam of the signal wave traveling in a specific direction. The phase shifters delay the signal so that each antenna radiates its wave later than the one below it. This causes the beam to be directed at an angle θ to the antenna's axis. By adjusting the phase shifts, we can steer the beam in the desired directions.

This means that there is a direct relationship between the phase shift ϕ and the beam angle θ, and we can leverage this relationship using the same theory when capturing acoustic signals. In other words, it can be used to estimate the direction-of-arrival (DOA) of sound waves by utilizing the time difference of arrival (TDOA) of the signals captured in an array of sensors as delayed time. This relationship between them can be described as [27]:

$$\theta = \arcsin\left(\frac{V_{sound} \times \tau}{f_{sample} \times d}\right), \tag{1}$$

where V_{sound} is the speed of sound (∼340 m/s); f_{sample} is the sampling frequency in Hz; d is the distance between microphones in meters; τ is the TDOA of the sound source in samples, and θ is angle of the coming sound to the microphone array's axis.

With this relationship, we can apply the DAS beamforming processing technique to signals captured in an array of microphones to synchronize back the signal as it came from a specific direction. For example, assuming

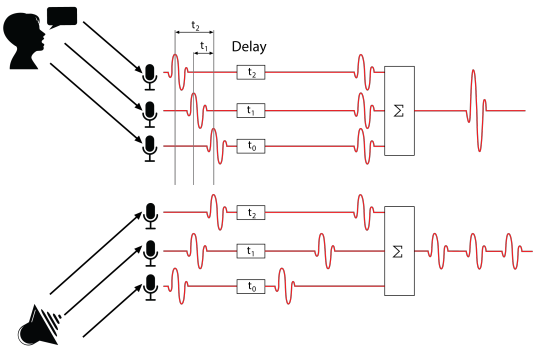


Fig. 2: Apply DAS beamforming technique to DOA estimation.

the sound DOA is approaching from a human speech source, as shown in the upper part of Fig. 2. Using (1) we can determine the TDOA between microphones as t_1, t_2. The primary purpose of the DAS beamformer is to artificially shift the signals to counter such time difference and then add the shifted signal to simulate the assumed sound source's original waveform. As the pre-assumed direction gets closer to the actual DOA, the alignment between shifted signals gets tighter, resulting in a more amplified output. Applying the same process, but with the input coming at a different angle than the pre-assumed DOA, as depicted in the lower part of Fig. 2, the shifted signals do not line up, and resulting in a scattered, smaller output. Thus, with a recorded sound frame, a polar coordinate DOA estimation map can be created by scanning through a set of pre-assumed DOAs, steering the beamformer toward it, and measuring its output power. Here's how it is done, assuming the sound DOA creates an angle θ with the microphone array's axis, and the output of the DAS beamformer is given as [27]:

$$\hat{s}_\theta[t] = \sum_{n=1}^{N} x_n[t - \tau_n(\theta)], \tag{2}$$

where $\hat{s}_\theta$ is the beamformer's output; t is the time bin; x_n is the signal received at microphone n; N is the number of microphones; and $\tau_n(\theta)$ is TDOA of the source in microphone n with the incoming angle θ.

The energy of the beamformer's output steered towards θ (E_θ) can be calculated as [27]:

$$E_\theta = \sum_{t=0}^{T} \hat{s}_\theta[t]^2. \tag{3}$$

where T is time windows. Then, the time delay of DAS can be converted to phase delay of the signals in the frequency domain. However, in reality, an audio signal is a wideband signal, so in order to shift the signal in the time domain, a frequency band separation is required. Each band should be treated differently with different amounts of phase shift. This can be accomplished by using Fourier transform as [27]:

$$X_{n_{\tau_n(\theta)}}[f] = e^{-j2\pi f_n \tau_n(\theta)} X_n[f], \tag{4}$$

where $X_n[f]$ is the Fourier transform of $x_n[t]$; f is the frequency bin; and $X_{n_{t_n}(q)}[f]$ is the Fourier transform of $x_n[t - \tau_n(\theta)]$. The time delay $\tau_n(\theta)$ is constrained by physical constraints based on microphone spacing and the speed of sound. The arrangement of shifts related to angle θ can be expressed as the complex-valued $N \times F$ matrix W_θ, as shown [27]:

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{-j2\pi f_1 \tau_2(\theta)} & e^{-j2\pi f_2 \tau_2(\theta)} & \cdots & e^{-j2\pi f_F \tau_2(\theta)} \\ e^{-j2\pi f_1 \tau_3(\theta)} & e^{-j2\pi f_2 \tau_3(\theta)} & \cdots & e^{-j2\pi f_F \tau_3(\theta)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-j2\pi f_1 \tau_N(\theta)} & e^{-j2\pi f_2 \tau_N(\theta)} & \cdots & e^{-j2\pi f_F \tau_N(\theta)} \end{pmatrix}, \tag{5}$$

where the f's are the frequency bin; N is the number of microphones; F is the frequency window size; and W_θ is the broadband steering matrix.

The Fourier transform of the output of the beamformer can be constructed via [27]:

$$\hat{S}_\theta[f] = W_\theta[f]^H X[f], \tag{6}$$

where $\{.\}^H$ is Hermitian transpose operator, X is a $N \times F$ matrix holding all of the X's elements in its rows, $W_\theta[f]$ is the column of W_θ holding the beamforming

weights for the frequency f, $X[f]$ is the column holding the frequency information in f out of N microphones, and $\hat{S}_\theta[f]$ is the beamformer's frequency domain output steered towards θ in the frequency f. Because $\hat{S}_\theta$ is the Fourier transform of $\hat{s}_\theta$, $\hat{s}_\theta = F^{-1}(\hat{S}_\theta)$. From $\hat{s}_\theta$ obtained, we can use (3) to get the energy value, and apply the same steps with different θ of the pre-assumed DOAs set.

However, employing beamforming for multiple DOA estimation has a crucial weakness. This process involves making predictions and subsequently evaluating those predictions. Accuracy depends on the density of the proposed DOA prediction set, and it requires one beamformer per proposed DOA. Moreover, due to the complexity of Fourier data and trigonometric Euler's function, executing this requires a substantial amount of computational resources. The following section presents our modified approach for multiple DOA estimation using beamforming.

TDOA is a feature that has proven to offer excellent advantages in DOA estimation. It can be calculated in various ways, typically cross-correlation. Furthermore, it can also be exploited in a custom way, based on the propagation pattern or specific array geometries to improve robustness as well as to reduce computational load [31, 32, 33]. Therefore, in this study, we will also use it in conjunction with an array of microphone geometry to simplify the algorithm.

Frequency sparsity is another property that provides excellent advantages in SSL with multiple sources or, furthermore, in blind source separation (BSS) [29]. When it comes to music or speech sources, when multiple sources play simultaneously, their signals are mixed on the time domain [28]. However, in the frequency domain, the signal strength is not evenly spread across the domain, but is concentrated at some points, i.e., the pitch's harmonic [30, 34]. We find that this particular property can be used for the performance improvement.

3. Proposed method to estimate direction of sound sources

The previous section explains the DAS beamforming, known as the basic principle of every beamforming-based approach. Section 1 also introduces various improvements developed from the conventional beamforming approach. However, these approaches mainly focus on increasing resolution, nullify side-lobe presence, or amplify robustness against noise and reverberation. This section will propose the adaptation from beamforming, which obtains advantages, in terms of unique, fast and light-weight.

3.1. New evaluation method

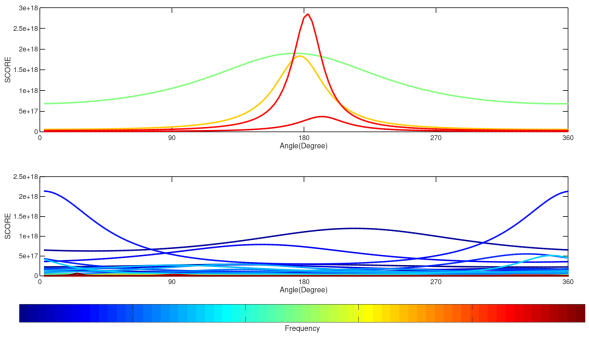


Fig. 3: Frequency domain processing demonstration.

To obtain the output energy E_θ of the proposed DOA from $\hat{S}_\theta$, we execute an Inverse Fourier transform (IFFT) $\hat{s}_\theta = F^{-1}(\hat{S}_\theta)$ and (3). However, as shown in Fig. 2, we have a direct relationship between the alignment of after-counter-shifted element signals and the amplitude of the output, the square root of E_θ. Therefore, we can take advantage of this relationship, and set up our evaluating method with a variable (called SCORE). The proposed SCORE can show us how well the DOA prediction performs without executing the IFFT. Indeed, the proposed SCORE is obtained by

$$SCORE_\theta = \sum_{f=1}^{F} \frac{\bar{E}[f]}{\sigma^2_{\hat{S}_\theta[f]}}, \tag{7}$$

where $\bar{E}[f]$ is the average energy captured in f, ${\sigma_{\hat{S}_\theta[f]}}^2$ denotes the variance of the after-counter-shifted signals steered toward θ, F is the frequency window size, and $SCORE_\theta$ is the evaluation score of the assumed DOA. We calculate the SCORE of the arrival angles, and then use the obtained SCORE value to find the angle at which the phase-shifted signals merge together, ensuring that the best direction is found.

The numerator is used to accentuate the actual sound away from the background noise. Here, we utilize the energy of the captured signal in each frequency bin which can be calculated in the frequency domain. As shown in (7), the signals in the frequency domain can be described in a complicated polar plane as follows: $X_n[f] = A_n[f]e^{jf_n[f]}$, where modulus $A_n[f] = \sqrt{Re(X_n[f])^2 + Im(X_n[f])^2}$. We then obtain the energy of the signal captured by microphone n in $f(E_n[f])$ as

$$E_n[f] = A_n[f]^2 = Re(X_n[f])^2 + Im(X_n[f])^2. \tag{8}$$

This captured energy can be uneven between microphones due to differences in distance from the source to a sensor or other hardware variations, but these differences are not significant. Therefore, we can use the mean $\bar{E}[f] = \frac{\sum_{n=1}^{N} E_n[f]}{N}$ as a measure of balance.

The denominator is the variance of $\hat{S}_\theta$ complex-values in the frequency domain. Its primary purpose is to evaluate the time synchronization of the after-shifted element signal as

$$\sigma^2_{\hat{S}_\theta[f]} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{S}_{\theta n}[f] - \bar{S}_\theta[f]\right)^2, \tag{9}$$

where $\hat{S}_\theta[f]$ is the Fourier transform of the output of beamformer in the frequency bin f, N is the number of microphones, $\bar{S}_\theta[f]$ is the mean of the output in f (i.e., $\bar{S}_\theta[f] = \frac{\sum_{n=1}^{N}\hat{S}_{\theta n}[f]}{N}$), and ${\sigma_{\hat{S}_\theta[f]}}^2$ is variance of the complex-valued output in f.

3.2. Metric in frequency domain

Let us take a closer look at (7), which contains two indicated components: θ is the scanning angle, and f is the frequency bin. At each candidate angle, the SCORE of a single frequency is calculated, and summed up repeatedly to obtain the results. This process creates a steering-response map, similar to beamforming, which can be considered a construct of multiple single frequency beamforming. However, these elements exhibit typical beam patterns, including the beam-width and side-lobes. When summed up, these characteristics can lead to results in the non-focus area, resulting in inaccurate and false peaks. Additionally, the peaks on the map are the places relied to decide the DOA of the sound sources. Hence, we can alternatively rearrange the construction order of the SCORE, perform the scan at each frequency, and focus only on its maximum value.

As mentioned in Section 2, the spoken audio signals in the frequency domain are not only stationary at one frequency but dispersed across a range of the related frequencies. Taking advantage of this feature, we have converted the evaluation method by using SCORE in the frequency domain:

$$SCORE_f = \max\left(\frac{\bar{E}_f}{\sigma^2_{\hat{S}_\theta,f}}\right), \forall\theta \in [0, 2\pi]. \tag{10}$$

Fig. 3 illustrates result of a multiple single-frequency execution with a speech signal played at an angle of $\phi = 180°$ on the azimuth plane. We set up a grid scan containing 128 scan points from 0° to 360°. The SCORE value of each frequency is calculated at each scan point, similar to performing multiple single-frequency beamforming. However, instead of using the output power as discussed in Section 2, we use our SCORE method. In other words, we break down (7) into a set of SCORE for each frequency. The resulting lines are color-coded resemble their frequencies, represented by blue to red corresponding to increasing frequency. To differentiate the results for easier analyzing, all the scans with SCORE's peak positions varing from 175° to 185° are plotted on the upper part, while the rest are plotted on the lower part. As we can see, the SCORE of the speech signal is relatively high, which is predicted since the signal's power is dominant. Additionally, not only one SCORE plot line matches the signal but four lines at different frequencies also match. However, there are only two results with moderately sharp peaks; the others are less sharp and can have a negative impact on the overall result. This pattern implies that using only the maximum SCORE can be simpler, more precise and open up more opportunities for further improvement.

3.3. Simplify the data type and function

The phase-shifting step in (4) shows that it only affects the argument $X_{n_{\tau_n(\theta)}}[f] = |X_n[f]|\, e^{j\phi_n[f]} e^{-j2\pi f_n \tau_n(\theta)} = |X_n[f]|\, e^{j(\phi_n[f] - 2\pi f_n \tau_n(\theta))}$. The uneven amplitude or modulus elements $A_n[f]$ should not be involved in time alignment evaluation ${\sigma_{\hat{S}_\theta[f]}}^2$, and it is also such a waste of computational resources to store and use complex data as shown in (9). Instead, we can use a much simpler argument data $\phi_{n,f}$ such that $\phi_{n,f} = \arctan(\mathrm{Im(X_{n,f})}, \mathrm{Re(X_{n,f})})$.

Next, the phase shifter function is reconstructed via simple addition/subtraction, as (11), without the need of complex-value matrices, Hermitian transpose or Euler's function:

$$\phi_{\theta_n,f} = (\phi_{n,f} - 2\pi f \tau_n(\theta)) \mod (2\pi), \tag{11}$$

where $\tau_n(\theta)$ is the TDOA of the source in microphone n related to the incoming angle θ; f is the frequency bin; $\phi_{\theta_n,f}$ is the argument of signal captured by microphone n in frequency domain in f; θ is the assumed DOA angle; and $\phi_{\theta_n,f}$ is the $\phi_{n,f}$ after-counter-shifted signal toward θ.

With this, equation (9) can be improved as

$$\sigma^2_{\phi_{\theta,f}} = \frac{1}{N}\sum_{n=1}^{N}\left(\phi_{\theta_n,f} - \bar{\phi}_{\theta,f}\right)^2, \tag{12}$$

with $\bar{\phi}_{\theta,f} = \sum_{n=1}^{N}\frac{\phi_{\theta_n,f}}{N}$. Therefore, we have the improved version of (10) as

$$SCORE_f = \max\left(\frac{\bar{E}_f}{\sigma^2_{\phi_{\theta,f}}}\right), \forall\theta \in [0, 2\pi]. \tag{13}$$

Combining with (11), (12) can also be expressed as

$$\sigma^2_{\phi_{\theta,f}} = \frac{1}{N}\sum_{n=1}^{N}\left[\phi_{n,f} - \bar{\phi}_f - 2\pi f\tau_n(\theta) + \sum_{n=1}^{N}\frac{2\pi f\tau_n(\theta)}{N}\right]^2. \tag{14}$$

However, the argument data that we are processing is in circular data form (data of a periodic nature). Any kind of linear treatment can lead us to

incorrect conclusions, including the usual arithmetic mean $\bar{\phi}_f = \frac{\sum_{n=1}^{N} \phi_{n,f}}{N}$. However, if we consider them as vectors, there is a natural way to calculate $\bar{\phi}_{\theta,f}$ through vector addition $\sum_{n=1}^{N} X_{n,f}$. After that, since we don't care about the magnitude, the angle difference can be calculated via multiplication as

$$\phi_{n,f} - \bar{\phi}_f = \arg(X_{n,f} \sum_{n=1}^{N} X_{n,f}), \tag{15}$$

where arg(.) denotes the argument of a complex number.

The $\phi_{n,f} - \bar{\phi}_f$ elements can be calculated as soon as the Fourier transformation complete. In (15), the function *arg* extracts the phase angle (or argument) of the complex number, representing the phase shift of the signal. This phase shift is critical for computing the phase difference between the signals received at different microphones, allowing for the alignment of the signals when estimating the direction of arrival (DOA).

3.4. Geometric meaning of microphone array

The previous modification leaves our SCORE with the $\tau_n(\theta)$ elements which have a direct relation with the DOA scan. Equation (1) mentioned this relationship as $\theta = \arcsin\left(\frac{V_{sound} \times \tau}{f_{sample} \times d}\right)$, but this only tells the angle created by the proposed DOA and the planar microphone array. These TDOAs are normally calculated based on a reference microphone acting as an origin point. Therefore, the TDOA links closely with the scan grid, microphone array, and can be pre-calculated, used as a lookup table. However, the planar array seems to suffer from an ambiguity problem that cannot specify the sound's direction from the front or the back of the array. Furthermore, a symmetrical array with the origin point at its center can nullify TDOA's summary value $\sum_{n=1}^{N} \tau_n(\theta)$ and keep the sum of its square value $\sum_{n=1}^{N} (\tau_n(\theta))^2$ at a constant in (14). This array type usage will simplify our method more and eventually lead to the calculation of the SCORE's maximum value without the DOA scan, which is a huge drawback of the overall beamforming-based method.

Assume that we have N points on a circle centered on O. Let $A_n(x_n, y_n)$ denote one of these points; a line passing $A_n(x_n, y_n)$ creates the angle α with the Ox axis, which can be considered as the sound direction. Let B_n be the intersection point of the normal line connecting O to that line. Let $D_n(\alpha)$ be the distance from the point A to the point B, which is the distance between microphone n and the origin point that sound waves have to travel more or less if they come from angle α. We then have

$$D_n(\alpha) = x_n \cos\alpha + y_n \sin\alpha, \tag{16}$$

where x_n, y_n are the microphone n coordinates in meters. In addition, we obtain that

$$\tau_n(\alpha) = -\frac{D_n(\alpha) \times f_{sample}}{V_{sound}}. \tag{17}$$

Equation (17) means that we can skip the DOA scan. Next, each frequency bin has its optimal DOA with an angle as

$$\alpha_f = \operatorname{arccot}\left(\frac{x_n}{y_n}\right) + k\pi,$$

where k is an integer. Thus, we have

$$\max \frac{\bar{E}_f}{\sigma^2_{\phi_\theta,f}} = SCORE_f.$$

Through experimentation, this SCORE tends to get boosted with a signal in the lower frequency since it contains the $4(2\pi f \times f_{sample} \times x_1/V_{sound})^2$ component. Moreover, this value, with $\sum_{n=1}^{N} (\phi_{n,f} - \bar{\phi}_f)^2$, can have different effects on the final prediction without any DOA information. Now, we consider the other SCORE function which relates to the DOA more strongly:

$$SCORE_f = \bar{E}_f \times \frac{\overline{\sigma^2_{\phi_\alpha,f}}}{\sigma^2_{\phi_\alpha,f}}. \tag{18}$$

where

$$\overline{\sigma^2_{\phi_\alpha,f}} = \frac{1}{N}\left[\sum_{n=1}^{N} (\phi_{n,f} - \bar{\phi}_f)^2 + 4\left(\frac{2\pi f \times f_{sample} \times x_1}{V_{sound}}\right)^2\right].$$

Finally, (18) is used to calculate and evaluate the direction of sound sources.

The main contributions of the proposal when compared with some previous studies as follows. The proposed method is designed to operate efficiently on low-resource devices, such as the STM32 microcontroller, which has limited computational power and memory. Unlike other approaches that often require powerful hardware like PCs, FPGAs, or larger microphone arrays, this method minimizes hardware requirements. By avoiding computationally expensive operations, such as matrix manipulations, Fourier transforms, or direction-of-arrival (DOA) scans, the method significantly reduces computational load through a simplified scoring technique (SCORE). It also stands out for its lightweight memory usage, unlike other methods that need larger memory resources for complex calculations or bigger microphone arrays. Furthermore, the method leverages frequency sparsity and simpler data representations, such as phase shifts, to reduce algorithmic complexity. This contrasts with traditional methods like GCC-PHAT or

beamforming, which rely on intensive signal processing and grid searches, making the proposed method a faster and more practical solution for real-time applications.

4. Experimental and evaluation

The experiment involves a simple 4-microphone grid array, as shown in Fig. 4, with a spacing $d = 5.5cm$. The type of microphone used is a CZN-15E Omnidirectional Microphone with a self-made amplifier to magnify the output. Since the sound source can appear anywhere, an omnidirectional microphone is handy because it can capture sound uniformly, regardless of the direction. The signal was sampled and processed, using an STM32F103c8t6 microcontroller (ARM Cortex-M3 32-bit RISC core operating at a $72MHz$ frequency), which can also connect to a computer through UART communication for testing and analysis. An LED-ring module at the center of the microphone array will indicate the DOA estimation when the device works independently.

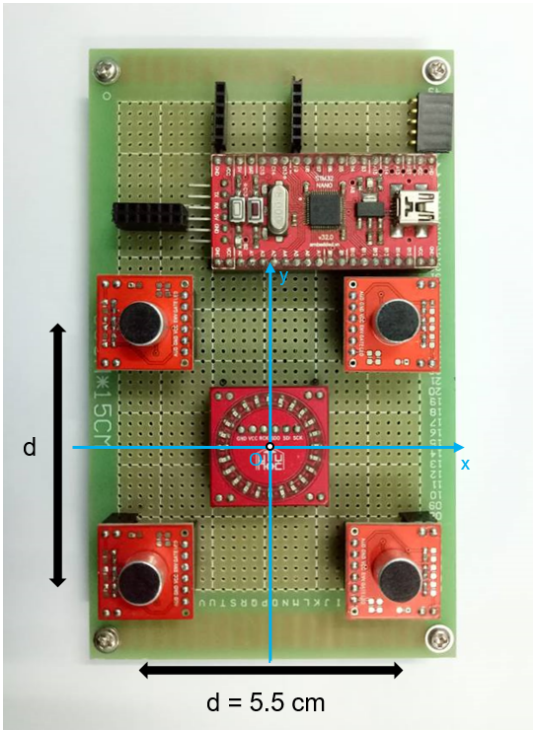


Fig. 4: The architecture of the capturing and processing device.

The capture signals were sampled at $f_s = 16kHz$ at 16-bit using the STM32's A/D converters. The frame size was set to $N = 256$, with 50% overlap between the frames. The speed of sound was set to 343.2 m/s. Due to the spacing of $d = 5.5cm$, these signals were passed through a low-pass filter with a cutoff frequency of $3120Hz$ to prevent the spatial aliasing. Additionally, to minimize DC bias and spectral leakage effect, a processing of DC Off-set Removals with Hann window was applied. Then, these signals were immediately processed by our proposed method. By incorporating some hardware-friendly algorithms such as CORDIC and some data management, our microcontroller was able to handle all the calculations and provide the real-time results.

4.1. Evaluation method

In this sub-section, several experiments were conducted to assess the device's ability to rapidly and accurately estimate multiple DOAs. All experiments were conducted in a 3.3m x 3m x 3m room, a reverberation time of $RT_{60} = 0.98ms$, background noise was introduced to simulate real-life scenarios. In each experiment, the number of sound sources varied from 1 to 4, with the sources playing simultaneously from different locations within a radius of $r = 0.4$m at the azimuth plane.

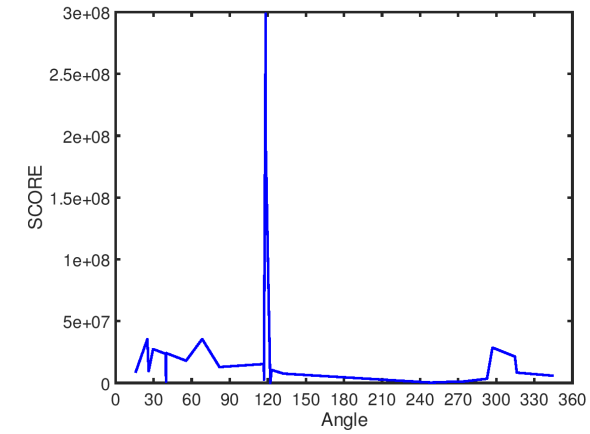
1) Ability examination of the proposed approach

In the first test, we investigated the ability to simultaneously identify multiple sound sources in the frequency domain. The test uses 1 to 3 sine wave sound sources, with different frequencies transmitted by loudspeakers at different room positions. The sound sources were positioned in the room at angles $\phi_1 = 120°$, $\phi_2 = 280°$, $\phi_3 = 0°$, with the frequencies at $1250Hz$, $625Hz$, and $1800Hz$, respectively. The results obtained at a time frame are extracted in Fig. 5 for evaluation.

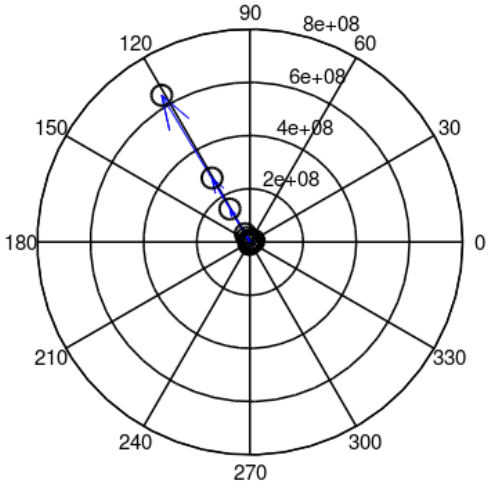
Next, another experiment was also performed to examine the device's ability to track a moving sound source at a specific frequency. A loudspeaker was moved around the device in a full circle, starting from $\phi = 0°$ and completing the circle in 20 seconds while broadcasting sine wave signal at $875Hz$. The actual movement of the sound source and the trajectory estimated by the device are shown in Fig. 6.

2) Localization with multiple static speech sources

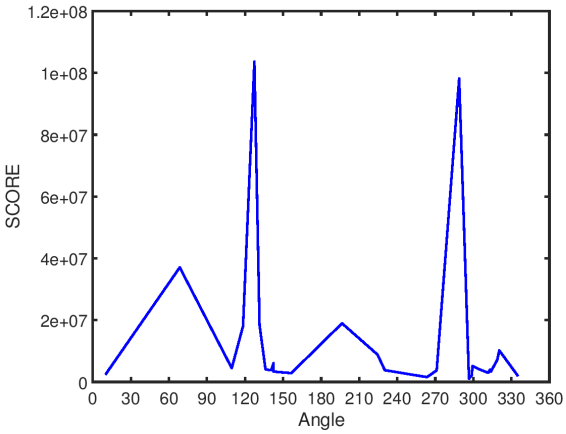
We evaluated the device's performance with multiple speech sources using four speech sources played by loudspeakers located at azimuths $\phi_1 = 30°$, $\phi_2 = 250°$, $\phi_3 = 130°$, and $\phi_4 = 310°$, as shown in Fig. 8. For the first session, only source 1 was played; the second one involves sources 1 and 2 played simultaneously, and so on. All the tests were carried out in 30 seconds. Since the potential source is rated by the SCORE value, all potential sources are plotted with the SCORE intensity on the decibel scale. Fig. 7 shows the visualization of the result and the potential DOA's SCORE distribution in each case. However, these cannot be considered as



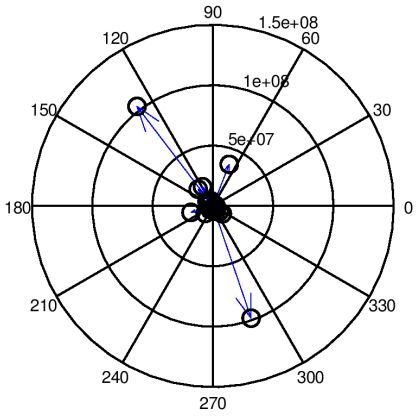
(a)



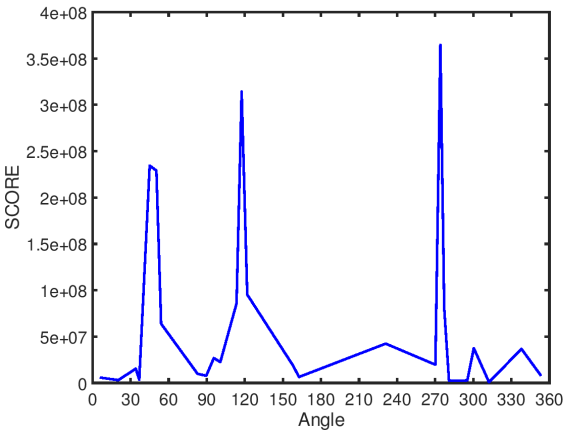
(b)



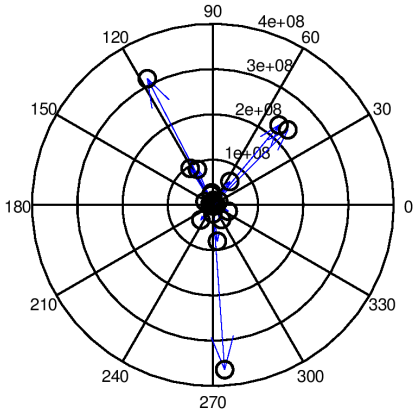
(c)



(d)



(e)



(f)

Fig. 5: Survey results with sound sources: (a)(b) One sound source; (c)(d) Two sound sources; and (e)(f) Three sound sources.

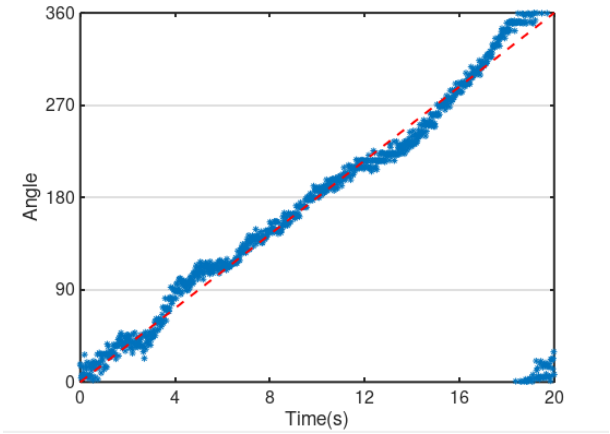


Fig. 6: Trace investigation of the sound source is a sine signal.

the final result since this is just raw data that requires a threshold to filter out the undesirable results, and we need some metrics to evaluate the device's performance. The device's recovery capabilities and accuracy were measured as follows because of the unique way of the estimating DOA.

If a potential source's DOA is estimated within a range of $\pm 15^\circ$ of an actual source in a single time frame, it is considered a real positive. If a source is estimated outside of that range, it is considered a false positive. If an actual sound source is not estimated at the recent time frame, it is considered a false negative. Using these metrics, the precision and recall rates with different thresholds ranging from 0 to the maximum SCORE value were calculated. Fig. 9 shows the Precision-Recall Curves (PRC) used to assess the device performance in each case.

There is a trade-off between precision and recall, and we do not prioritize any parameters but need to flexibly calculate the suitability for each case to achieve the goal of obtaining results as accurately as possible. To find the threshold that produces the most optimal results, in terms of accuracy and responsiveness, we used the balanced harmonic mean of precision and recall, known as F1, as a measurement. Fig. 10 shows the relationship between F1 scores and thresholds. An ideal threshold would maximize F1 and gave us the optimum precision and recall. Then, the Mean Absolute Error (MAE) is calculated for every accurately optimistic prediction from the actual DOA. These metrics are presented in Table 1 for every sound source in each case.

3) Localization with multiple mobile speech sources

The similar setup and performance metrics from localization with multiple static speech sources are applied in this case. However, at this time, the sound source will be moving around with different trajectories. In the first scenario, source 1 with $\phi_1 = 160^\circ \rightarrow 320^\circ$, the second one involving 2 sound sources, with $\phi_1 = 70^\circ \rightarrow 250^\circ$ and $\phi_2 = 290^\circ \rightarrow 110^\circ$, and the third one is 3 sources, with $\phi_1 = 80^\circ \rightarrow 170^\circ$, $\phi_2 = 130^\circ \rightarrow 220^\circ$, and $\phi_3 = 180^\circ \rightarrow 0^\circ \rightarrow 280^\circ$ (Fig. 11). Fig. 12 shows the result for each considered case. Indeed, with the same threshold as in the previous 'Localization with multiple static speech sources' scenario, we evaluate the performance of the device through Table 2, with the same metrics F1, precision, recall and average error from the reference of the expected trajectories, instead of a static value.

4) Computational load

The implementation of the proposed method resulted in a 30KB program using a 14KB of data and BSS, which is a total of 44KB of memory occupied in an STM32F103c8t6 microcontroller. Additionally, when active, the device took an average of 20ms to provided results, this includes 16ms of data sampling process and 4ms to execute all the calculations.

5) Comparison with other methods

To provide context for these results, Table 3 presents a comparison between our proposed approach and other state-of-the-art methods in terms of hardware usage and performance. Given the uniqueness of our implementation, we selected studies that provide sufficient detail in both hardware and performance aspects as references. One such study is the Steered Response Power with Phase Transform (SRP-PHAT), which employs a scanning approach enhanced by the Hierarchical Search technique with a Directivity model and Automatic calibration (HSDA) [26]. Another method is the Intensity Difference approach, implemented on a bio-bot equipped with lightweight capture hardware designed for the bot's constraints. However, the data processing for this method still relies on a laptop [2]. Additionally, the Circular Integrated Cross Spectrum (CICS) method, which incorporates a time-frequency (TF) assumption, was included due to its similarity in leveraging sparsity principles [35]. Finally, two variations of the MUSIC algorithm, GVSD and GEVD, combined with Hierarchical SSL (H-SSL) and the Minimum Akaike Information Criterion Estimate (MAICE), were selected for their demonstrated effectiveness in handling multiple source sound source localization (SSL) scenarios [7, 8].

5. Results and Discussion

Firstly, as can be expected, the performance of the proposed device decreases as more sources appear in the environment, which can be expected. However, increas-

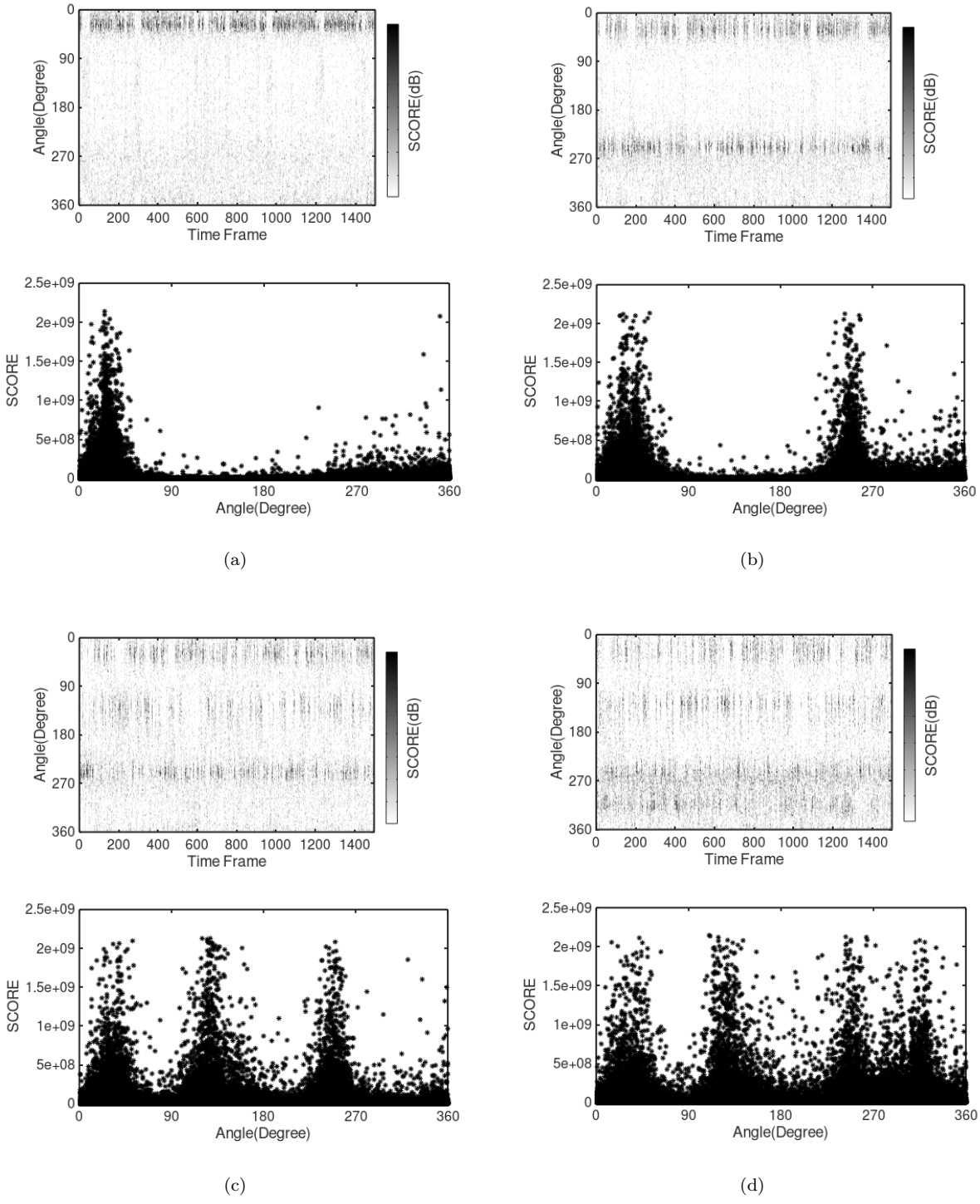


Fig. 7: Potential DOA distribution with static sources: (a) A static source; (b) Two static sources; (c) Three static sources; and (d) Four static sources.

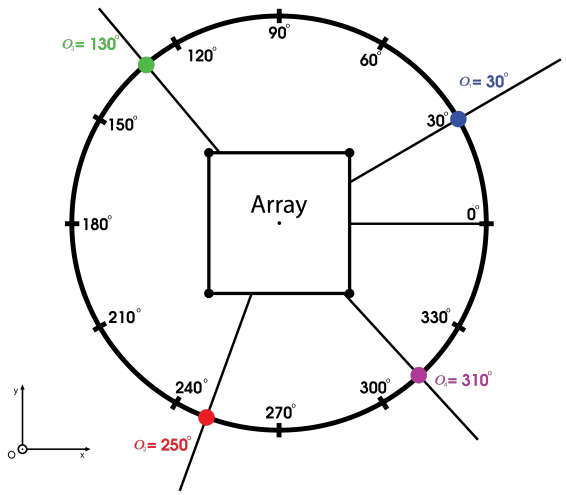


Fig. 8: Setup for static sources.

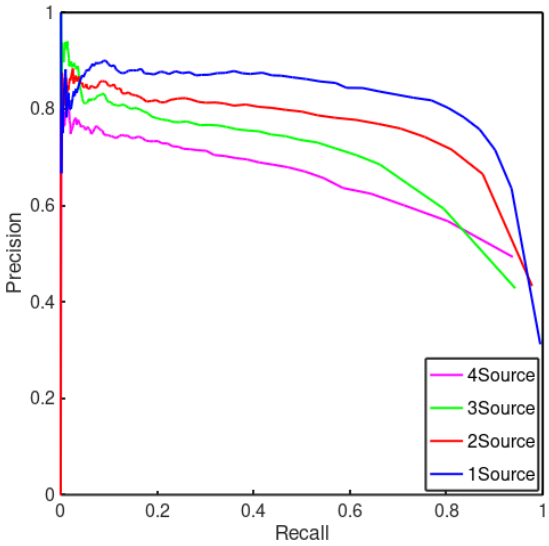


Fig. 9: Precision-Recall Curves of static sound source scenarios.

ing the number of sources does not affect the precision and recall equally. More specifically, the precision rate is not as significantly impacted, as the drop-down rate is much slower. Additionally, the PRC for cases 1 and 2 show an almost horizontal line and only curve downwards at the end. This implies that when there are one or two sound sources in the environment, the device's precision is high and stable. Also, Fig. 7 shows that the distribution peak is narrow, resulting in a small Average Error in most case ($\leq 7°$) (Table 1,2).

In contrast, the recall drops at a higher rate than precision. One explanation for this is observation sparsity, which means that the speech sound sources have to share the potential sources, and the estimation at each frequency bin can not provide multiple values. We can see in Fig. 7 and Fig. 12, when more sound sources

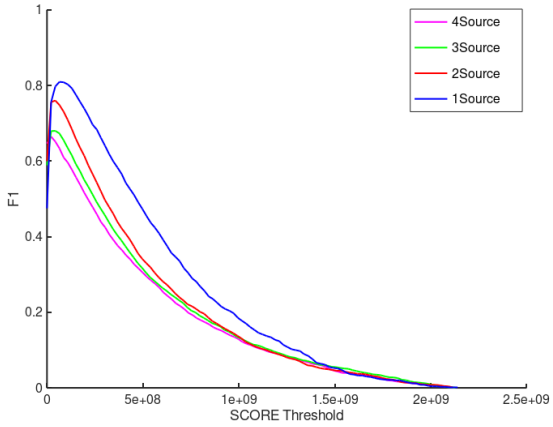


Fig. 10: F1 and threshold correlation.

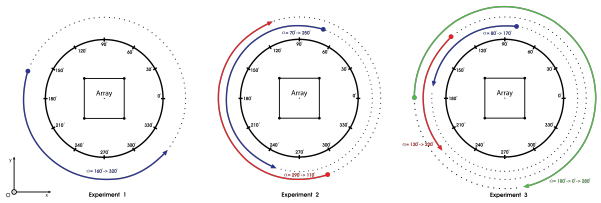


Fig. 11: Setup for mobile sources.

were present in the environment, the signal started to become less dense and detached in the time domain, and causing the estimation to tend towards other sources. This leads to an increase in false negatives, resulting in a lower recall rate. However, due to the short interval time between each result, which is only 20ms, we can still observe the continuity of the sources.

Furthermore, when the sources began to move around with different trajectories, both parallel and crossing each other, the differences in metric values can be seen as insignificant (Table 2). This means that the movement of the sources does not significantly affect the system performance. In conclusion, the proposed approach can carry out real-time multi-DOA estimations in a realistic environment, using a microcontroller with fair accuracy.

Other methods don't provide detail to make a direct side-by-side technical comparison. However, with general benchmark, there are 5 criteria can be assessed: hardware requirements, processing time, computational load, memory usage, algorithmic simplicity. As compared to the other research methods, as we can see in Table 3, the medium and heavy hardware methods [7, 8, 26] are the ones that provide the accurate and quick results. Lighter hardware methods [2] tend to perform worse than the others. This pattern has been around for quite a long time in this particular field. Our proposed approach is an exception, being the fastest in terms of processing time, requiring only an

Tab. 1: Device performance with static sources.

Case	Sound Scource	F1	Precision	Recall	Average Error (degree)
1	1	0.81	0.76	0.87	5.46
2	1	0.75	0.74	0.75	6.22
	2				5.18
3	1	0.67	0.68	0.66	6.23
	2				5.79
	3				6.43
4	1	0.63	0.62	0.64	7.02
	2				6.30
	3				6.85
	4				6.93

Tab. 2: Device performance with mobile sources.

Case	Sound Scource	F1	Precision	Recall	Average Error (degree)
1	1	0.82	0.78	0.86	6.34
2	1	0.73	0.74	0.71	6.73
	2				6.73
3	1	0.69	0.70	0.68	6.21
	2				6.58
	3				6.96

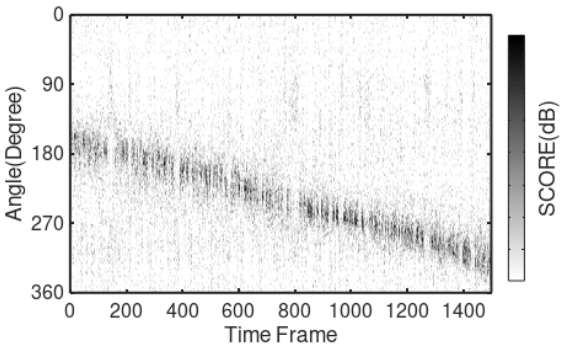
Tab. 3: State-of-the-art method in comparison.

Method	Hardware		Performance	
	Capture	Processor	Accuracy	Time
SRP-PHAT -HSDA[26] (2019)	Medium-Heavy (8-16mic)	Light-Medium (12-42% In compare with Raspberry Pi3 single CPU core)	5 sources, Medium-High	Not mentioned
Intensity Difference[2] (2015)	Light (3 unidirectional mic (PUM-3046LR))	Heavy (Matlab on a Laptop)	1source, Low (27.3°)	720ms
CICS + TF sparsity[35] (2013)	Medium (8mic) (Shure SM93)	Heavy (Standard PC, 2.4Ghz CPU, 2GB RAM)	High ($\leq$ 2.5°)	Not clearly mentioned
MUSIC + MAICE[8] (2012)	Medium (8mic)	Heavy (FPGA-Virtex4)	3sources, High (1°-5°)	22ms
GSVD-MUSIC + H-SSL[7] (2012)	Medium (8mic in circular array)	Heavy (Laptop, Core i7 2Ghz CPU, 8GB SDRAM)	Medium-High ($\leq$ 10°)	GVSD MUSIC: 5.52ms H-SSL: 18.4-20.8ms
Our propose	Light (4 omnidirectional mic (CZN-15E))	Light (44KB program in STM32F103)	4 sources, Medium-High($\leq$ 7°)	20ms

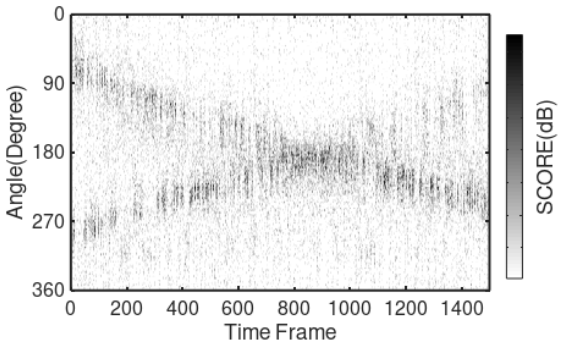
average of 20ms for capturing, sampling (16ms), and calculations (4ms). It also uses the lightest hardware, occupying only 44KB of memory in an STM32F103 microcontroller. Moreover, it employs a small number of microphones (only four) while maintaining medium to high accuracy.

It is important to note that among all of these above approaches, the designers will select the appropriate ones, depending on the final product's practical application. For example, in a search and rescue mission where a swarm of biobots uses sound to navigate and locate survivors [2], highly precise results may not be necessary. Instead, a compact capturing device is essential to allow the robot to carry it while exploring deep inside rubble. Our proposed approach has demonstrated its performance capabilities and provides flexibility, allowing the precision and recall rates to be focused on separately depending on the goals. This is why its potential is significant, with room for further development in various applications.

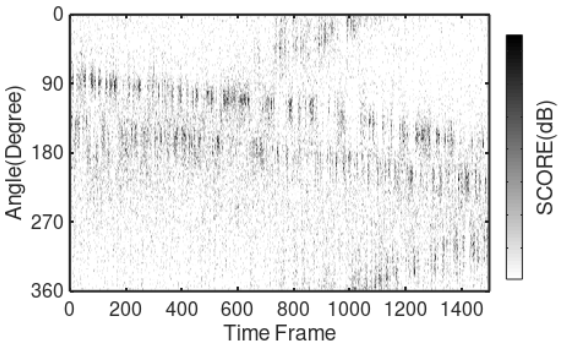
One ongoing challenge with this work is that the sources abruptly discontinue when three or more sound sources are present in the environment. We observed that this is due to observation sparsity and can potentially be addressed by increasing the density of frequency bins. In future work, the proposed method could incorporate a standard and lightweight tracking method such as the Kalman filtering technique to track the movement of sources. Additionally, this approach could be extended to create a real-time wireless sensor network capable of tracking movement through sound. Furthermore, the detection plane of the method can be expanded to both azimuth and elevation planes. Research could also focus on source separation methods to create a comprehensive speech recognition module for broader applications.



(a)



(b)



(c)

Fig. 12: Potential DOA distribution with mobile sources: (a) One mobile source; (b) Two mobile sources; and (c) Three mobile sources.

In this study, the impact factors such as environmental noise, limited microphone array size, and simplification of SCORE data can reduce the performance and accuracy of the device. Some solutions to improve the performance will be further studied such as: applying advanced filtering techniques, larger microphone arrays, multi-frequency integration, machine learning algorithms, and source tracking methods [36], [37], [38].

6. Conclusion

This paper introduced a novel method for estimating the directions of multiple sound sources based on three primary techniques: beamforming, time difference of arrival (TDOA), and frequency sparsity. The results of this new method are presented through a novel computing system that we have proposed and fully demonstrated using mathematical theory and signal processing techniques. The effectiveness of frequency-domain beamforming in sound source localization (SSL) has been established. Additionally, by leveraging a symmetrical geometry array, we reduced the complexity of the entire method. We also implemented the proposed approach on a compact device using the popular ARM STM32 microcontroller, ensuring resource efficiency and real-time performance. Preliminary surveys and evaluation results have provided evidence of the method's accuracy. Despite limitations in our testing equipment, which involved a high amount of substrate and low-quality components, the received signal contained significant noise, further confirming the method's robustness. These promising results indicate that the proposed method can be deployed in real-time IoT systems (as well as edge computing equipments), information acquisition, robotics, and location-based applications.

Author Contributions

N. T. H. made a complete contribution to this paper by developing the theoretical formalism, performing the analytic calculations, conducting the numerical simulations, synthesizing the results, and writing the first draft. K. Y. provided comments on the layout and edited some of the explanations.

References

[1] HWANG, S., Y. PARK, Y. S. PARK. Sound direction estimation using an artificial ear for robots. *Robotics and Autonomous Systems.* 2011, vol. 59, iss. 3-4, pp. 208–217. DOI: 10.1016/j.robot.2010.12.005.

[2] LATIF, T., E. WHITMIRE, T. NOVAK, A. BOZKURT. Sound localization sensors for search and rescue biobots. *IEEE Sensors Jour-*

nal. 2016, vol. 16, no. 10, pp. 3444–3453. DOI: 10.1109/JSEN.2015.2477443.

[3] DENG, F., S. GUAN, X. YUE, X. GU, J. CHEN, L. JIANYAO, J. LI. Energy-Based Sound Source Localization with Low Power Consumption in Wireless Sensor Networks. *IEEE Transactions on Industrial Electronics*. 2017, vol. 64, no. 6, pp. 4894–4902. DOI: 10.1109/TIE.2017.2652394.

[4] SCHMIDT, R. Energy-Based Sound Source Localization with Low Power Consumption in Wireless Sensor Networks. *IEEE Transactions on Antennas and Propagation*. 1986, vol. 34, no. 3, pp. 276–280. DOI: 10.1109/TAP.1986.1143830.

[5] ASONO, F., H. ASOH, T. MATSUI. Sound source localization and signal separation for office robot jijo-2. *International Conference on Multisensor Fusion and Integration for Intelligent Systems. MFI'99 (Cat. No.99TH8480), Taipei, Taiwan*. 1999, pp. 243–248. DOI: 10.1109/MFI.1999.815997.

[6] NAKAMURA, K., K. NAKADAI, F. ASANO, G. INCE. Intelligent Sound Source Localization and its application to multimodal human tracking. *IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA*. 2011, pp. 143–148. DOI: 10.1109/IROS.2011.6094558.

[7] NAKAMURA, K., K. NAKADAI, G. INCE. Real-time super-resolution Sound Source Localization for robots. *IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal*. 2012, pp. 694–699. DOI: 10.1109/IROS.2012.6385494.

[8] LUNATI, V., J. MANHES, P. DANES. A versatile System-on-a-Programmable-Chip for array processing and binaural robot audition. *IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal*. 2012, pp. 998–1003. DOI: 10.1109/IROS.2012.6386144.

[9] KNAPP, C., G. CARTER. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1976, vol. 24, no. 4, pp. 320–327. DOI: 10.1109/TASSP.1976.1162830.

[10] ARGENTIERI, S., P. DANES, P. SOUERES. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech and Language*. 2015, vol. 34, iss. 1, pp. 87–112. DOI: 10.1016/j.csl.2015.03.003.

[11] LIU, H., Z. FU, X. LI. A two-layer probabilistic model based on time-delay compensation for binaural sound localization. *IEEE International Conference on Robotics and Automation, Karlsruhe, Germany*. 2013, pp. 2705–2712. DOI: 10.1109/ICRA.2013.6630949.

[12] GRONDIN, F., F. MICHAUD. Noise mask for TDOA sound source localization of speech on mobile robots in noisy environments. *IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden*. 2016, pp. 4530–4535. DOI: 10.1109/ICRA.2016.7487652.

[13] RASCON, C., G. FUENTES, I. MEZA. Lightweight multi-DOA tracking of mobile speech sources. *EURASIP Journal on Audio, Speech, and Music Processing*. 2015, vol. 11, pp. 1-16. DOI: 10.1186/s13636-015-0055-8.

[14] CHIARIOTTI, P., M. MARTARELLI, P. CASTELLINI. Acoustic beamforming for noise source localization–Reviews, methodology and applications. *Mechanical Systems and Signal Processing*. 2019, vol. 120, pp. 422–448. DOI: 10.1016/j.ymssp.2018.09.019.

[15] LI, Y., H. MA, D. YU, L. CHENG. Iterative robust Capon beamforming. *Signal Processing*. 2016, vol. 118, pp. 211–220. DOI: 10.1016/j.sigpro.2015.07.004.

[16] DOUGHERTY, R. P. Functional beamforming for aeroacoustic source distributions. *20th AIAA/CEAS aeroacoustics conference*. 2014. DOI: 10.2514/6.2014-3066.

[17] SARRADJ, E. A fast signal subspace approach for the determination of absolute levels from phased microphone array measurements. *Journal of Sound and Vibration*. 2010, vol. 329, iss. 9, pp. 1553–1569. DOI: 10.1016/j.jsv.2009.11.009.

[18] SIJTSMA, P. Clean based on spatial source coherence. *International journal of aeroacoustics*. 2007, vol. 6, iss. 4, pp. 357–374. DOI: 10.1260/147547207783359459.

[19] BROOKS, T. F., W. M. HUMPHREYS. A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays. *Journal of sound and vibration*. 2006, vol. 294, iss. 4–5, pp. 856–879. DOI: 10.1016/j.jsv.2005.12.046.

[20] MALGOEZAR, A. M. N., M. SNELLEN, R. MERINO-MARTINEZ, D. G. SIMONS, P. SIJTSMA. On the use of global optimization methods for acoustic source mapping. *The Journal of the Acoustical Society of America*. 2017, vol. 141, iss. 1, pp. 453–465. DOI: 10.1121/1.4973915.

[21] HALD, J. Basic theory and properties of statistically optimized near-field acoustical holography. *The Journal of the Acoustical Society of America*. 2009, vol. 125, iss. 4, pp. 2105–2120. DOI: 10.1121/1.3079773.

[22] ANTONI, J. A bayesian approach to sound source reconstruction: Optimal basis, regularization, and focusing. *The Journal of the Acoustical Society of America*. 2012, vol. 131, iss. 4, pp. 2873–2890. DOI: 10.1121/1.3685484.

[23] PEREIRA, A., J. ANTONI and Q. LECLERE. Empirical Bayesian regularization of the inverse acoustic problem. *Applied Acoustics*. 2015, vol. 97, pp. 11–29. DOI: 10.1016/j.apacoust.2015.03.008.

[24] LECLERE, Q., A. PEREIRA, C. BAILLY, J. ANTONI, C. PICARD. A unified formalism for acoustic imaging techniques: illustrations in the frame of a didactic numerical benchmark. *the 6th Berlin Beamforming Conference*. 2016, pp. 1–17.

[25] SUZUKI, T. L1 generalized inverse beam-forming algorithm resolving coherent/incoherent, distributed and multipole sources. *Journal of Sound and Vibration*. 2011, vol. 330, iss. 24, pp. 5835–5851. DOI: 10.1016/j.jsv.2011.05.021.

[26] GRONDIN, F., F. MICHAUD. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robotics and Autonomous Systems*. 2019, vol. 113, pp. 63–80. DOI: 10.1016/j.robot.2019.01.002.

[27] RASCON, C., I. MEZA. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*. 2017, vol. 96 pp. 184–210. DOI: 10.1016/j.robot.2017.07.011.

[28] BOFILL, P., M. ZIBULEVSKY. Underdetermined blind source separation using sparse representations. *Signal processing*. 2001, vol. 81, iss. 11, pp. 2353–2362. DOI: 10.1016/S0165-1684(01)00120-7.

[29] SADHU, A., S. NARASIMHAN, J ANTONI. A review of output-only structural mode identification literature employing blind source separation methods. *Mechanical Systems and Signal Processing*. 2017, vol. 94, pp. 415–431. DOI: 10.1016/j.ymssp.2017.03.001.

[30] ZHANG, W., B. D. RAO. A two microphone-based approach for source localization of multiple speech sources. *IEEE Transactions on Audio, Speech, and Language Processing*. 2010, vol. 18, iss. 8, pp. 1913–1928. DOI: 10.1109/TASL.2010.2040525.

[31] KIM, U. H., K. NAKADAI, H. G OKUNO. Improved sound source localization in horizontal plane for binaural robot Audition. *Applied Intelligence*. 2015, vol. 42, iss. 1, pp. 63–74. DOI: 10.1007/s10489-014-0544-y.

[32] DANES, P., J. BONNAL. Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme. *IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan*. 2010, pp. 1976–1981. DOI: 10.1109/IROS.2010.5651249.

[33] PAVLIDI, D., M. PUIGT, A. GRIFFIN, A. MOUCHTARIS. Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan*. 2012, pp. 2625–2628. DOI: 10.1109/ICASSP.2012.6288455.

[34] REDDY, A. M., B. RAJ. Soft Mask Methods for Single-Channel Speaker Separation. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007, vol. 15, no. 6, pp. 1766–1776. DOI: 10.1109/TASL.2007.901310.

[35] PAVLIDI, D., A. GRIFFIN, M. PUIGT, A. MOUCHTARIS. Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array. *IEEE Transactions on Audio, Speech, and Language Processing*. 2013, vol. 21, no. 10, pp. 2193–2206. DOI: 10.1109/TASL.2013.2272524.

[36] SVOBODOVA, H., E. VAVRINSKY, D. TURONOVA, M. DONOVAL, M. DARICEK, P. TELEK, M. KOPANI. Optimization of the position of single-lead wireless sensor with low electrodes separation distance for ECG-derived respiration. *Advances in Electrical and Electronic Engineering*. 2018, vol. 16, no. 4. DOI: 10.15598/aeee.v16i4.2773.

[37] DE RANGO, F., N. PALMIERI, S. RANIERI. Spatial correlation based low energy aware clustering (leach) in a wireless sensor networks. *Advances in Electrical and Electronic Engineering*. 2015, vol. 13, no. 4. DOI: 10.15598/aeee.v13i4.1496.

[38] CHEN, L., G. CHEN, L. HUANG, Y. CHOY, W. SUN. Multiple Sound Source Localization, Separation, and Reconstruction by Microphone Array: A DNN-Based Approach. *Applied Sciences*. 2022, vol. 12, no. 7. DOI: 10.3390/app12073428.