MASARYK
UNIVERSITY

FACULTY OF INFORMATICS

Sound source localization from a microphone array

Bachelor's Thesis

JAKUB PEKÁR

Brno, Spring 2023

MASARYK UNIVERSITY

FACULTY OF INFORMATICS

Sound source localization from a microphone array

Bachelor's Thesis

JAKUB PEKÁR

Advisor: doc. RNDr. Petr Matula, Ph.D.

Department of Visual Computing

Brno, Spring 2023


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Jakub Pekár

Advisor: doc. RNDr. Petr Matula, Ph.D.

Acknowledgements

I am deeply grateful to my advisor, doc. RNDr. Petr Matula, Ph.D., for his unwavering support and guidance throughout the entire duration of my research. His constructive feedback has played a pivotal role in shaping the quality and refinement of this thesis. I would also like to extend my gratitude to RNDr. Miloš Liška, Ph.D., for introducing me to the field of sound propagation and acquisition. I also thank OpenAI for developing the ChatGPT application, which helped me enhance my thesis's English. Lastly, I would like to acknowledge Šimon Čecháček for generously lending me the necessary microphones for the acquisition of real-world data.

Abstract

This thesis investigates the problem of indoor speaker localization, particularly concentrating on the localization of speaking students within a classroom setting, a specialized scenario in the expansive field of sound source localization (SSL). A spectrum of SSL methods was scrutinized, encompassing traditional algorithms and machine learning techniques, underpinned by a comprehensive theoretical exploration of SSL. A bespoke dataset was created due to the lack of an apt public dataset featuring recordings from four diverse rooms utilizing a circular microphone array with four uniformly spaced microphones. Furthermore, two potential microphone placements were assessed for localizing speaking students within a classroom setting. The efficacy of each method was gauged using both synthetic and real-world datasets, considering both proposed microphone arrangements. Despite none of the evaluated methods achieved localisation performance sufficient for localising students within classroom, two conventional methods, Steered-Response Power Phase Transform and Generalized Cross-Correlation Phase Transform with multilateration via Particle Swarm Optimization, demonstrated notable localization accuracy using ceiling-attached microphones on real-world evaluation data. The thesis concludes by elucidating the constraints of the tested methods, emphasizing those necessitating training, and discusses the potential influence of discrepancies between real-world and synthetic data on model performance.

Keywords

Sound Source Localization, Acoustic Source Localization, Time Difference Of Arrival, Neural Networks, 3D Localization, Indoor Speaker Localization

Contents

List of Tables

List of Figures

1 Introduction

Source localization is a critical task with widespread applications that significantly influence various aspects of daily life. Source localization techniques are implemented across diverse domains, including global positioning systems (GPS) [1], phone localization [2], sonar [3], and surveillance [4]. Remarkably, numerous living organisms, including humans, exhibit the innate ability to estimate the position of proximate auditory stimuli [5]. Sound source localisation (SSL) refers to the problem of identifying the spatial coordinates of an acoustic-emitting object within a specified environment through the utilization of sound receivers, such as microphones. The estimated source location is determined in relation to one or multiple sensing devices.

Nevertheless, the term "source location" within the context of SSL is inherently ambiguous, potentially denoting the source's position in a given space or solely the relative direction towards the acoustic source. While the estimation of the direction of arrival (DOA) is sufficient for specific applications, including videoconferencing [6], comprehensive localization is indispensable for others, such as systems facilitating human-device interaction, like home assistants [7], which enable advanced interactions between the system and human [8]. In practical SSL implementations, detrimental effects, including noise and reverberation, can diminish performance, particularly in indoor environments.

Despite the extensive investigation into SSL by the scientific community over several decades, achieving complete 3D localization remains a formidable challenge. The advent of machine learning (ML) and deep neural networks (DNN) has spurred renewed interest in SSL research in recent years. ML methodologies have surpassed traditional algorithms in a wide array of related and disparate fields, thereby encouraging incorporation of neural networks into contemporary SSL algorithms [9]. However, even with their successes, a majority of ML-based SSL techniques continue to rely on conventional algorithms for feature extraction.

In this thesis, I focus on indoor speaker localization (ISL) in the particular setup of localizing speaking pupils within an educational setting, specifically a classroom environment. The primary objective of

this research is to determine the most promising method for localizing speaking pupils during classes. This investigation was initiated by researchers at the Department of Educational Sciences, Faculty of Arts, Masaryk University, who necessitate a tool to quantify pupils' engagement during lectures. In this context, an active student is an individual who actively participates in dialogues with the lecturer.

The problem at hand can be conceptualized as an SSL problem, as I assume that the most prominent sound in the classroom emanates from either the speaking student or the lecturer. I further assume that there is no overlap between the speakers, ensuring only a single dominant sound source exists at any moment.

In order to ascertain the most appropriate technique for localizing active pupils, various methods, such as time of arrival and machine learning-based approaches, are examined. I will assess the performance of each method on a custom dataset and choose the technique yielding the highest accuracy in localizing speaking students. It is imperative to acknowledge that the classroom's acoustics may influence the effectiveness of the selected method; thus, factors such as room dimensions and ambient noise levels must be considered during the selection and evaluation process.

The organization of this thesis is as follows: post the theoretical background, where the focus is placed on signal characteristics within indoor environments, and a review of selected published SSL methods, two suitable microphone arrangements are proposed for the localization of speaking pupils within an educational setting. Subsequently, the creation process of synthetic and real-world datasets to evaluate various localization methods' performance is delineated. Then I describe the implementation details of the chosen methods. Section 4 presents the localization results considering two viable microphone arrangements. The results are then discussed and interpreted. Finally, conclusions are drawn in Section 6.

2 Theoretical Background

Various general methods are employed to estimate the location of acoustic sources, leveraging the distinct characteristics of sound propagation in specific environments [10, 5, 11]. In general, the problem of SSL is not limited to any particular environment. Nevertheless, the medium through which the acoustic wave propagates significantly influences the propagation properties of sound [11, 12, 13, 14, 15]. While most of the existing SSL research concentrates on air as the transition medium [16, 17, 9], other environments, such as water, have been considered in some applications [11, 18]. In the air, the predominant techniques employed for SSL are grounded in the fundamental sound propagation properties described below.

This section first familiarizes the reader with the theoretical background of sound propagation, subsequently delineating general strategies for the SSL task. An exposition of tools apt for simulating sound propagation complements this. Finally, a collection of SSL methods spanning a broad spectrum of general approaches is elucidated.

2.1 Principal Properties of Sound Propagation

The essential properties governing sound propagation and its interaction with the environment can be classified into two categories. The first category comprises the sound signal's inherent properties, such as frequency, wavelength, and amplitude [19]. The second category includes the properties of the environment through which the sound travels, encompassing the speed of sound propagation, sound absorption, reflection, and diffraction [20, 21]. The foundations of SSL methods primarily revolve around two critical aspects: the speed of sound and its energy attenuation [22].

2.1.1 Speed of Sound

A primary property of sound propagation in air, which most SSL methods rely on, is the speed of sound [9]. This speed is influenced by the transmission medium and factors such as temperature and pressure [12, 14, 23]. Within the context of indoor localization, standard

temperature, and pressure conditions are typically assumed [10, 5], resulting in an estimated speed of sound of 343 m/s [10, 5, 14].

$$v_{\text{sound}} = 343 \text{ m/s} \tag{2.1}$$

2.1.2 Signal Energy Attenuation

The energy of an acoustic wave decreases with the increase in the distance it travels [22, 24]. This phenomena is called the attenuation of the signals energy or intensity [22], what is another crucial property of sound propagation. The intensity decay adheres to an inverse square law proportional to the distance traversed by the acoustic wave. This relationship can be formulated mathematically as [22, 24]

$$e_m(t) = g_m \frac{S(t)}{d_m^2} + \epsilon_m(t), \tag{2.2}$$

where $e_m(t)$ stands for the captured signal energy at the receiver m, t denotes the time parameter, d_m represents the distance between the source and receiver m, g_m signifies the gain factor of the receiver, $\epsilon_m(t)$ refers to the additive noise, and $S(t)$ denotes the energy of the emitted sound signal one meter from the sound source [4].

2.1.3 Signal Propagation in Indoor Environments

Despite the assumption of a direct sound propagation path from the sound-emitting object to the receivers, meaning there are no obstacles on the direct path from the source to any microphone, non-negligible disparities exist in the signals captured by spatially separated receivers. These disparities stem from the distinct path of the emitted acoustic wave toward each receiver, leading to significant variations in the captured signals [9].

The time shift of the received signal or the alteration in captured signal intensity are the distinctions of interest. However, we must consider other modifications caused by reverberation and noise in indoor environments. As the acoustic wave propagates through space, it reflects off walls and other objects, generating reverberation in the captured signal [9]. Noise may also be introduced by the receivers or processing devices, as well as other background sound emitters

such as televisions or radios. Both reverberation and noise contribute to the non-negligible differences among spatially separated receivers, resulting from the unique propagation paths.

In the context of SSL, *a priori* knowledge of the environment, signal, and noise is limited. In many cases, only the position of receivers and the type of environment are known. At the same time, other parameters such as Signal-to-Noise Ratio (SNR), the spectrum of the source signal, or reverberation are either unknown or only approximately known.

2.1.4 Acoustic Far-Field and Near-Field

The concepts of acoustic far-field and near-field pertain to the sound propagation characteristics, which are influenced by the distance between the sound source and the receiver [25, 26]. This distance impacts the behavior of the radiated acoustic energy from the sound source. The far-field region is typically considered to commence at approximately one wavelength distance from the sound source and extends infinitely [27, 26]. Typically, a distance of one meter is deemed sufficient to ensure far-field propagation for the most critical frequencies [25]. In this region, the sound pressure and acoustic particle velocity are in phase, and the sound pressure level adheres to the inverse square law [26]. The acoustic wave propagates spherically. However, when the distance between receivers is small enough in relation to the distance to the sound source, in the far-field context, the acoustic wave can be approximated as a plane wave due to the negligible curvature [26].

In contrast, when the sound source is in close proximity to the receiver, the sound energy oscillates between the vibrating surface of the source and does not propagate away from the source [25]. Increasing the distance between the source and receiver up to the far-field boundary, a portion of the sound field circulates while the rest propagates away. Consequently, no fixed relationship exists between pressure and distance in the near-field scenario [25].

2.2 Captured Signal

Consider an array comprising M sound receivers, each with an index m. The signal captured by the m-th receiver at time t can be expressed mathematically as [28, 29]:

$$s_m(t) = (h_m * s)(t - \tau_m) + n_m(t), \tag{2.3}$$

where $h_m(t)$ represents the Room Impulse Response Function (RIRF), $s(t)$ denotes the source signal, $n_m(t)$ signifies any introduced additive white noise, and τ_m indicates the delay between the emitted signal $s(t)$ and the received signal at receiver m. The operator $*$ represents convolution. The delay τ_m depends solely on the distance between the source and receiver m. As elucidated in [30], the RIRF is unique for each receiver and source position. Under ideal conditions, such as an anechoic environment that is noise-free and devoid of reverberation, the received signal s_m would be a delayed version of the emitted signal $s(t)$, with delay τ_m.

2.3 General SSL Approaches

The subsequent sections present a concise exposition of the extensively employed techniques to determine the spatial location of an acoustic source by utilizing recordings from multiple receivers [31, 4]. These techniques can be fundamentally categorized into indirect and direct approaches [32]. Indirect approaches typically adhere to a two-step process: they initially conduct measurements between pairs of microphones, subsequently estimating the source position based on the geometry of the array and the acquired measurements. In contrast, direct approaches facilitate source localization in a single, unified step.

2.3.1 Time of Arrival

One prevalent time-based localization approach is Time of Arrival (TOA). TOA estimates the location of an object by calculating the arrival time between the source and receiver, which defines the radius of a sphere [33]. The location estimate is obtained using trilateration [34], employing several TOAs from different pairs of receivers and sources.

Two different setups are supported by TOA: one that involves multiple sources and a single receiver (as in GPS) and the other that utilizes multiple receivers and a single source. However, the calculation of arrival time with TOA necessitates knowledge of the source emitting time, which is often unavailable in most SSL problems.

Time Difference of Arrival

In order to determine the position of a sound source when the emission time is unknown, the Time Difference of Arrival (TDOA) technique is commonly employed [35]. As the acoustic wave propagates through the air at a fixed speed, the signal captured by a receiver experiences a delay from the original signal, as demonstrated by equation (2.3). The delay, denoted by τ_m, depends on the distance between the sound source and the receiver. By utilizing signals captured by multiple receivers, the TDOA between a pair of receivers can be computed as:

$$\tau_{kl} = \tau_k - \tau_l = \frac{\|x_s - x_k\| - \|x_s - x_l\|}{v_{\text{sound}}}. \tag{2.4}$$

Here, τ_{kl} represents the TDOA between receivers k and l and is obtained as the difference between the arrival times at each receiver. In the equation (2.4), x_s represents the position of the sound source, and x_k and x_l denote the locations of receivers k and l, respectively. This approach allows for precise estimation of the sound source location, even when the emission time is unknown.

Multilateration

Multilateration is a widely employed technique for estimating the location of an acoustic source based on TDOA measurements [36]. TDOA between each pair of receivers defines a hyperboloid in 3-dimensional space [37]. The foci of the hyperboloid are located at the positions of the receivers, while the TDOA determines the vertex's position. The intersection of several hyperboloids identifies the location of the sound source. However, in real-world scenarios, TDOA estimates are often imprecise, resulting in hyperboloids with multiple intersections or none at all. Consequently, optimization algorithms are typically employed to estimate the location of the sound source [36]. A depiction of the multilateration in a 2-dimensional case is illustrated in Figure 2.1.

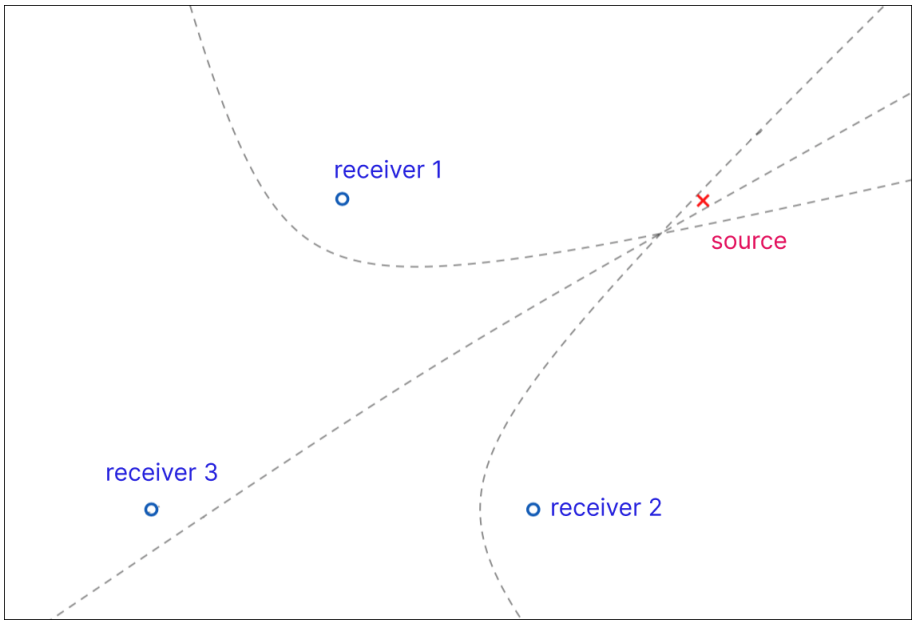

Figure 2.1: An illustration of multilateration in a 2-dimensional space, utilizing three spatially separated receivers. The TDOAs generate a set of hyperbolas. The intersection points of these hyperbolas determine the location of the source.

2.3.2 Signal Energy

The energy-based localization of acoustic sources relies on the attenuation of signal intensity, as described in Section 2.1.2. The energy ratio between a pair of receivers k and l is defined as the ratio of the captured signal energy and can be formulated as follows [24, 38]:

$$m_{kl} = \frac{g_k / g_l}{E_k / E_l} = \sqrt{\frac{\|x_s - x_k\|}{\|x_s - x_l\|}} \tag{2.5}$$

Here, E_k and E_l denote the energies captured by the receivers k and l, respectively, and g_k and g_l represent the gains of the respective receivers. Assuming that all receivers have equal gains, the energy ratio can be simplified as follows:

$$m_{kl} = \frac{E_l}{E_k} \tag{2.6}$$

The position of the source x_s is situated on a hypersphere defined as [24, 38]:

$$\|x_s - c_{kl}\|^2 = \rho_{kl}^2, \tag{2.7}$$

where c_{kl} represents the center of the hypersphere and ρ_{kl} denotes its radius. The center and radius can be calculated using the following equations [24, 38]:

$$\begin{aligned} c_{kl} &= \frac{x_k - m_{kl} x_l}{1 - m_{kl}} \\ \rho_{kl} &= \frac{\sqrt{m_{kl}} \|x_k - x_l\|}{1 - m_{kl}} \end{aligned} \tag{2.8}$$

To determine the location of the source, the intersection of hyperspheres defined by several pairs of receivers is considered. An illustration is presented in Figure 2.2. However, environmental properties such as noise and reverberation can lead to sub-optimal energy ratio estimations, necessitating the use of optimization methods [4].

2.3.3 Direction of Arrival

The Direction of Arrival (DOA) technique is widely utilized in addressing the challenge of SSL within the context of indoor environments

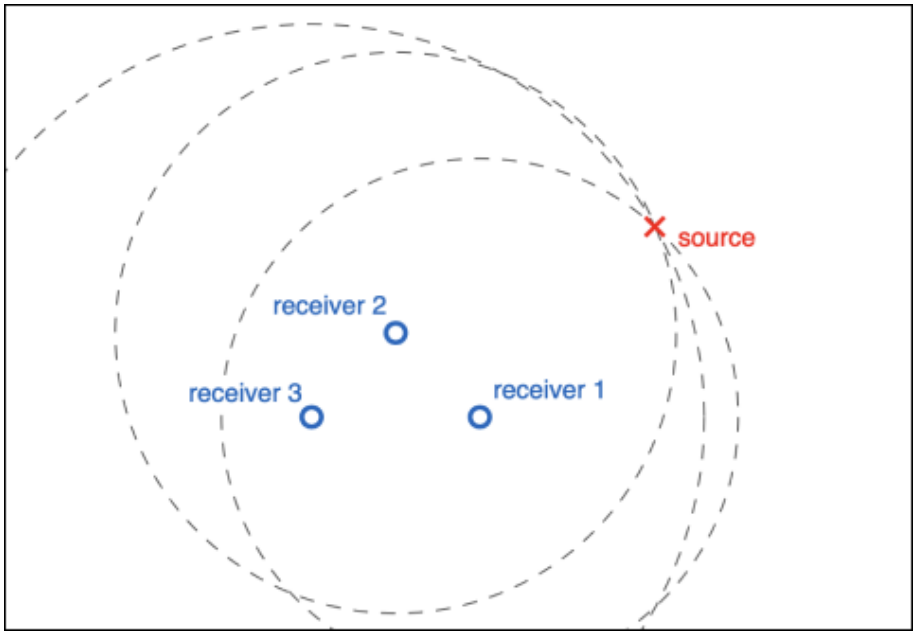

Figure 2.2: An illustration of localization using the intersection of hyperspheres in a 2-dimensional space, utilizing three spatially separated receivers.

[9, 17, 39, 40, 41, 42]. It is crucial to recognize that, in most instances, the primary focus is on calculating the direction from the receivers to the sound source(s) while neglecting the aspect of distance. This observation is substantiated by recent SSL challenges, such as the LOCATA Challenge [17] and the DCASE 2019-2022 Challenges [39, 40, 41, 42]. Consequently, DOA estimation is typically expressed in terms of azimuth (θ) and elevation (ψ) angles toward the sound source(s). The estimation of DOA is generally achieved by adopting the far-field sound propagation model, which relies on planar wave propagation as opposed to spherical wave propagation [43]. A visual representation of the far-field planar wave model is depicted in Figure 2.3.

It is essential to note that a single DOA corresponds to a line in space. The employment of triangulation techniques [44], which involve the intersection of multiple lines defined by DOAs, is instrumental in ascertaining the location of the sound source. Nevertheless, the estimation of a singular DOA necessitates the use of at least three receivers. Interestingly, mammals exhibit the capability to estimate the DOA utilizing merely two sound receivers (ears), owing to the Head Related Impulse Response (HRIR) [45]. The HRIR, respectively the Head Related Transfer Function (HRTF), derived from the Fourier Transform of HRIR [5], elucidate the alterations that sound waves experience while propagating from the sound source to the ears. These modifications are influenced by factors such as the size and shape of the head, torso, and outer ears, which are unique to each individual [45, 46].

2.3.4 Beamforming

Acoustic beamforming belongs to single-step SSL methodologies. It is a technique of acoustic imaging that generates an acoustic power map using data obtained from microphone arrays and a sound propagation model [47]. Acoustic imaging's fundamental concept is to render an acoustic representation of the environment. Beamforming utilizes the procured data and parameters from the sound propagation model, which constructs a directive "beam" of sound energy projected towards the source, thereby providing the propagated sound's direction of arrival.

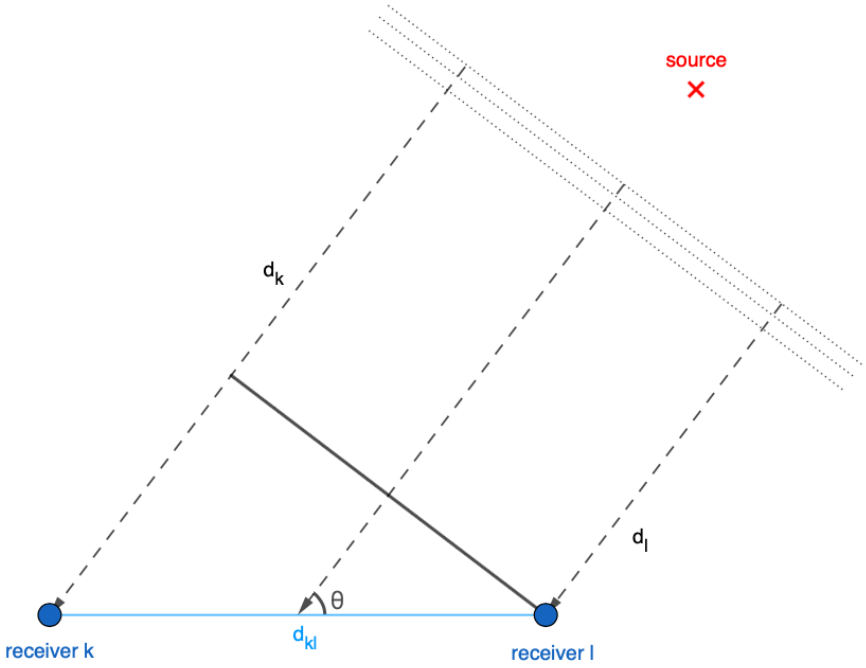

Figure 2.3: A visual representation of the far-field planar wave sound propagation model. Here, d_l and d_k denote the distances between the source and receivers k and l, respectively. The angle of arrival is represented by θ.

Generally, beamforming algorithms estimate the bearing of received signals by amplifying signals from specific directions while attenuating signals from other directions. This is accomplished by manipulating the amplitude and phase of each received signal. Despite beamforming's robustness against disturbances such as noise and reverberation, it is generally perceived that its spatial resolution is inferior [4]. Furthermore, the substantial computational demands of beamforming represent a notable limitation.

2.3.5 Neural Networks

Recent studies have witnessed a surge in interest in utilizing Neural Networks (NNs) for addressing SSL challenges [9]. NNs have demonstrated remarkable advancements over traditional algorithms across numerous research areas, including SSL. A significant proportion of contemporary methods for SSL rely on NNs, either exclusively or in conjunction with conventional techniques [9, 48, 49].

As delineated by Grumiaux *et al.* in their survey of sound source localization with deep learning methods [9], the SSL problem can be partitioned into three distinct phases: signal preprocessing, measurement computation (such as TOA or DOA), and location estimation. NNs can be implemented at any of these stages or in any combination thereof. Alternatively, an end-to-end solution can be constructed solely using NNs, wherein the raw signal from multiple receivers is input into a NN, which subsequently produces the estimated sound source location.

Convolutional Neural Networks (CNNs) are frequently employed for SSL [9]. However, other models based on self-attention mechanisms, transformers, Recurrent Neural Networks (RNNs), or Variational Autoencoders (VAEs) can also be utilized. The SSL problem mirrors trends observed in related fields concerning NNs and Deep Neural Networks (DNNs) applications, such as computer vision.

2.4 Sound Receivers

The acquisition and recording of audio signals are essential prerequisites for addressing the problem of SSL. In this context, the selection

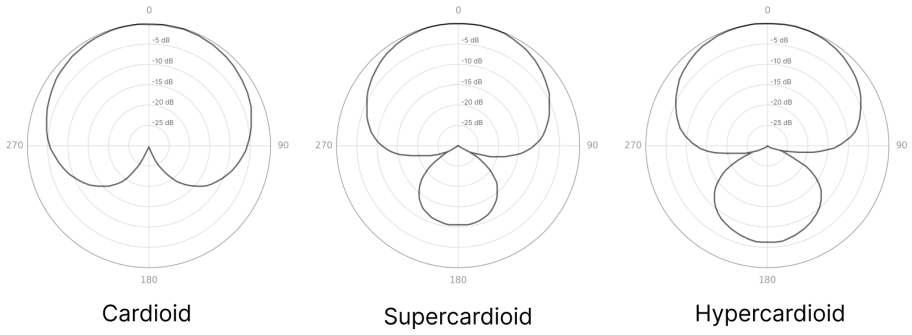

Figure 2.4: Polar charts visualizing various microphone directivity patterns.

of appropriate sound receivers is of paramount importance. In electronics, microphones serve as the primary receivers for sound waves.

Microphones are classified according to type and directivity [50]. The three main types of microphones are dynamic, condenser, and ribbon microphones, each employing a unique method for converting sound waves into electrical signals. Among these types, condenser microphones exhibit superior sensitivity, making them a favored choice for numerous SSL applications.

The directivity of a microphone pertains to its capacity to capture sounds from various directions [51]. Microphones that collect sound waves uniformly from all directions are classified as omnidirectional. Conversely, directional microphones are designed to reject incoming sounds from specific directions. Several subtypes of directional microphones exist, each differing in the degree of directivity, such as cardioid, supercardioid, and hypercardioid. Figure 2.4 visually represents these directivities.

The performance of SSL systems is contingent upon the selection of suitable microphones, as well as the number and arrangement of these devices. In the context of three-dimensional localization, a common approach involves the deployment of four microphones [5, 8]. While incorporating additional microphones may yield improved results or a more robust solution, it entails increased hardware and computational costs.

2.5 Sound Propagation Simulation

The training of a NN necessitates an extensive dataset encompassing a broad range of variations. In the context of methods incorporating NNs, there is a need for simulation of sound propagation in indoor environments, enabling the generation of synthetic data of suitable quality. Several software applications have been developed to simulate sound propagation, considering effects such as diffraction, reflection, and absorption [52, 53, 54]. These applications employ the image source method, introduced in [55], and the ray-tracing algorithm proposed in [56] to simulate sound propagation. However, not all software applications offer adequate interfaces for generating synthetic data suitable for training and evaluating SSL algorithms.

Pyroomacoustics, an open-source Python framework, provides tools for simulating sound propagation in three-dimensional models of acoustic environment [54]. It has been designed to provide accessible tools for generating synthetic audio data to train and evaluate ML models. This framework facilitates the creation of arbitrary convex rooms, with walls, floors, and ceilings composed of diverse materials to emulate the acoustic properties of realistic environments. Additionally, sound sources and microphones can be positioned arbitrarily within the room, making Pyroomacoustics a valuable asset for SSL research. Moreover, custom sound signals with the directivity of both sound sources and receivers can be implemented [54], rendering this framework an ideal solution for addressing SSL challenges.

2.6 Selected SSL Algorithms

This section presents some of commonly utilized algorithms for estimating the position of a sound source. These algorithms employ various general approaches, as discussed in Section 2.3. Acknowledging that the environment and source are unlikely to remain stationary is crucial. Hence, these methods must accommodate a finite observation time denoted by T, typically lasting hundreds of milliseconds. A summary of the selected algorithms is provided at the end of this section for quick reference.

2.6.1 Cross-Correlation

Cross-correlation (CC) is a well-established signal-processing technique used to measure the similarity between two signals as a function of the time lag applied to one of them [57]. For two receivers k and l the source signal $s(t)$, and noises $n_k(t)$ and $n_l(t)$ are real, jointly stationary random processes, where the signal $s(t)$ is assumed to be uncorrelated with noise $n_k(t)$ and $n_l(t)$. Therefore, the time lag τ_{kl} that maximizes the CC function

$$R_{kl}(\tau_{kl}) = E[s_k(t)s_l(t-\tau_{kl})] \tag{2.9}$$

serves as an estimate of the TDOA between two receivers k and l. Here, E denotes the expectation operator, while $s_k(t)$ and $s_l(t)$ represent the signals captured at the two receivers, as expressed in equation (2.3). Unfortunately, due to the finite observation time, the obtained TDOA estimate may be inaccurate. Therefore it is desirable to prefilter the signals $s_k(t)$ and $s_l(t)$ before evaluating the CC to enhance the accuracy of CC [57].

Generalized cross-correlation (GCC) is an extension of CC that employs FFT prefiltering on the signals $s_k(t)$ and $s_l(t)$. The observed signals (2.3) in the frequency domain is formed as follows:

$$S_m(f) = H_m(f)S(f) + N_m(f). \tag{2.10}$$

Here $H_m(f)$, $S(f)$, and $N_m(f)$ represent the The Discrete Time Fourier Transforms (DTFTs) of $h_m(t)$, $s(t)$, and $n_m(t)$, respectively. The parameter f denotes the frequency. The discrete-time GCC function is thus defined as

$$R_{kl}(\tau_{kl}) = \frac{1}{F}\sum_{f=0}^{F-1} \psi_{kl}(f)G_{kl}(f)e^{\frac{i2\pi f\tau_{kl}}{F}}, \tag{2.11}$$

where $e^{\frac{i2\pi f\tau_{kl}}{F}}$ realizes the phase shift, F represents the maximum frequency, $\psi_{kl}(f)$ denotes a general weighting function and

$$G_{kl}(f) = S_k(f)S_l^*(f), \tag{2.12}$$

where the operator $(\cdot)^*$ denotes the complex conjugate.

The CC function is thus the GCC function with the weighting function $\psi_{kl}(f) = 1$.

2.6.2 Generalized Cross-Correlation Phase Transform

While various weighting functions exist [57], the Phase Transform (PHAT) is the most widely used, as it has demonstrated superior performance in sound signal processing [5]. The PHAT weighting function is defined as follows:

$$\hat{\psi}_{kl}(f) = \frac{1}{|G_{kl}(f)|}. \tag{2.13}$$

Substituting (2.13) into (2.11) results in the Generalized Cross-Correlation Phase Transform (GCC-PHAT):

$$\hat{R}_{kl}(\tau_{kl}) = \frac{1}{F}\sum_{f=0}^{F-1}\frac{G_{kl}(f)}{|G_{kl}(f)|}e^{\frac{i2\pi f\tau_{kl}}{F}}. \tag{2.14}$$

The GCC-PHAT is a widely employed technique for estimating TDOAs between pairs of receivers.

2.6.3 CNN-based Parametrized GCC-PHAT

The accuracy of TDOA estimation can be improved by combining the power of CNNs and the GCC-PHAT [29]. Salvati *et al.* proposed a method for TDOA estimation using a parametrized GCC-PHAT (PGCC-PHAT) and regression-based CNN. The PGCC-PHAT is an extension of the traditional GCC-PHAT utilizing an additional parameter β, which modifies the influence of the PHAT weighting in the GCC-PHAT. The β-parametrized PHAT is defined as:

$$\tilde{\psi}_{kl}(\beta, f) = \hat{\psi}^{\beta}_{kl}(f). \tag{2.15}$$

By combining (2.15) and (2.11), we obtain the PGCC-PHAT, which is defined as:

$$\tilde{R}_{kl}(\beta, \tau_{kl}) = \frac{1}{F}\sum_{f=0}^{F-1}\frac{G_{kl}(f)}{|G_{kl}(f)|^{\beta}}e^{\frac{i2\pi f\tau_{kl}}{F}}. \tag{2.16}$$

Salvati *et al.* used the PGCC-PHAT as an input feature extractor for their CNN. By applying different values of β to the PGCC-PHAT on a pair of raw signals, they obtained the feature matrix

$$\tilde{\mathbf{R}}_{kl} = \begin{bmatrix} \tilde{R}_{kl}(\beta_1, -\tau_{\max_{kl}}) & \tilde{R}_{kl}(\beta_1, -\tau_{\max_{kl}}+1) & \dots & \tilde{R}_{kl}(\beta_1, \tau_{\max_{kl}}) \\ \tilde{R}_{kl}(\beta_2, -\tau_{\max_{kl}}) & \tilde{R}_{kl}(\beta_2, -\tau_{\max_{kl}}+1) & \dots & \tilde{R}_{kl}(\beta_2, \tau_{\max_{kl}}) \\ \vdots & \vdots & \dots & \vdots \\ \tilde{R}_{kl}(\beta_B, -\tau_{\max_{kl}}) & \tilde{R}_{kl}(\beta_B, -\tau_{\max_{kl}}+1) & \dots & \tilde{R}_{kl}(\beta_B, \tau_{\max_{kl}}) \end{bmatrix}$$

with dimensions of $B \times (2\tau_{\max_{kl}} + 1)$, where $\tau_{\max_{kl}}$ is the maximum possible TDOA given in samples and is defined as:

$$\tau_{\max_{kl}} = \left\lceil \frac{\|x_k - x_l\| f_s}{v_{\text{sound}}} \right\rceil, \tag{2.17}$$

where $\|x_k - x_l\|$ is the distance between the two microphones, and f_s is the sampling frequency of the processed signal. Without any further modification, they directly fed the feature matrix R_{kl} into the CNN to output the estimate of TDOA.

The regression CNN proposed by Salvati *et al.* comprises five convolutional blocks, each implementing a two-dimensional convolution with a kernel size of 3×3. Batch normalization and Rectified Linear Unit (ReLU) activation functions are applied following each convolution layer. The number of feature maps doubles with each block, commencing with 32 feature maps. Lastly, three fully connected layers with a dropout rate of 0.2 are employed to enhance the model's accuracy.

Salvati *et al.* conducted the training and evaluation of their approach using entirely synthetic data. The experimental setup involved two spatially separated microphones positioned 20 cm apart. The raw signal, sampled at a rate of 16 kHz and consisting of 1024 frames corresponding to 64 ms, was processed by PGCC-PHAT employing 11 distinct values of β, uniformly distributed within the range 0 to 1, inclusive.

For ease of notation, the abbreviation PGCC-PHAT will henceforth refer to Salvati et al.'s entire method, encompassing CNN.

2.6.4 Extended GCC-PHAT using SE NN

A promising approach to improve the precision of traditional methods such as GCC-PHAT involves signal preprocessing before applying conventional techniques. One possibility is the utilization of NNs as signal preprocessors, as demonstrated in a recent study by Berg *et al.*

[58]. The authors employed a Shift Equivariant (SE) CNN to prefilter signals from a pair of receivers, generating L distinct versions of each signal. With its non-linear filters, the SE NN effectively mitigates noise and reverberation effects [58]. When $L > 1$, the network can learn various non-linear filters that capture unique input signal properties, as Berg *et al.* observed.

Preserving time information in the signal is essential, necessitating using an SE preprocessing network. Shift equivariance ensures that the time lag of the original signal is equivalent to the time lag of the output signal [58]. If the SE NN preprocessing function is represented as $f_{\theta_{\text{SE}}}$ and $s_k(t)$ denotes the signal received at the receiver k (2.3), shift equivariance is defined as follows:

$$f_{\theta_{\text{SE}}}(s_k(\tau_k)) = s_k(f_{\theta_{\text{SE}}}(\tau_k)). \tag{2.18}$$

Berg *et al.* opted to use SincNet as the signal preprocessing CNN, which was proposed by Ravanelli *et al.* and designed for processing audio signals [59]. SincNet employs parametrized sinc functions as filters instead of traditional convolutional layers with fixed-size filters, allowing for greater flexibility in capturing frequency information in audio signals. For more information on the SincNet architecture, see [59].

Upon SincNet preprocessing, the filtered signal pairs form two matrices, each containing L versions of the original signal. Pairs of signals with the same index (i-th index in both matrices) are evaluated using GCC-PHAT, generating a single matrix with L entries as input for the second CNN, which produces the TDOA estimate.

The CNN employed by Berg *et al.* is a classification model consisting of $2\tau_{\text{max}} + 1$ output neurons, each representing the time lag within the range $(-\tau_{\text{max}}, \tau_{\text{max}})$. The TDOA estimating CNN comprises three convolutional blocks, each implementing 1D convolution with 128 feature maps, followed by 1D batch normalization, LeakyReLU activation, and a dropout rate of 0.5. The sizes of convolutional filters were 11, 9, and 7, respectively. An additional convolutional layer is applied with a kernel size of 5 and a single output feature map representing the log probabilities of each possible TDOA. All convolutional layers incorporate padding relative to the size of the convolutional kernel to prevent shrinking.

The proposed algorithm was trained and evaluated on synthetic data generated by a room sound simulation framework. The experimental setup included two microphones positioned 50 cm apart, with the simulated signal sampled at a rate of 16 kHz. The input signal to the algorithm was cropped to a length of 2048 frames, corresponding to 128 ms.

For the sake of brevity and clarity, this method will henceforth be referred to as NGCC-PHAT.

2.6.5 End-to-End CNN

Vera-Diaz *et al.* were the first to propose an End-to-End (E2E) approach employing DL techniques for the task of sound source localization [8]. In essence, they designed a CNN that accepts raw signals from an array of receivers as input and produces the corresponding coordinates of the sound source as output. The architecture of their E2E-CNN consists of five convolutional blocks followed by two fully connected layers. Max pooling operations are performed after the first, third, and fourth convolutional layers, with the kernel size matching that of the preceding convolutional layer. The output layer comprises three neurons that represent the spatial coordinates of the sound source. The input matrix for the E2E-CNN was constructed using recordings from two pairs of spatially separated microphones, with an intra-pair distance of 10 cm and an inter-pair separation of 80 cm. The model was trained and tested with three distinct window sizes at a sampling rate of 16 kHz, with the network's accuracy improving as the window size increased. The evaluated window durations were 80 ms, 160 ms, and 320 ms. To assess the performance of the E2E-CNN, the authors employed the Audio-Visual Corpus for Speaker Localization and Tracking (AV16.3) [60], a dataset consisting of real recordings of human speakers within a specified room. Vera-Diaz *et al.* initially trained their CNN on a synthetic dataset and refined the model's performance by fine-tuning it using a portion of the AV16.3 dataset.

2.6.6 Steered-Response Power Phase Transform

The Steered-Response Power Phase Transform (SRP-PHAT) is a beamforming-based approach utilized for sound source localization [32]. This method

searches for the candidate position that maximizes the output of a steered delay-and-sum beamformer. The candidate positions consist of all possible locations of the sound source within the room, divided into a finite grid with a specific resolution, typically 10 centimeters. The beamformer algorithm produces a power map based on which a grid search is conducted, identifying the source locations as peaks in the power map [4].

The SRP-PHAT power map is constructed using GCC-PHAT. The power map is computed as the sum of the GCC-PHAT outputs between all pairs of microphones within the array for each candidate position [32]:

$$P(x) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \hat{R}_{kl}(\tau_{kl}(x)), \tag{2.19}$$

where x represents the candidate's position and $\tau_{kl}(x)$ is the discrete TDOA for the signal emitted at the location x to microphones k and l.

$$\tau_{kl}(x) = \left\lfloor \frac{\|x - x_k\| - \|x - x_l\|}{v_{\text{sound}}} f_s \right\rceil \tag{2.20}$$

The source position estimate is then computed using the grid search. Let ς denote the set of all candidate positions. The estimated source position is:

$$\hat{x}_s = \arg\max_{x \in \varsigma} P(x). \tag{2.21}$$

The SRP-PHAT algorithm is commonly used as a baseline for evaluating other methods [8, 61] as it is robust to noise and reverberation.

2.6.7 TDE-ILD

Methods based on TDOA estimation typically require a minimum of four receivers to accomplish 3D SSL. However, the Intensity Level Difference (ILD) approach, also known as the Interaural Level Difference, provides the potential to reduce the number of required microphones. An approach proposed by Pourmohammad *et al.* demonstrates the capability of achieving complete 3D localization utilizing only three microphones. Their method combines Time Delay Estimation (TDE), ILD, and Head-Related Transfer Function (HRTF) analysis.

To achieve complete 3D localization, Pourmohammad *et al.* enveloped one of the microphones with a hemispherical shell containing slits and subsequently compared the signals captured by the enveloped microphone with those captured by the uncovered microphone, utilizing HRTF. Without the HRTF hemispherical covering, their algorithm is capable of half-plane 3D localization as it generates two mirror points on the y-axis. The TDE-ILD algorithm, which omits the use of HRTF, is delineated below. For a comprehensive understanding of their approach and the derivation of the necessary equations, refer to [35].

The specific arrangement of microphones utilized by Pourmohammad *et al.* played a pivotal role in formulating their algorithm. In their setup, the microphones were placed at coordinates $[R, 0, 0]$, $[-R, 0, 0]$, and $[R, 0, R]$. Two microphones were aligned horizontally, separated by a distance of $2R$, while the other two were aligned vertically, distanced by R.

The TDE-ILD algorithm, as presented by Pourmohammad *et al.*, is provided below:

1. Compute the energy ration m between microphone 1 and 2 within the observed time T:

$$m = \frac{E_1}{E_2}, \tag{2.22}$$

   where E_m is defined as

$$E_m = \int_0^T s_m^2(t)dt. \tag{2.23}$$

2. Calculate $\hat{\tau}_{21}$ and $\hat{\tau}_{31}$ using GCC-PHAT.

3. Compute the sound source location estimate using the following equations:

   (a)

$$\begin{aligned} \phi &= \cos^{-1}\left(\frac{\hat{\tau}_{21} v_{\text{sound}}}{2R}\right) \\ \theta &= \cos^{-1}\left(\frac{\hat{\tau}_{31} v_{\text{sound}}}{R}\right) \end{aligned} \tag{2.24}$$

(b)

$$\begin{aligned} r_1 &= \frac{\hat{\tau}_{21} v_{\text{sound}}}{1 - \sqrt{m}} \\ r_2 &= \frac{\hat{\tau}_{21} v_{\text{sound}} \sqrt{m}}{1 - \sqrt{m}} \end{aligned} \tag{2.25}$$

(c)

$$\begin{aligned} x &= \frac{r_2^2 - r_1^2}{4R} \\ y &= \pm\sqrt{r_1^2 - (x - R)^2} \end{aligned} \tag{2.26}$$

(d)

$$r = \sqrt{x^2 + y^2} \tag{2.27}$$

4. Finally, the estimated source location is obtained as:

$$\hat{x}_s = \begin{cases} x_s = \cos(\phi) \sin(\theta) r \\ y_s = \sin(\phi) \sin(\theta) r \\ z_s = \cos(\theta) r. \end{cases} \tag{2.28}$$

Pourmohammad et al. conducted outdoor experiments to examine their method, utilizing a loudspeaker as the sound source. It is important to acknowledge that the indoor environment, in comparison to outdoor settings, typically exhibits greater reverberation, which may unfavorably impact the captured energy at the microphones and consequently result in an inaccurate estimation of the energy ratio m.

2.6.8 Multilateration via Particle Swarm Optimization

Some earlier-mentioned algorithms are specifically designed to estimate the TDOA between pairs of receivers without directly providing an estimate of the source location. However, leveraging TDOA values garnered from an array of microphones (a minimum of four in a 3D scenario) enables the estimation of the source location via multilateration algorithms. One viable approach involves the application of an

optimization algorithm such as Particle Swarm Optimization (PSO) [62].

PSO is an iterative technique where each particle symbolizes a potential solution to the multilateration conundrum. Each particle preserves a record of its optimum positions throughout the process. During every iteration, a particle's new position is computed considering its optimal position as well as the position of the particle demonstrating the most favorable fitness within its local neighborhood.

In the context of multilateration employing PSO, the fitness function is defined as follows:

$$f_{\text{PSO}}(x) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} (\tau_{kl}(x) - \hat{\tau}_{kl})^2, \tag{2.29}$$

$\hat{\tau}_{kl}$ denotes the TDOA estimate derived from a TDOA estimation algorithm, while $\tau_{kl}(x)$ represents the TDOA value corresponding to the particle's position x. The fitness function quantifies the disparity between the estimated and actual TDOA values, thereby enabling PSO to optimize the particle positions and converge toward the actual source location.

2.6.9 Overview of Selected SSL methods

The methods delineated in this section encompass diverse general approaches to SSL. While certain methods are indirect—computing only a subset of the desired results—they contribute significantly to the SSL field. A succinct summary of these methods is presented in Table 2.1 for efficient comprehension. Each enumerated method has been designed for specific receiver configurations and signal characteristics, precisely the sampling frequency and window size (duration). Table 2.2 delineates each method's signal characteristics and microphone arrangements. The dashes in the table 2.2 denote that the method is independent of the specific configuration of the respective parameter. Furthermore, Table 2.3 specifies the training configurations utilized for methods that necessitate training.

In order to conduct a comprehensive evaluation, I have reimplemented all of the methods delineated in this section. I reimplemented

Method	Method Utilizes	Output
CC		TDOA
GCC-PHAT	CC	TDOA
PGCC-PHAT	PGCC-PHAT, CNN	TDOA
NGCC-PHAT	GCC-PHAT, CNN, SincNet	TDOA
E2E-CNN	CNN	Location
SRP-PHAT	GCC-PHAT	Location
TDE-ILD	GCC-PHAT	Location
PSO		Location

Table 2.1: A concise overview of selected methods for SSL.

Method	Number of Receivers	Receiver Arrangement	Sampling Frequency	Window Length
CC	2	-	-	-
GCC-PHAT	2	-	-	-
PGCC-PHAT	2	20 cm	16 kHz	64 ms
NGCC-PHAT	2	50 cm	16 kHz	128 ms
E2E-CNN	4	10 cm, 80 cm	16 kHz	360 ms
SRP-PHAT	-	-	-	-
TDE-ILD	3	1 m, 2 m	96 kHz	100 ms

Table 2.2: Employed microphone arrangements and signal characteristics.

Method	Criterion	Optimizer	Batch Size	Epochs
PGCC-PHAT	MSE	Adam	128	50
NGCC-PHAT	CE	Adam	32	30
E2E-CNN	MSE	Adam	100	250

Table 2.3: Suggested training configurations employed for the respective methods.

these methods based on the description provided in the papers where each method was proposed. Nevertheless, I drew inspiration and guidance from the existing reimplementation of PGCC-PHAT and NGCC-PHAT based on the code provided in the publication introducing NGCC-PHAT [58].

Necessary modifications tailored to meet the unique requirements of this study have been incorporated into these methodologies. Detailed descriptions of these modifications and their rationale are presented in Section 3.3. It is crucial to emphasize that I performed these changes judiciously, ensuring they do not distort the underlying principles of the original methods yet adapt them effectively for the specific context of this research.

3 Methodology

The subsequent sections detail the process of selecting and positioning microphones for data acquisition, as well as the creation of synthetic and real-world datasets. Furthermore, the necessary alterations to the selected algorithms, as delineated in section 2.6, are elucidated. These adaptations predominantly focus on adapting the methods to account for discrepancies in microphone arrangement.

3.1 Microphone Array Setup

Various SSL problems may necessitate varying numbers of microphones or alternative microphone arrangements. This section outlines the detailed process of selecting the most suitable microphone setup for the specific task of localizing active students in a classroom. Considering both cost and computational time is a necessity, as providing an affordable setup is a crucial expectation.

3.1.1 Microphone Quantity

Distinct methods for SSL rely on varying quantities of receivers; however, it is generally acknowledged that the minimal number of receivers for localization in an n-dimensional space is $n + 1$ [35]. Considering the objective of this thesis to localize sound sources in 3D spaces, a minimum of four receivers is mandatory. An increased number of receivers typically enhances accuracy since substantial errors in partial computations, such as TDOA estimation, can be mitigated by other partial computations. Nevertheless, this simultaneously results in elevated hardware and computational costs.

As microphone signals undergo processing on digital computational devices, it is imperative to employ a suitable audio card to convert analog signals from microphones into digital format. In order to guarantee simultaneous sampling of all signals, a singular audio card must be utilized. Thus, the audio card should have a minimum of M microphone inputs. Employing multiple audio cards is ill-advised since it cannot assure that all signals will be sampled concurrently,

which is a prerequisite for TDOA estimation and other related techniques.

Although extracting the position of a sound source in a 3D setting with a single receiver is feasible [30], such method relies on a complex analysis of the RIRF, which is distinct for each room and subsequently not portable [30]. Given that the portability of a solution with minimal or no intervention is crucial for the objectives of this thesis, an alternative approach is necessary.

Considering hardware costs and the requirements of the algorithms selected, as elaborated upon in Section 2.6, I employed four microphones for this study. This quantity represents the minimum expected to fulfill most of the algorithms designed for 3D SSL.

3.1.2 Microphone Arrangement

Microphone arrangement plays a pivotal role in SSL applications, influencing the precision and robustness of localization. Different studies have used diverse configurations, each tailored to the specific demands of the application. Nevertheless, there are two prevalent microphone arrangement concepts: a surround configuration where the sound source is encircled by microphones and another where the sound source can be situated anywhere around the microphones. These concepts are illustrated in Figure 3.1.

A fundamental assumption for many SSL methodologies is the direct sound propagation pathway connecting the sound source and all receivers. Thus, the configuration of microphones should not violate this postulate. This study aims to localize students within a classroom setting where all pupils are oriented toward the front board. As such, all microphones should be positioned at the front of the room to ensure unobstructed propagation pathways from students to the microphones. Two optimal microphone configuration placements uphold the direct propagation pathways assumption: an "on-wall" configuration aligning all microphones with the front wall and an "on-ceiling" configuration aligning all microphones with the ceiling.

In order to enhance the adaptability of the microphone setup, I designed a standardized configuration of microphones, which is independent of the room's geometry. This configuration facilitates deployment in any room and renders it suitable for various classroom

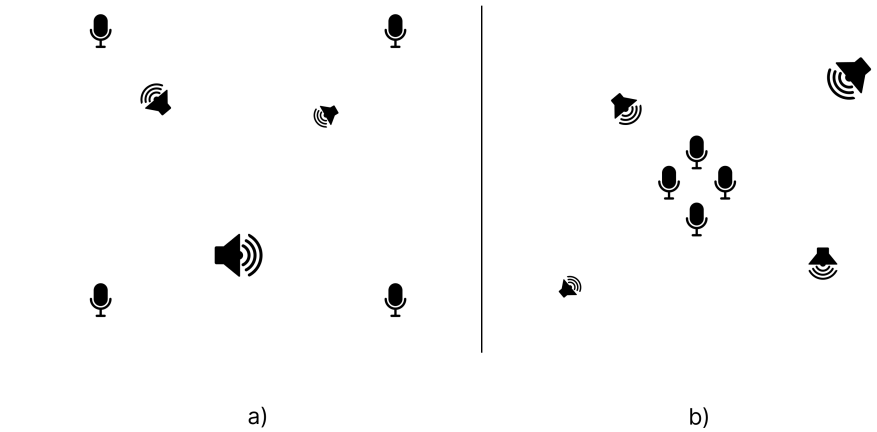

Figure 3.1: Two prevalent concepts of microphone arrangements: a) Denotes a concept where the sound source(s) are surrounded by microphones, whereas b) depicts a concept where sound source(s) appear outer the microphone array.

environments. For the "on-wall" configuration, a vertical circular array comprising four equidistantly spaced microphones was established, as illustrated in Figure 3.2. Microphones were oriented along the vertical and horizontal axes at a distance, denoted R, from the center of the circular microphone array. An identical microphone arrangement was utilized for the "on-ceiling" configuration, albeit rotated by $-90°$ along the horizontal axis (aligned with the ceiling rather than the front wall).

Presuming that all students face the front of the classroom, the center of the microphone array should be positioned in the middle of the room's width to evenly partition the room into horizontal halves. Moreover, the "on-wall" configuration should be positioned above the students to ensure direct propagation paths. It is also necessary to situate the microphone array in a location that does not disrupt the lecture, such as above the board.

Unfortunately, due to the limited height range of the camera tripod used as a stand for the microphone array during data collection, the center of the "on-wall" microphone array was fixed at a height of 1.5 meters. Conversely, the "on-ceiling" microphone array was fixed to

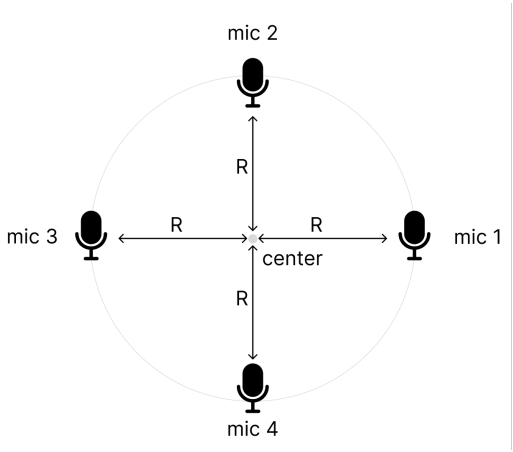

Figure 3.2: Proposed circular microphone array configuration comprising four uniformly spaced microphones.

a 1.5-meter distance from the front wall. The distance between all microphones within the array and the nearest wall or ceiling, denoted as d_{wall}, was estimated to be 0.3 meters, considering the size of the microphones and the connected cables. Therefore, in a room with dimensions $[x, y, z]$, where x represents the length, y the width, and z the height, the center of the microphone array was situated at coordinates $[0.3, y/2, 1.5]$ and $[1.5, y/2, z - 0.3]$ for the "on-wall" and "on-ceiling" configurations, respectively.

It is crucial to underscore that both proposed microphone configurations only support 3D half-plane localization as they produce two mirror points along the x-axis for the "on-wall" and the z-axis for the "on-ceiling" configuration. Fortunately, this half-plane consideration is sufficient as all sound sources should only manifest within one half-plane.

3.1.3 Microphone Directivity

Considering the proposed microphone configurations, I deemed the selection of unidirectional cardioid microphones appropriate. These microphones exhibit a cardioid directivity pattern, capturing sound predominantly from the front and sides while attenuating acoustic

signals from the rear. As rear-propagating sound primarily comprises reflected acoustic waves, employing unidirectional cardioid microphones may mitigate reverberation and noise within the recorded signal. Consequently, the directional attributes of these microphones aim to reduce the influence of extraneous noise and undesirable sound reflections on the acquired audio data.

3.1.4 Microphone Distance

The utilized microphone array forms a circular configuration with four uniformly spaced microphones at a distance of R from the array's center. The array configuration is depicted in Figure 3.2. To determine the optimal value of R, I created a synthetic dataset using the Pyroomacoustics framework [54]. This Distance Evaluation (DEval) dataset comprises 4,332 recordings obtained from four "on-wall" circular microphone arrays with the radius of R set to 0.1 m, 0.25 m, 0.5 m, and 1 m. I provide further details of the DEval dataset in Section 3.2.

The GCC-PHAT algorithm is employed in all the selected methods except for one, as discussed in Section 2.6. Therefore, to determine the optimal radius R, I assessed the performance of GCC-PHAT on each radius R incorporated in the DEval dataset. Because GCC-PHAT is a TDOA estimate method, I extracted all unique pairs of microphones for each circular array in the DEval dataset, resulting in 6 unique pairs for each microphone array. All four distances of R were evaluated using 25,992 samples; each randomly cropped to a length equivalent to 1 second.

Additionally, I evaluated the influence of sampling frequency on TDOA estimation using the GCC-PHAT algorithm, utilizing the DEval dataset. Given that the DEval dataset was generated using audio recordings sampled at a rate of 48 kHz, I chose sampling frequencies of 8 kHz, 16 kHz, 24 kHz, and 48 kHz to evaluate their respective impact on the GCC-PHAT algorithm's TDOA estimation performance.

The results of the TDOA estimation error as a function of the microphone distance and sampling frequency are presented in Table 3.1. It is evident from the table that the Root Mean Square Error (RMSE) increases with the microphone distance and decreases with increasing sampling frequency. However, it should be noted that the microphone distance and the sampling frequency define the density

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	0.214	0.564	1.329	2.928
16 kHz	0.202	0.537	1.232	2.784
24 kHz	0.187	0.525	1.212	2.73
48 kHz	**0.18**	0.515	1.196	2.707

Table 3.1: The impact of microphone distance and sampling frequency on the TDOA estimate RMSE in milliseconds for GCC-PHAT, utilizing a one-second window. The distance R represents the radius of a circular microphone configuration with four microphones positioned uniformly on the horizontal and vertical axes.

of hyperboloids, as described in Section 2.3.1. Consequently, a greater TDOA estimation error might result in more accurate localization as the density of hyperboloids increases with distance and sampling frequency.

In a 2D space, TDOA measurement delineates a hyperbola with foci situated at the microphones. Section 2.3.1 provides an in-depth discussion on this subject. To evaluate localization capabilities based on the acquired TDOA estimation errors, I assessed the distance between the y-axis and the hyperbola, or its asymptote, with foci at the x-axis. The radius R determines the hyperbola's eccentricity. At the same time, the length of its primary axis is derived from the translation of the measured errors to distance, analogous to the TDOA translation outlined in Section 2.3.1. These findings are presented in Table 3.2.

Furthermore, I have incorporated the results of GCC-PHAT TDOA estimation, employing the Mean Average Error (MAE) metric. The pertinent measurements are displayed in Tables 3.3 and 3.4. To expedite analysis, I converted the errors in Tables 3.1 and 3.3 from milliseconds to absolute errors in samples concerning the sampling frequency. These outcomes are exhibited in Tables 3.5 and 3.6. I observed that the RMSE results consistently surpassed the MAE results across all assessed distances, which can be ascribed to the RMSE's emphasis on the influence of outliers, as demonstrated in Tables 3.5 and 3.6. These observations suggest that the GCC-PHAT exhibited substantial outliers in its estimations.

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	0.395	0.42	0.512	0.581
16 kHz	0.369	0.396	0.466	0.543
24 kHz	0.338	0.386	0.457	0.53
48 kHz	**0.324**	0.378	0.45	0.524

Table 3.2: Table presenting the distance between the asymptote of hyperbolas and the y-axis at the coordinate point $y = 1$, wherein the hyperbolas' foci are located on the x-axis, with an eccentricity equivalent to the distance R of the microphone, and whose main axis is acquired through the translation of the measured TDOA RMSE to distance.

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	0.128	0.330	0.846	2.047
16 kHz	0.111	0.295	0.748	1.872
24 kHz	0.099	0.282	0.723	1.806
48 kHz	**0.092**	0.275	0.706	1.777

Table 3.3: The impact of microphone distance and sampling frequency on the TDOA estimate MAE in milliseconds for GCC-PHAT, utilizing a one-second window. The distance R represents the radius of the circular microphone configuration with four microphones positioned uniformly on the horizontal and vertical axes.

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	0.226	0.232	0.303	0.375
16 kHz	0.193	0.207	0.266	0.339
24 kHz	0.173	0.197	0.256	0.326
48 kHz	**0.161**	0.192	0.250	0.320

Table 3.4: Table presenting the distance between the asymptote of hyperbolas and the y-axis at the coordinate point $y = 1$, wherein the hyperbolas' foci are located on the x-axis, with an eccentricity equivalent to the distance R of the microphone, and whose main axis is acquired through the translation of the measured TDOA MAE to distance.

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	**1.7**	4.5	10.6	23.4
16 kHz	3.2	8.6	20.0	44.5
24 kHz	6.0	16.8	38.8	87.4
48 kHz	11.5	33.0	76.5	173.3

Table 3.5: The relative RMSE from Table 3.1, transformed to samples with respect of the sampling frequency.

Sampling Frequency	Circular Microphone Array Radius R [m]			
	0.1	0.25	0.5	1
8 kHz	**1.0**	2.6	6.8	16.4
16 kHz	1.8	4.7	12.0	30.0
24 kHz	2.4	6.8	17.4	43.4
48 kHz	4.4	13.2	33.9	85.3

Table 3.6: The relative MAE from Table 3.3, transformed to samples with respect of the sampling frequency.

Sampling Frequency	Microphone Distance d [m]			
	0.2	0.5	1	2
8 kHz	9	23	47	93
16 kHz	19	47	93	187
24 kHz	27	69	139	279
48 kHz	55	139	279	559

Table 3.7: The number of hyperboloids between two microphones distanced d meters.

Based on the measured data, I deduced that a microphone distance R of 0.1 meters within the circular array setup yields promising TDOA estimation results. Decreasing the radius below 0.1 meters is anticipated to yield smaller TDOA estimation errors; however, this would necessitate an increased sampling frequency to achieve accurate localization due to the diminishing density of hyperboloids as the distance decreases, while the density increases with the sampling frequency. Table 3.7 offers a comparative analysis of hyperboloid densities for varying microphone distances and sampling frequencies.

Despite the relatively low hyperboloid density observed between two microphones positioned 0.2 meters apart with a 48 kHz sampling frequency, the density should be sufficient for locating active pupils during classes. Therefore, considering this factor, I determined the radius R of the circular microphone array configuration to be 0.1 meters.

3.2 Datasets

This section comprehensively describes the construction of synthetic and real-world datasets for training and evaluating the selected SSL methods in indoor environments concerning two proposed microphone arrangements. A summary of the utilized datasets created in the context of this research is presented at the end of this section.

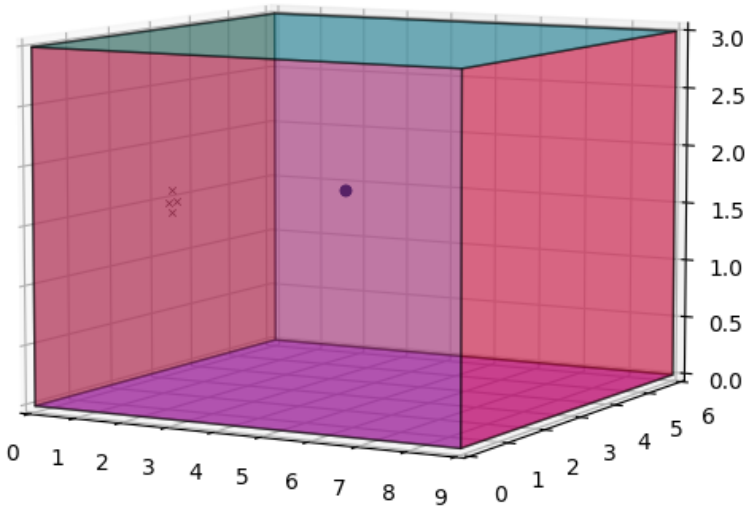

Figure 3.3: A schematic representation of a synthetic Shoebox-shaped room with dimensions of 9 meters in length, 6 meters in width, and 3 meters in height. The room contains an omnidirectional sound source located at coordinates [5, 2, 1.8] and the "on-wall" circular microphone array with microphones positioned at coordinates [0.3, 2.9, 1.5], [0.3, 3.1, 1.5], [0.3, 3, 1.4], and [0.3, 3, 1.6].

3.2.1 Synthetic Datasets

For this study, I employed the Pyroomacoustics framework [54] to generate a total of five distinct synthetic datasets. These datasets encompassed a dataset for ascertaining the optimal inter-microphone distance (termed the DEval dataset), two training datasets, and two synthetic evaluation datasets. Each synthetic dataset encapsulated simulations of sound propagation in Shoebox-shaped rooms [54], with a circular microphone array exhibiting cardioid directivity. A representative example of the simulated room featuring the "on-wall" microphone configuration is illustrated in Figure 3.3.

To augment the simulations' authenticity, variations in the signal-to-noise ratio (SNR) and reverberation times were integrated. The location of the omnidirectional sound source was uniformly generated within the room, adhering to a constraint of a minimum distance of

0.3 m from the front wall and elevation between 1.2 and 2 meters above the floor level for the "on-wall" microphone configuration. For the "on-ceiling" configuration, the source's location was regulated by an elevation constraint between 1.2 and 2 meters above the floor. The microphone array's center was anchored to the same relative coordinates as delineated in Section 3.1.

The Pyroomacoustics framework provides the capability to simulate custom audio signals, thereby enabling the inclusion of a diverse set of speech recordings. This feature ensured the creation of synthetic datasets that were both varied and realistic. Given these attributes, this approach was adjudged to be suitable for training and evaluating the efficacy of the methodologies under investigation.

DEval Dataset

The DEval synthetic dataset comprises sound propagation simulations executed in rooms of random dimensions, specifically length, width, and height. These dimensions were selected as integers from the ranges [4, 10], [3, 8], and [2, 4], respectively. The dataset was generated by leveraging speech recordings from the TSP speech database [63], which consists of 1444 utterances vocalized by 24 speakers, sampled at a frequency of 48 kHz. Each TSP audio signal was used in three individual simulations, culminating in a dataset containing 4,332 simulated instances. For every simulation, reverberation times T_{60} and SNR levels were randomly sampled within [0.2, 0.8] seconds and [0, 30] dB, respectively.

The DEval dataset employs four different "on-wall" circular microphone arrays, each characterized by radii of 0.1 m, 0.25 m, 0.5 m, and 1 m. The chosen radii for the microphone arrays in the DEval dataset were influenced by prior studies that utilized similar configurations for specific SSL tasks [29, 35, 58].

Training Datasets

The construction of the training datasets mirrors that of the DEval dataset, featuring simulations of varying SNR levels and reverberation times within a synthetic room of dimensions contingent on the utilized microphone configuration. For the "on-wall" configuration,

the synthetic room is 9 meters long, 8 meters wide, and 3 meters high. In contrast, for the "on-ceiling" microphone setup, the room measures 8 meters in length, 8 meters in width, and 3.3 meters in height.

Both training datasets were generated utilizing a circular microphone array with a radius of 0.1 m, with five simulations conducted per each TSP audio signal. This resulted in datasets containing 7,220 data points each.

Synthetic Evaluational Datasets

Synthetic evaluation datasets were constructed for each microphone configuration to scrutinize the impact of varying reverberation times and SNRs on the selected methods. These datasets encompass a broad spectrum of SNRs $\in \{-10, -5, 0, 5, 10, 15, 20, 25, 30\}$ and reverberation times $\mathrm{RT}_{60} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, employed during the simulation process.

In creating these evaluation datasets, 15 custom speech utterances were recorded, each with an approximate duration of 4 seconds, and sampled at a rate of 48 kHz. These utterances functioned as the simulated audio signals. For the "on-wall" configuration, a synthetic room with dimensions of $[7, 5, 3]$ meters was created, whereas for the "on-ceiling" configuration, the synthetic room was modeled with dimensions of $[7, 5, 3.3]$ meters.

Within both simulated environments, fifteen random positions were chosen, each assigned a unique recorded utterance. An anechoic room simulation was conducted for each of the fifteen positions per SNR level, followed by a simulation of varying reverberation times with a fixed SNR level set to 30 dB. Consequently, each Synthetic Evaluation dataset comprises 270 distinct data points.

3.2.2 Real-World Dataset

This section delineates the procedure for constructing a real-world dataset derived from recordings obtained in four distinct indoor environments, utilizing the "on-wall" circular microphone array with a radius of 0.1 meters. Initially, I explicate the hardware utilized for capturing the audio signals. Subsequently, the methodology employed for extracting speech segments is elucidated, followed by a compre-

hensive description of the environments and the speaker's locations within each setting.

Microphones

The challenge of localizing active pupils in a classroom, where considerable distances exist between speakers and receivers, necessitates using sensitive microphones. In this context, the selection of an appropriate microphone type is critical. Compared to dynamic microphones, condenser microphones are highly favored due to their enhanced sensitivity. Moreover, I consider the cardioid directivity suitable for the microphone placement configuration presented in Section 3.1. Consequently, I employed four RØDE M5[1] microphones, featuring a small-diaphragm condenser design with a cardioid polar pattern and a sensitivity of -34 dB re 1V/Pa. These attributes render the RØDE M5 microphone highly appropriate for the objectives of this study.

Audio Card

A crucial parameter of the audio card, significantly impacting its functionality, is its sampling frequency. In addition to requiring at least four microphone inputs, the selected audio card must support a sampling frequency that accommodates data at the sampling rate of 48 kHz, as indicated in Section 3.1. Therefore, the audio card must support a minimum sampling frequency of 48 kHz. The Behringer U-PHORIA UMC404HD[2] is a suitable audio-card that meets these criteria, featuring four phantom-powered microphone inputs and supporting data sampling frequencies up to 192 kHz.

Microphone Array Holder Construction

I manufactured a simple construction for holding the microphones to obtain real-world data. The construction comprised a primary holder consisting of four uniformly spaced beams, each forming a right angle with its neighboring beams. I used the whorl rods for the beams. Standard microphone holders were attached to each beam using nuts,

1. https://rode.com/en/microphones/studio-condenser/m5
2. https://www.behringer.com/product.html?modelCode=P0BK1

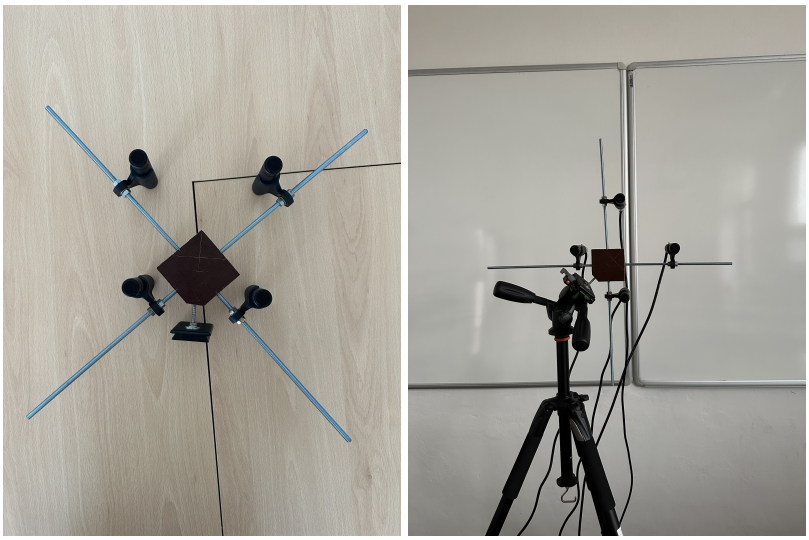

Figure 3.4: An illustration of used microphone holder.

providing the flexibility to adjust the distance between the microphones and configure them in the desired circular arrangement. This feature proved particularly beneficial in compensating for any potential inaccuracies that may have arisen during the manufacturing process. A camera tripod served as the base for the microphone array, allowing for adjustable height and rotation. The construction of the microphone holder utilized for acquiring real-world data is illustrated in Figure 3.4.

Acoustic Signal Extraction

The unavailability of a wireless controller for managing the recording process during real-world data acquisition led to procuring a continuous audio file encompassing multiple speech segments. In this context, a speech segment refers to an utterance from a specific location. To isolate individual speech segments from the continuous audio file, I employed a neural speaker diarization technique, pyannote.audio [64]. However, I, unfortunately, found the precision of the segmentation process to be insufficient. Therefore, I utilized the pyannote.audio

method for data preprocessing, and segmented the utterances with manual intervention.

Data Annotation

I cut the segmented speech utterances creating a multitude of four-channel audio files, each comprising a single utterance originating from a specific location. The dataset is organized into distinct folders corresponding to individual acquisition rooms, containing all extracted speech utterances in the form of audio files. Furthermore, the dataset contains a JSON annotation file *<room-id>-labels.json* for each acquisition room. These annotation files encompass room dimensions, spatial coordinates of all microphones, and for each extracted speech segment, the coordinates of the sound origin, the path to the corresponding audio file, and the transcript of the spoken utterance. The sound source coordinates are provided relative to the center of the microphone array. The sequence of the four channels in the audio file corresponds to the positions of the microphones listed in the annotation file. An exemplar JSON label file is presented below:

```
{
  "room-dimensions": [8.2, 6.85, 3],
  "receivers": [
    [0.7, 3.53, 1.5],
    [0.7, 3.43, 1.6],
    [0.7, 3.33, 1.5],
    [0.7, 3.43, 1.4]
  ],
  "data": [..., {
    "source-position": [2.18, -3.1, 0.12],
    "label": "The sun rises in the east and sets
        in the west.",
    "signal": "real-data/A320/26.flac"
  }, {
    "source-position": [2.11, -1.68, 0.12],
    "label": "Water is essential for life on
      Earth.",
    "signal": "real-data/A320/27.flac"
  }, ...]
```

Room ID	Length [m]	Width [m]	Height [m]	Positions
A320	8.2	6.85	3	25
KD	13.95	11.8	7	106
C511	8.15	5.3	3.25	13
C525	9	4.9	3.25	14

Table 3.8: Dimensions of real-world rooms and number of unique positions from which utterances were spoken in each room

}

Dataset Acquisition

I chose four unique indoor environments for the acquisition of real-world data. Three of these environments, specifically rooms A320, C525, and C511, are academic classrooms, while the fourth, designated KD, is a hall within the House of Culture, representing a spacious environment. All rooms adhere to a conventional shoebox shape except for KD, which deviates from the standard rectangular geometry due to a stage and an indoor balcony. Consequently, the dimensions provided for KD correspond to the size of its dance floor. Table 3.8 delineates the dimensions of each room, while visual depictions of the rooms can be found in Figures 3.5, 3.6, 3.7, and 3.8.

Throughout the data collection process, an effort was made to position the microphones as proximal to the front wall as practicable to minimize deviations from the estimated microphone distance from the wall, as discussed in Section 3.1. However, the breadth of the tripod stand utilized for the microphone array and obstructions within the rooms resulted in actual wall distances of 0.7, 1, 0.37, and 0.38 meters for rooms A320, KD, C511, and C525, respectively. The respective annotation files document the exact locations of all microphones within each room.

Within each room, I chose multiple locations where the speaker stood or sat facing the microphone array, delivering short utterances generated by the ChatGPT chatbot[3]. The utterances were spoken in

3. `https://chat.openai.com/chat`

Room ID	Number of Data Points	Total [s]	Mean [s]	Min [s]	Max [s]
A320	93	303	3.3	1.0	7.4
KD	241	610	2.5	1.0	5.3
C511	54	296	5.5	1.9	10.6
C525	58	294	5.1	1.5	9.5

Table 3.9: Detailed information of the data points in the created dataset.

both Czech and English languages. Transcripts of all spoken utterances can be found in the annotation files. Due to the absence of a tracking device to monitor the speaker's precise location during the recording phase, minor deviations between the actual and reported location may have occurred, amounting to a few centimeters. Nonetheless, I assume these discrepancies between the provided location and the true location to be negligible. The number of positions selected within each room summarizes the Table 3.8 and visual representations of the selected locations in each room can be found in Figures 3.9, 3.10, 3.11, and 3.12. Table 3.9 provides detailed information about the dataset, including the number of data points obtained in each room, total duration, mean data point duration, and other relevant metrics.

Data acquisition was performed at a sampling rate of 192 kHz, which is the maximum frequency supported by the used audio card. While a sampling rate of 48 kHz would have sufficed for the aims of this research, the higher rate of 192 kHz may offer potential advantages for future applications. The real-world dataset (ISL-Dataset) created as part of this study is publicly available on GitHub[4].

3.2.3 Overview of Constructed Datasets

Table 3.10 delineates all the datasets generated employing the "on-wall" microphone configuration. The datasets created using the "on-ceiling" microphone configuration only encompass the training and

4. `https://github.com/JakubPekar/ISL-Dataset`

Figure 3.5: Photo of the A320 university classroom.

Figure 3.6: Photo of KD spacious room.

Figure 3.7: Photo of the classroom C511.

Figure 3.8: Photo of the classroom C525.

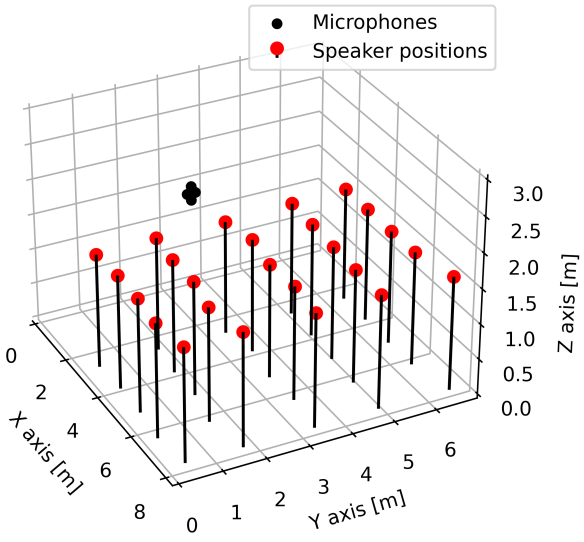

Figure 3.9: 3D coordinate map of speaker locations in the room A320

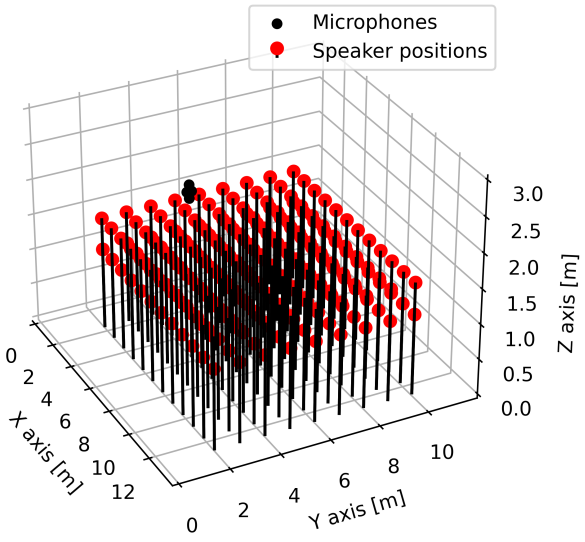

Figure 3.10: 3D coordinate map of speaker locations in the room KD

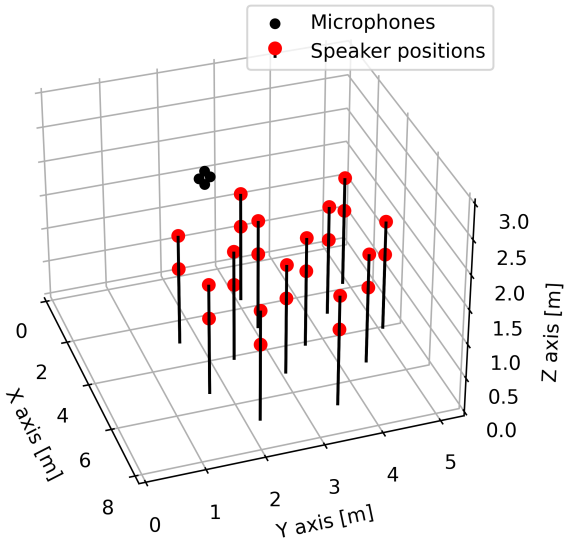

Figure 3.11: 3D coordinate map of speaker locations in the room C511

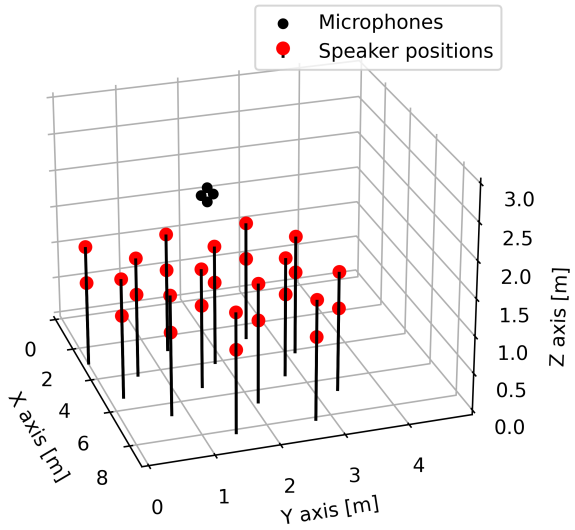

Figure 3.12: 3D coordinate map of speaker locations in the room C525

ID	Synthetic	Sampling Frequency [kHz]	Number of Data Points
DEval	✓	48	4332
Training	✓	48	7220
Eval	✓	48	1350
A320	✗	192	93
KD	✗	192	241
C511	✗	192	54
C525	✗	192	58

Table 3.10: Summary of all created datasets using "on-wall" microphone setup.

ID	Number of Data Points	Total Duration [s]
A320	17	49.7
KD	40	96.0
C511	8	45.5
C525	0	0

Table 3.11: Summary of the ISL-Dataset extracted subset.

synthetic evaluation datasets. Regrettably, I could not create a real-world evaluation dataset employing the "on-ceiling" microphone placement. Nevertheless, the "on-ceiling" microphone array essentially represents a translated version of the "on-wall" configuration, with the x and z axes interchanged. As such, I resolved to extract a subset from the ISL-Dataset, which was subsequently used to evaluate both microphone configurations. I selected all data points with a relative source position on the x-axis (room's length) in the [1, 2.2] meters interval. Using this methodology, I acquired a subset of 65 unique data points with an overall duration of 191.2 seconds. Unfortunately, all acquisition in room C525 was conducted with the relative position on the x-axis greater than 2.2 meters. Therefore no data points from this room were selected. Table 3.11 illustrates the quantity and duration of the selected data points within each respective room.

3.3 Implementation of Selected Algorithms

Based on the empirical findings regarding the influence of sampling frequency on the accuracy of TDOA estimation using the GCC-PHAT method, as discussed in Section 3.1, a fixed sampling frequency of 48 kHz was chosen for the input signal. It is well-established that enlarging the duration of the input signal, typically referred to as the window size, can enhance the precision of localization methods, provided that the environment exhibits stationary properties during the observation period. Hence, I set the window length T to 1 second, which I assume is a duration in which the classroom environment exhibits stationary behavior.

Although some methods are accessible as pre-trained models [58], their practical application was impeded by inconsistencies in the microphone configuration, sampling frequency, and input signal duration, as detailed in Table 2.2. Therefore I re-implemented selected methods and, if necessary, retrained the method to meet the specific requirements of this study. The implementation was conducted using the PyTorch framework [65], which I deemed the most suitable tool for this purpose.

The modifications applied to the selected methods were tailored to a circular microphone array with a radius R of 0.1 m, an input signal duration T of 1 second, and a sampling frequency of 48 kHz to maintain consistent evaluation conditions for the selected methods. In cases where data points in both the simulated and ISL-Dataset fell short of one second, they were supplemented with zeros. The subsequent sections describe the adjustments made to each chosen method.

Methods requiring training underwent training on a synthetic training dataset as described in Section 3.2. I partitioned the dataset into 80 % segments for training and 20 % segments for validation. Throughout the training and validation stages, the input signals were arbitrarily cropped to a 1-second duration or padded with zeros when required.

Due to variations in the size of training and evaluation rooms, all the methods that provide location estimates present them relative to the microphone array. The relative positions are particularly signifi-

cant in the case of E2E-CNN, which lacks positional encoding of the microphones.

The source code for this research is publicly available on GitHub[5].

3.3.1 CC and GCC-PHAT

The CC and GCC-PHAT operate independently of microphone distance, sampling frequency, and signal duration. However, these factors do impact the precision of the algorithms. The TDOA estimation between two receivers, denoted as k and l, is determined by the parameter τ_{kl} that maximizes equation 2.9, which is mathematically expressed as:

$$\hat{\tau}_{kl} = \arg\max_{\tau_{kl} \in \mathbb{Z}} R_{kl}(\tau_{kl}). \tag{3.1}$$

Nonetheless, a plausible TDOA estimate between receivers k and l resides within the interval $(-\tau_{\max_{kl}}, \tau_{\max_{kl}})$. To avoid potentially unrealistic estimates, the TDOA estimate in between microphones k and l in both CC and GCC-PHAT methods is computed as follows:

$$\hat{\tau}_{kl} = \arg\max_{\tau_{kl} \in (-\tau_{\max_{kl}}, \tau_{\max_{kl}})} R_{kl}(\tau_{kl}). \tag{3.2}$$

3.3.2 PGCC-PHAT

The PGCC-PHAT is a TDOA estimation technique that utilizes the GCC-PHAT as a feature extractor. In the context of the proposed circular microphone arrays, variations in the input signal parameters only affect the size of the input feature matrix $R_{kl}^{(\beta\text{-PHAT})}$. Nevertheless, the proposed circular microphone arrays comprises two distinct inter-microphone distances. The horizontal and vertical microphone spacing is $2R = 0.2$ m, while the distance between adjacent microphones on the circle approximates $\sqrt{2R^2} \approx 0.14$ m.

Initially, I tried to train a single PGCC-PHAT network using an input feature matrix with dimensions 11×57, appropriate for microphones spaced 0.2 meters apart. Given that PGCC-PHAT is a TDOA estimation technique and the suggested microphone arrays incorporates four microphones, a microphone pair was randomly selected

5. `https://github.com/JakubPekar/isl-algorithms`

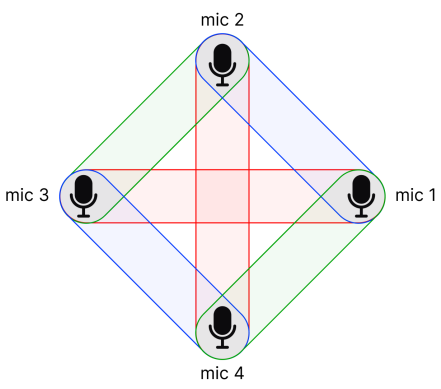

Figure 3.13: Three distinct two-pair microphone combinations in the proposed microphone array are indicated by the colors red (R), green (G), and blue (B). These combinations encompass all six unique microphone arrangements in the proposed array.

from six unique combinations for each sample during the training and validation phases. The training process utilized the Mean Squared Error (MSE) loss function, Adam optimizer with a learning rate of 0.001, a batch size of 128 samples, and output normalization. These training hyperparameters are consistent with those presented in the research of Salvati *et al.* [29]. Regrettably, this network did not demonstrate effective learning.

Three separate PGCC-PHAT networks were trained to rectify this learning deficit, each corresponding to a distinct microphone pair combination. These combinations are graphically represented in Figure 3.13, with each color denoting a unique pairs of microphones: red (R), green (G), and blue (B).

The input matrices for each PGCC-PHAT network possess dimensions of 11×57, 11×41, and 11×41 for PGCC-PHAT-R, PGCC-PHAT-G, and PGCC-PHAT-B, respectively. All three networks were trained using the same set of hyperparameters previously delineated.

3.3.3 NGCC-PHAT

In this study, the NGCC-PHAT was adapted by adjusting the SincNet parameters to match a sampling frequency of 48 kHz and a signal

length of 48,000 samples, equivalent to a duration of 1 second. Consequently, modifications were made to the subsequent CNN, particularly concerning the size of the output classification layer.

Similar to the PGCC-PHAT network, I attempted to train a single network with an output classification layer of size $2\tau_{\max} + 1$. Here, $\tau_{\max}$ represents the maximum TDOA among all combinations of receivers and is mathematically articulated as:

$$\tau_{\max} = \max_{k,l \in M} \left\lceil \frac{\|x_k - x_l\| f_s}{v_{\text{sound}}} \right\rceil. \tag{3.3}$$

The network training incorporated identical hyperparameters as delineated in the referenced literature [29]. Considering that both NGCC-PHAT and PGCC-PHAT function as TDOA estimators, a similar training strategy was adopted for NGCC-PHAT. However, the output normalization stage was excluded due to its inherent dependency on classification.

Regrettably, I discovered that the NGCC-PHAT network was incapable of simultaneous training on all unique combinations of microphones. Consequently, three distinct NGCC-PHAT networks, specifically NGCC-PHAT-R, NGCC-PHAT-G, and NGCC-PHAT-B, were established following the procedure delineated in Section 3.3.2. The dimensions of the output classification layer for each network were ascertained based on the maximum feasible TDOA between pairs of microphones.

3.3.4 E2E-CNN

The E2E-CNN technique, originally formulated for a microphone setup containing four microphones, necessitated alterations to the input parameters, particularly impacting the model's dimensions. The model training process exploited identical hyperparameters as outlined in the study by Vera-Diaz *et al.* [8].

3.3.5 SRP-PHAT

The SRP-PHAT method does not impose any explicit restrictions on parameters such as sampling frequency, window size, or microphone configuration. Consequently, this methodology does not necessitate

any inherent modifications. Nevertheless, I have reduced the set of all candidate positions ς that the SRP-PHAT algorithm explores to optimize the method's performance.

This reduction strategy involved eliminating candidate positions that reside outside the area where a sound source might be potentially located, as considered during the creation of synthetic datasets. The determination of this area is dependent on the microphone configuration, as elaborated in Section 3.2. By limiting the range of potential candidate positions, the search efficiency of the SRP-PHAT algorithm can be significantly improved.

3.3.6 TDE-ILD

The TDE-ILD technique initially employed three microphones, two arranged vertically, and two arranged horizontally, as delineated in Section 2.6.7. Even though proposed microphone configurations deviate in its arrangement, it consists of a pair of horizontally arranged and a pair of vertically arranged microphones, visually denoted in Figure 3.13 using the color red. Since the vertically arranged microphones from my configuration can exclusively be used for estimating the angle of elevation (θ) without any adverse effects on the algorithm, they can be interchanged with the originally employed vertical microphones. However, due to the alignment discrepancy between my configuration and the TDE-ILD technique, it was necessary to interchange the x and y axes in their algorithm to implement the TDE-ILD methodology in my setup accurately.

3.3.7 PSO

Analogous to SRP-PHAT, Multilateration via PSO is a universal algorithm indifferent to specific microphone configurations. The algorithm's input comprises TDOA estimates between all unique pairs of microphones and is independent of signal characteristics. The search space where particles are generated was restricted similarly to the context of SRP-PHAT.

3.4 Evaluation Metodology

As articulated in Section 3.2, a subset of the ISL-Dataset was employed to assess the localization capabilities of the selected methods. This subset encapsulates recordings conducted at position on the x-axis in the range of [1, 2.2] meters relative to the placement of the microphone array. The ISL-Dataset was generated using an "on-wall" setup. To derive a dataset compatible with the "on-ceiling" placement, I transformed the dataset by interchanging the x and z axes during the evaluation of the "on-ceiling" setup. Despite the transformed dataset not being optimal for the "on-ceiling" setup assessment, the transformation does not adversely impact the data. Consequently, this modified dataset was employed for the "on-ceiling" setup evaluation.

The transformation, however, results in rooms spanning several meters in height, with the sound sources positioned on the z-axis in the range [-2.2, -1] meters relative to the microphone array. To address this issue, the search space of SRP-PHAT and PSO methods was adjusted relative to the location of the microphone array. More specifically, the bounds impacting the elevation were altered, such that the new bound was [-2.2, -1] meters relative to the microphone array's position.

Throughout the evaluation phase, each data point across all datasets earmarked for evaluation was divided into segments of one second in duration, each with a 50 % overlap. The duration of data points varies, resulting in a variable number of segments for each data point. Each segment, comprising signals from four microphones, was subsequently employed as inputting all the chosen methods.

The TDOA estimation methods require two signals from a pair of microphones as input. Given that the proposed setup includes four microphones, the input segment was partitioned into six pairs of signals, which were then utilized as inputs to the TDOA estimation methods. Owing to the creation of three variants of the PGCC-PHAT and NGCC-PHAT method, each designed for specific pairs of microphones within the proposed microphone array, the six combinations of input signals were segmented, evaluated by the relevant network, and the outputs were amalgamated, thereby providing TDOA estimates for the input signals. These TDOA estimates were subsequently input into the PSO algorithm, which yielded the estimated source location.

For the evaluation metric, I utilized the Multiple Object Tracking Precision (MOTP), which was formulated under the CHIL project and is defined as [61]:

$$MOTP = \frac{\sum_{i=1}^{N} \|\hat{x}_i - x_i\|}{N}, \tag{3.4}$$

where N represents the number of predictions (segments), $\hat{x}_i$ is the ground truth sound location, and x_i is the predicted location. The MOTP quantifies the average distance between the predicted and actual locations.

4 Results

This section elucidates the results of the localization performance of the chosen methods employing two distinct microphone configurations: "on-wall" and "on-ceiling." These configurations are detailed in Section 3.1. Initially, the selected methods were evaluated on synthetic evaluation datasets to inspect the influence of varying SNR levels and reverberation times on the performance of these methods. Since half of the chosen methods do not directly calculate localization but estimate the TDOA between a pair of microphones, the TDOA estimation performance of these methods was primarily evaluated. As the TDOA estimation evaluation metric, I used the RMSE. Figure 4.1 illustrates the RMSE results under varying SNR conditions in an anechoic room, as well as the effects of different reverberation times with SNR fixed at the level of 30 dB for both implemented microphone configurations.

The TDOA estimation results indicate that the performance of TDOA estimations declines with the increment in reverberation time. This outcome aligns with expectations, as a more ample reverberation time signifies increased acoustic wave reflections in the captured signal. The GCC-PHAT method consistently enhanced over CC across all analyzed reverberation times. This superior performance indicates why GCC-PHAT is generally favored over traditional CC in signal-processing applications.

In the context of the "on-wall" setup, both PGCC-PHAT and NGCC-PHAT notably outperformed the conventional GCC-PHAT across all tested SNR levels and reverberation times. The mean reduction of RMSE compared to the conventional GCC-PHAT was 0.14 ms for PGCC-PHAT and 0.09 ms for NGCC-PHAT at various SNR levels. There was also an improvement of 0.17 ms and 0.14 ms for PGCC-PHAT and NGCC-PHAT, respectively, in different reverberation times.

The "on-ceiling" configuration often results in sound sources exhibiting a more substantial misalignment relative to the microphone positions, which induces signals to arrive at larger angles. This discrepancy potentially incites a minor degradation in the TDOA estimation performance compared to the "on-wall" setup. An exception to this trend is the NGCC-PHAT network, which manifested a significantly

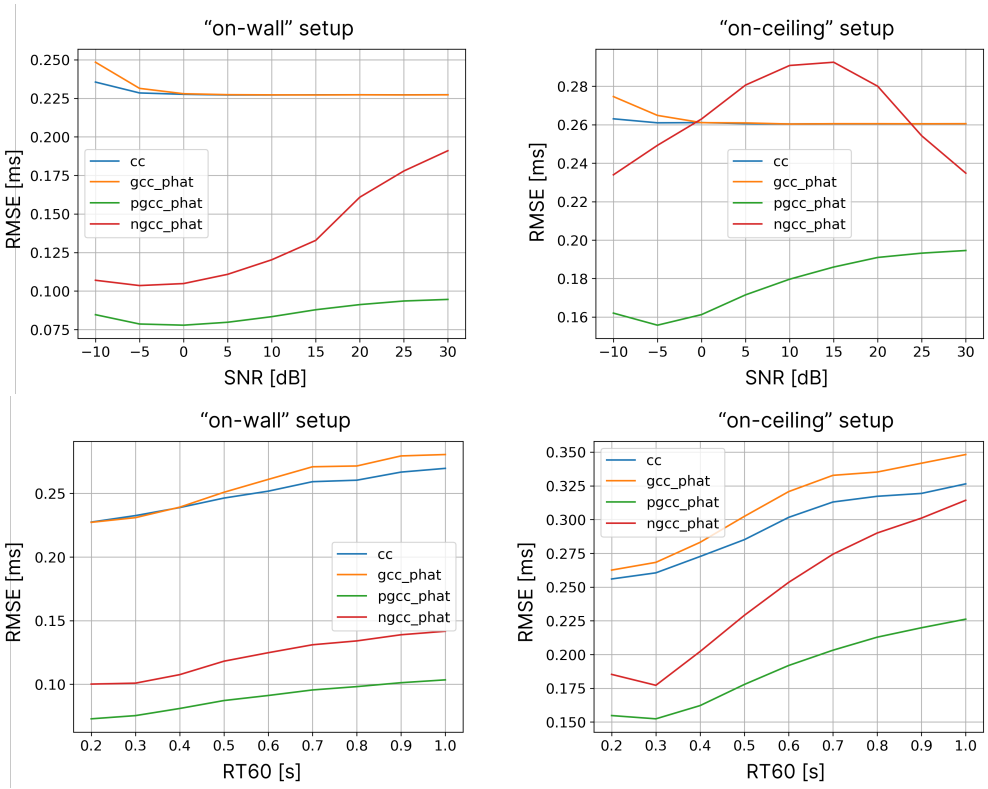

Figure 4.1: Influence of various SNR levels and reverberation times (RT_{60}) on TDOA estimation performance of different methods utilizing two microphone arrangements.

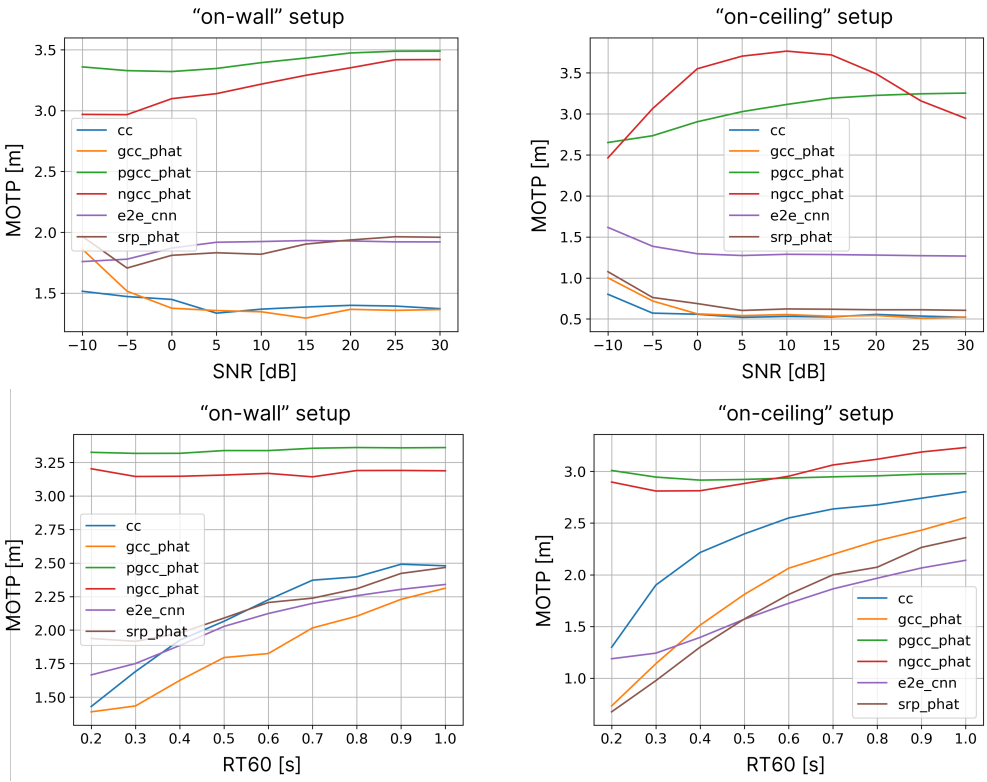

Figure 4.2: Influence of various SNR levels and reverberation times (RT_{60}) on the localization performance of different methods utilizing two microphone arrangements measured by the MOTP.

larger decline in TDOA estimation performance. This decrease could be ascribed to the insufficiency of the training dataset.

The localization potential of the selected methods was subsequently assessed on synthetic evaluation datasets utilizing the MOTP loss metric. Additionally, localization accuracy is provided as the percentage of segments whose estimated location fell within a 1-meter radius of the actual location. These results are visually represented in Figures 4.2 and 4.3. The TDE-ILD method was omitted from the graphical representations due to its inadequate MOTP performance (exceeding 25 meters) and an average localization accuracy below 1 %.

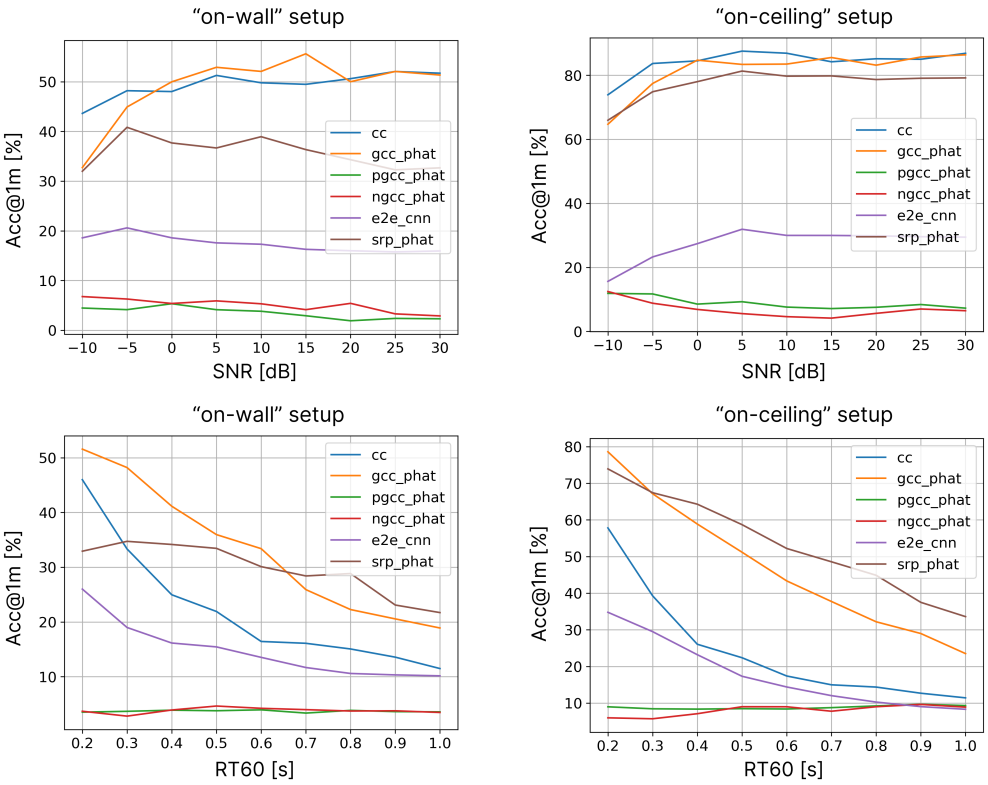

Figure 4.3: Influence of various SNR levels and reverberation times (RT_{60}) on the localization accuracy of different methods utilizing two microphone arrangements.

The MOTP results infer that elevating the SNR level does not influence the localization performance unless the SNR levels fall below 0 dB. However, the reverberation time significantly impacts the investigated methods' localization performances. PGCC-PHAT and NGCC-PHAT contravene this trend, but these methods exhibit poor localization capabilities with an average MOTP exceeding 2.5 meters. The localization performance of methods other than PGCC-PHAT, NGCC-PHAT, and TDE-ILD was contingent on the employed microphone array. In the case of the "on-wall" setup, the GCC-PHAT algorithm outperformed others with an average MOTP of 2.0 m and average localization accuracy greater than 33 %. Conversely, in the "on-ceiling" microphone setup, SRP-PHAT excels with a mean localization accuracy of 47 % and mean MOTP of 1.67 meters. In the anechoic environment, employing the "on-ceiling" microphone configuration, three methods, namely CC, GCC-PHAT, and SRP-PHAT, demonstrated localization accuracy exceeding 80

The TDOA estimation performances and localization outcomes of the selected methods are presented, utilizing a subset of the ISL-Dataset. The TDOA estimation results are depicted in Figure 4.4, whereas Figure 4.5 displays the MOTP and localization accuracy of the chosen methods for both implemented microphone configurations. Due to its subpar performance, with an average MOTP error of 24.5 meters and an average localization accuracy below 1 %, the TDE-ILD method was excluded from the graphical representation.

Amongst the assessed methods, the GCC-PHAT algorithm displayed superior TDOA estimation performance, providing location estimates with an accuracy of 50 % for each room within the evaluated subset of the ISL-Dataset, when utilizing the "on-ceiling" microphone arrangement. However, shifting to the "on-wall" microphone placement decreased its localization effectiveness, yielding an average MOTP of 2.0 meters and an average localization accuracy of 25 % across all tested real-world rooms.

The results derived from real-world data indicate that the "on-ceiling" setup yields localization outcomes that are up to twice as accurate as those obtained from the "on-wall" setup. In both configurations, the top-performing methods were CC, GCC-PHAT, and SRP-PHAT. These methods significantly outperformed all evaluated machine learning-based methods regarding localization results. Three

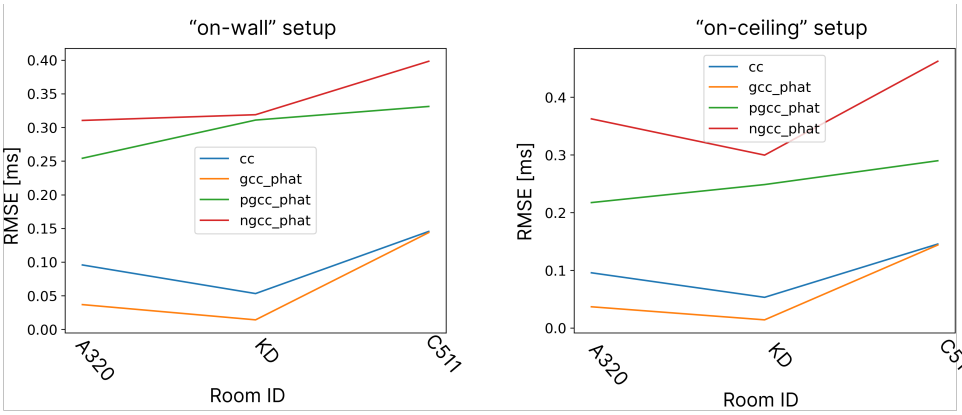

Figure 4.4: The TDOA estimation performance of selected algorithms utilizing two microphone arrangements on the ISL-Dataset subset.

methods achieved a localization accuracy greater than 50 % in at least one of the ISL-Dataset rooms using the "on-ceiling" configuration. Consequently, the localization accuracy results within a 0.5 meters radius are provided for these methods, as displayed in Figure 4.6.

Notably, the GCC-PHAT utilizing the PSO multilateration provided the highest overall localization accuracy.

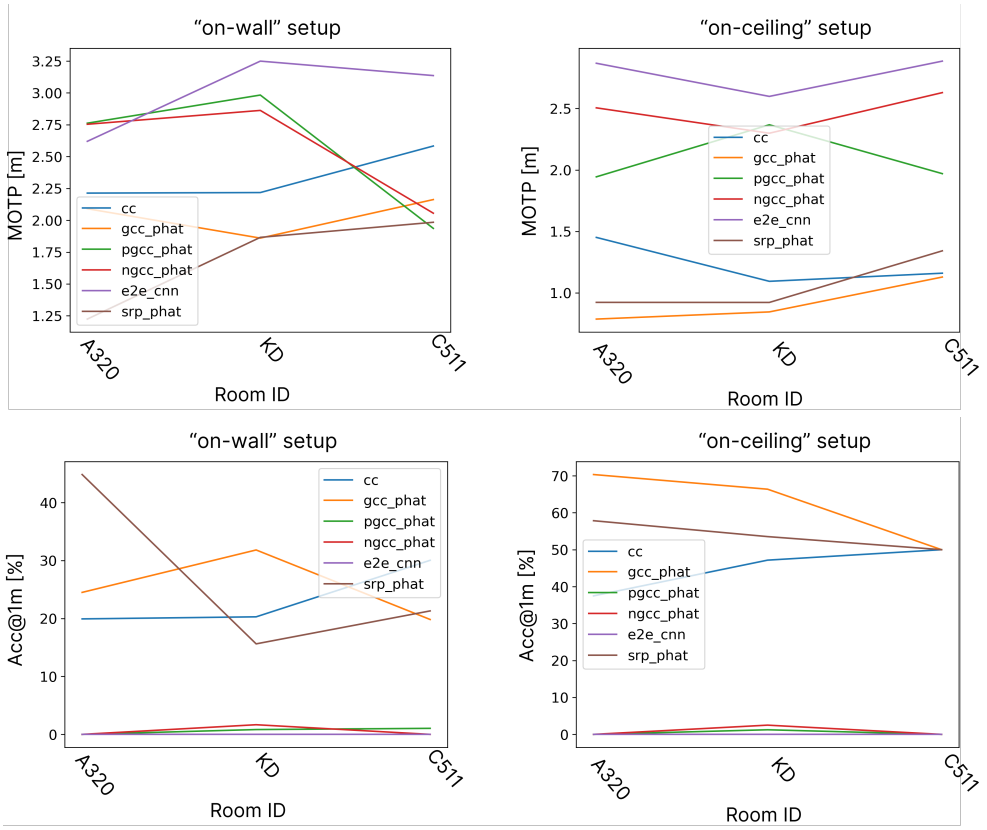

Figure 4.5: The MOTP and localization accuracy of the selected methods utilizing two microphone arrangements on the ISL-Dataset subset.

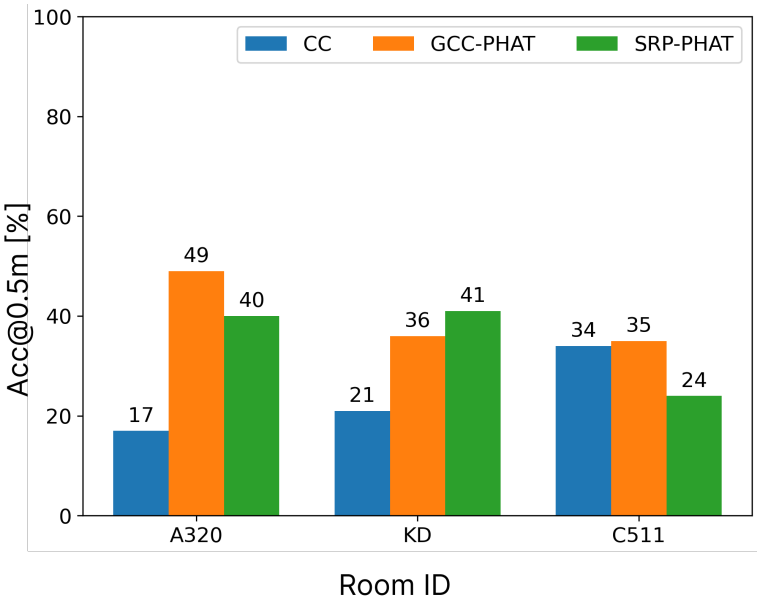

Figure 4.6: Localisation accuracy within the radius of 0.5 meters of the three best-performing methods on the ISL-Dataset subset utilizing the "on-ceiling" microphone configuration.

5 Discussion

The efficacy of sound localization methods is contingent upon the configuration of the microphone array. In both proposed microphone arrangements, the microphones are organized in a circular array with a radius of 0.1 meters. Nonetheless, it is not always feasible to ascertain the distance between the sound source and the microphone array based on the TDOA information gleaned from all pairs of microphones. For instance, for a sound source situated arbitrarily distant from the microphone array along the axis to the microphone array, the TDOA among all microphone pairs is invariantly zero, irrespective of the distance between the sound source and the microphone array. By displacing the sound source off the microphone axis within a prescribed space and grid density, the concurrent sound source location, exhibiting identical TDOAs, exclusively depends on the circular microphone array radius and the sampling frequency.

Regrettably, for the proposed circular microphone array and signal sampled at a rate of 48 kHz in both training and evaluation rooms, numerous concurrent locations are exhibiting identical TDOAs. Figure 5.1 illustrates an example of concurrent positions for a sound source situated at coordinates [6, 4, 2] meters in a room with dimensions [8, 6, 3] meters and grid density of 0.1 meters utilizing the "on-wall" microphone array setup. The demonstrated example shows multiple concurrent locations starting near the microphone array and ranging to the room boundaries. Consequently, without incorporating energy information, it is only feasible to extract directional information from the acquired TDOAs.

Given that all selected methods, except for TDE-ILD, are designed to rely solely on TDOA information, their measured localization performances were influenced by concurrent locations. Prime examples are the PGCC-PHAT and NGCC-PHAT methods, which have demonstrated significant TDOA estimation improvements over the GCC-PHAT algorithm on synthetic data. However, they yielded poor localization accuracy on both synthetic and real-world data.

The circular microphone array with a radius of 0.1 meters is employed in both presented microphone setups. However, due to the vertical limit of the space where the sound source may appear, the

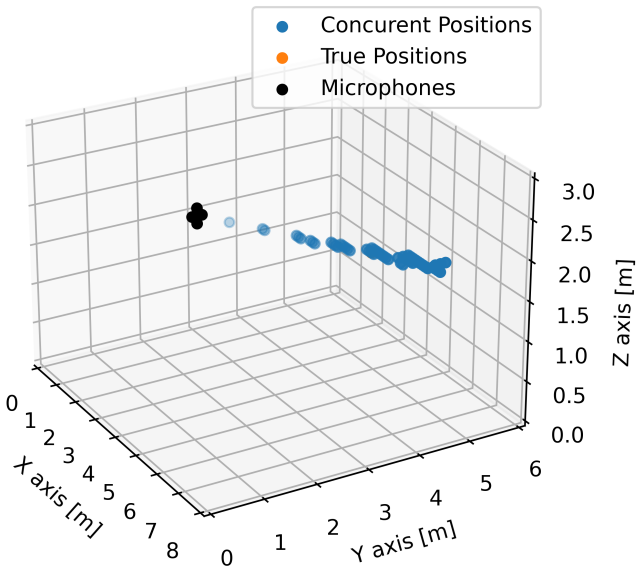

Figure 5.1: Concurent locations within a room of dimension [6, 4, 2] meters utilizing the "on-wall" microphone setup with sound source at coordinates [6, 4, 2] meters.

"on-ceiling" setup permits a greater focus on the direction towards the sound source rather than the distance factor. The results indicate that the SRP-PHAT and GCC-PHAT methods significantly improved localization accuracy for real-world recordings' "on-ceiling" setup.

Concerning the performance of methods necessitating training, the E2E-CNN demonstrates suboptimal localization capabilities across all evaluation data. This can likely be attributed to the microphone arrangement and the training dataset. As detailed by Vera-Diaz *et al.* [8], the E2E-CNN method was designed to concentrate on the TDOA information within the captured signal. However, due to the inability to extract distance information from TDOA estimates, the E2E-CNN solely learned to output the coordinates of the approximate center of the room in the training dataset. Both PGCC-PHAT and NGCC-PHAT exhibited a significant decline in TDOA estimation performance on real-world data compared to synthetic data. In contrast, GCC-PHAT and CC display a considerable improvement in TDOA estimation on real-world data relative to synthetic data. This leads to the conclusion that substantial discrepancies exist between synthetic and real-world signals, which have not been accounted for in the network training and consequently impact TDOA prediction capabilities.

The TDE-ILD method is the sole selected approach incorporating signal energy information, theoretically enabling it to provide the most accurate localization estimates. Regrettably, as evidenced in the results, this method exhibited the poorest localization performance with an average MOTP exceeding 25 meters, which is well outside the bounds of all evaluation rooms. Further examination of the method revealed that it suffers from inaccurately estimated energy ratios between two signals, as explained in Section 2.6.7. This is likely due to the small distance between microphones. For context, the microphones were placed two meters apart in the original setup. Another contributing factor could be reverberation, but even in an anechoic chamber devoid of reverberation, the TDE-ILD method performed poorly.

Two distinct methodologies have demonstrated noteworthy localization efficacy on synthetic and real-world data sets, employing the "on-ceiling" microphone placement. These include a beamforming-based technique, SRP-PHAT, and the TDOA-based technique, GCC-PHAT, which leverages PSO for multilateration. Both techniques achieved the MOTP of approximately one meter across all tested real-world

rooms and roughly half a meter on synthetic anechoic data. Given the complexity of the task and the potential deficiencies of the employed microphone arrays, as previously discussed, both methodologies have demonstrated a high degree of localization precision. Regrettably, the localization error remains too substantial to pinpoint speaking students within an educational setting accurately.

Acknowledging that various approaches exist for localizing or identifying a speaking student within an educational setting is crucial. The SSL approach represents one possible method; however, alternative techniques could potentially be employed. In the context of measuring speaking time for students in educational environments, direct speaker detection is more desirable. One option for detecting speaking students is speaker diarization, a process that separates distinct speakers within an audio signal [66]. Another possibility is implementing Active Speaker Detection (ASD) algorithms using audio-visual recordings (e.g., from a camera) [67].

The ASD approach may be the most suitable, as its outcome typically consists of face crops for all individuals in the video and a label indicating the probability of a person being the source of speech. Applying face recognition algorithms to the face crops enables sharing of student data across different classrooms or settings. Currently, an annual challenge sponsored by Google focuses on ASD [68], with participating Deep Neural Network (DNN) models demonstrating promising results. Unfortunately, increasing the number of faces in the video raises the computational demands of ASD models [69, 70], rendering current ASD models unsuitable for detecting speaking students in classrooms.

6 Conclusion

This thesis aimed to find the most promising approach for localizing speaking pupils within an educational environment using four microphone audio recordings. Given the absence of a prescribed microphone configuration, I designed two potential microphone setups that could be implemented within educational settings.

Regrettably, to the best of my knowledge, there were no publicly accessible datasets consisting of real-world recordings from a similar microphone setup. Consequently, I created a dataset of recordings from four rooms utilizing the "on-wall" microphone configuration. This dataset includes recordings of spoken utterances from multiple positions and also contains the transcripts of the spoken utterances, making it potentially useful for other tasks such as speech processing.

Several methods incorporating various general approaches to sound source localization were then selected and reimplemented, including approaches based on cross-correlation, phase correlation, and neural networks. Since some of the chosen methods required training, synthetic training datasets were generated using the sound propagation simulation software Pyroomaccoustics [54], and the models were subsequently trained.

The chosen methods and microphone configurations were assessed using synthetic and real-world data. I discovered that all the methods requiring training were incapable of handling real-world data, presumably due to significant discrepancies between the real-world and synthetic data used for training. However, the traditional approaches, except for the TDE-ILD method, demonstrated substantial localization accuracy with real-world evaluation data. The "on-ceiling" placement exhibited superior localization capabilities compared to the "on-wall" microphone configuration. This is primarily due to the inability to determine the distance between the sound source and the microphone array, where the "on-ceiling" placement benefits from the constraints imposed on the expected location of the speaker's height.

This research established that two specific localization techniques, namely, SRP-PHAT and the GCC-PHAT methods, which utilize PSO for multilateration, displayed solid localization efficacy on synthetic and real-world data when paired with the "on-ceiling" microphone

placement. Both methods exhibited a MOTP error of approximately one meter with real-world data. However, this localization error remains too significant for the accurate localization of speaking students within an educational setting.

In summary, three evaluated SSL methods demonstrated commendable localization performance on localizing single dominant sound sources in low reverberation rooms, using acoustic signals from a circular microphone array composed of four uniformly spaced microphones. However, none of the methods suits the localization of speaking students within an educational setting.

In future research, I intend to focus on different microphone placements that could be more suitable for complete 3D localization using DNNs. Given the remarkable performance of neural networks across various domains, I remain confident in their potential for sound source localization. The failure of the utilized neural networks in this research context can likely be attributed to inadequate training datasets and substantial discrepancies between the real-world and synthetic signals.

Bibliography

1. VATANSEVER, Saffet; BUTUN, Ismail. A broad overview of GPS fundamentals: Now and future. In: *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*. 2017, pp. 1–6. Available from DOI: 10.1109/CCWC.2017.7868373.
2. WANG, S.S.; GREEN, M.; MALKAWA, M. E-911 location standards and location commercial services. In: *2000 IEEE Emerging Technologies Symposium on Broadband, Wireless Internet Access. Digest of Papers (Cat. No.00EX414)*. 2000, p. 5. Available from DOI: 10.1109/ETS.2000.916514.
3. CARTER, G. Clifford. Time delay estimation for passive sonar signal processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1981, vol. 29, pp. 463–470.
4. MALMGREN, Anna. *Sound Source Localization for an Urban Outdoor Setting - A Systematic Review*. 2021.
5. ÇATALBAŞ, Cem; DOBRIŠEK, Simon. Dynamic speaker localization based on a novel lightweight R-CNN model Dynamic Speaker Localization Based on a Novel Lightweight R-CNN Model. *Neural Computing and Applications*. 2023. Available from DOI: 10.1007/s00521-023-08251-3.
6. WANG, H.; CHU, P. Voice source localization for automatic camera pointing system in videoconferencing. In: *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1997, vol. 1, 187–190 vol.1. Available from DOI: 10.1109/ICASSP.1997.599595.
7. AN, Inkyu; SON, Myungbae; MANOCHA, Dinesh; YOON, Sung-Eui. Reflection-Aware Sound Source Localization. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 66–73. Available from DOI: 10.1109/ICRA.2018.8461268.
8. VERA-DIAZ, Juan; PIZARRO, Daniel; MACIAS-GUARASA, Javier. Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates. *Sensors*. 2018, vol. 18, no. 10, p. 3418. Available from DOI: 10.3390/s18103418.

9. GRUMIAUX, Pierre-Amaury; KITIĆ, Srđan; GIRIN, Laurent; GUÉRIN, Alexandre. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*. 2022, vol. 152, no. 1, pp. 107–151. Available from DOI: 10.1121/10.0011809.

10. ANNIBALE, Paolo; FILOS, Jason; NAYLOR, Patrick A.; RABENSTEIN, Rudolf. TDOA-Based Speed of Sound Estimation for Air Temperature and Room Geometry Inference. *IEEE Transactions on Audio, Speech, and Language Processing*. 2013, vol. 21, no. 2, pp. 234–246. Available from DOI: 10.1109/TASL.2012.2217130.

11. POSTOLACHE, Octavian; PEREIRA, José; GIRÃO, Pedro. Underwater Acoustic Source Localization and Sounds Classification in Distributed Measurement Networks. In: *Advances in Sound Localization*. 2011. ISBN 978-953-307-224-1. Available from DOI: 10.5772/15487.

12. BOHN, Dennis. Environmental effects on the speed of sound. *Journal of the audio engineering society*. 1988, vol. 36, no. 4, pp. 223–231.

13. DUSHAW, Brian; WORCESTER, Peter; CORNUELLE, Bruce; HOWE, Bruce. On equations for the speed of sound in seawater. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*. 1992, vol. 91. Available from DOI: 10.1121/1.403304.

14. CRAMER, Owen P. The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and CO2 concentration. *Journal of the Acoustical Society of America*. 1993, vol. 93, pp. 2510–2516.

15. KUPERMAN, W.A. Acoustics, Deep Ocean. In: COCHRAN, J. Kirk; BOKUNIEWICZ, Henry J.; YAGER, Patricia L. (eds.). *Encyclopedia of Ocean Sciences (Third Edition)*. Third Edition. Oxford: Academic Press, 2019, pp. 296–307. ISBN 978-0-12-813082-7. Available from DOI: https://doi.org/10.1016/B978-0-12-409548-9.11386-7.

16. DCASE. *Sound event localization and detection evaluated in real spatial sound scenes*. 2022. Available also from: https://dcase.community/challenge2022/task-sound-event-localization-

and-detection-evaluated-in-real-spatial-sound-scenes. Accessed on 20. 4, 2023.

17. EVERS, Christine; LOLLMANN, Heinrich W.; MELLMANN, Heinrich; SCHMIDT, Alexander; BARFUSS, Hendrik; NAYLOR, Patrick A.; KELLERMANN, Walter. The LOCATA Challenge: Acoustic Source Localization and Tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2020, vol. 28, pp. 1620–1643. Available from DOI: 10.1109/taslp.2020.2990485.

18. FINFER, D.C.; WHITE, Paul; LEIGHTON, Timothy; HADLEY, M.; HARLAND, E. On clicking sounds in UK waters and a preliminary study of their possible biological origin. In: *Conference: 4th International Conference on Bio-Acoustics*. 2007.

19. MAHER, Robert C. Fundamentals of Audio Signals and Systems. In: *Principles of Forensic Audio Analysis*. Cham: Springer International Publishing, 2018, pp. 3–28. ISBN 978-3-319-99453-6. Available from DOI: 10.1007/978-3-319-99453-6_2.

20. HEWETT, David Peter. 2010. Available also from: https://www.ucl.ac.uk/~ucahdhe/10_Hewett_DPhilThesis.pdf. PhD thesis.

21. KUTTRUFF, Heinrich. *Room Acoustics*. 2016. ISBN 9781315372150. Available from DOI: 10.1201/9781315372150.

22. SHENG, Xiaohong; HU, Yu Hen. Energy Based Acoustic Source Localization. In: *Information Processing in Sensor Networks*. 2003, vol. 2643, pp. 551–551. ISBN 978-3-540-02111-7. Available from DOI: 10.1007/3-540-36978-3_19.

23. RABENSTEIN, Rudolf; ANNIBALE, Paolo. Acoustic Source Localization under Variable Speed of Sound Conditions. *Wireless Communications and Mobile Computing*. 2017, vol. 2017, pp. 1–17. Available from DOI: 10.1155/2017/9524943.

24. MENG, Wei; XIAO, Wendong. Energy-Based Acoustic Source Localization Methods: A Survey. *Sensors*. 2017, vol. 17, no. 2. ISSN 1424-8220. Available from DOI: 10.3390/s17020376.

25. DISW, Siemens. *Sound Fields: Free versus Diffuse Field, Near versus Far Field*. 2020. Available also from: https://community.sw.siemens.com/s/article/sound-fields-free-versus-diffuse-field-near-versus-far-field. Accessed on 13. 4, 2023.

26. SIANO, Daniela; VISCARDI, Massimo; PANZA, Maria Antonietta. Experimental acoustic measurements in far field and near field conditions: characterization of a beauty engine cover. In: *12th International Conference on Fluid Mechanics & Aerodynamics (FMA 2014) and 12th International Conference on Heat Transfer, Thermal Engineering and Environment (HTE 2014)*. 2014.

27. E-BOOK, The microflown. *Near Field Acoustics*. 2009. Available also from: https://www.microflown.com/resources/e-books/e-book-the-microflown-e-book. Accessed on 13. 3, 2023.

28. MANDLIK, Michal; BRÁZDA, Vladimír. Sound Source Location Method. *Perner's Contacts*. 2011, vol. 6, no. 4, pp. 197–204. Available also from: https://pernerscontacts.upce.cz/index.php/perner/article/view/916.

29. SALVATI, Daniele; DRIOLI, Carlo; FORESTI, Gian. Time Delay Estimation for Speaker Localization Using CNN-Based Parametrized GCC-PHAT Features. In: *Interspeech 2021*. 2021, pp. 1479–1483. Available from DOI: 10.21437/Interspeech.2021-988.

30. TUKULJAC, Helena Peic; LISSEK, Herve; VANDERGHEYNST, Pierre. *Localization of Sound Sources in a Room with One Microphone*. 2017. Available from eprint: arXiv:1707.04504.

31. LIAQUAT, Muhammad; MUNAWAR, Hafiz; RAHMAN, Amna; QADIR, Zakria; KOUZANI, Abbas; MAHMUD, M. A. Localization of Sound Sources: A Systematic Review. *Energies*. 2021, vol. 14, p. 3910. Available from DOI: 10.3390/en14133910.

32. COBOS, Maximo; MARTI, Amparo; LOPEZ, Jose J. A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling. *IEEE Signal Processing Letters*. 2011, vol. 18, no. 1, pp. 71–74. Available from DOI: 10.1109/LSP.2010.2091502.

33. GREWAL, Mohinder S.; ANDREWS, Angus P.; BARTONE, Chris G. Fundamentals of Satellite Navigation Systems. In: *Global Navigation Satellite Systems, Inertial Navigation, and Integration*. 2020, pp. 21–41. Available from DOI: 10.1002/9781119547860.ch2.

34. SAND, Stephan; MENSING, Christian; DAMMANN, Armin. Positioning in Wireless Communications Systems – Introduction and Overview. In: *18th Wirless World Research Forum (WWRF) Meeting*. 2007.

35. POURMOHAMMAD, Ali; AHADI, Seyed Mohammad. N-dimensional N-microphone sound source localization. *EURASIP Journal on Audio, Speech, and Music Processing*. 2013, vol. 2013, p. 27. Available from DOI: 10.1186/1687-4722-2013-27.

36. KITIĆ, Srđan; GAULTIER, Clément; PALLONE, Grégory. *A Comparative Study of Multilateration Methods for Single-Source Localization in Distributed Audio*. 2020. Available from arXiv: 1910.10661 [cs.SD].

37. TORRES-HERRERA, Luis; VASQUEZ-MONROY, Dainer. Acoustic source localizer using wireless sensor networks. In: *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*. 2017, pp. 1–6. Available from DOI: 10.1109/CHILECON.2017.8229587.

38. DENG, Fang; GUAN, Shengpan; YUE, Xianghu; GU, Xiaodan; CHEN, Jie; LV, Jianyao; LI, Jiahong. 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. Available from DOI: 10.1109/TIE.2017.2652394.

39. ADAVANNE, Sharath; POLITIS, Archontis; VIRTANEN, Tuomas. A Multi-room Reverberant Dataset for Sound Event Localization and Detection. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, NY, USA, 2019, pp. 10–14. Available also from: https://dcase.community/workshop2019/proceedings.

40. POLITIS, Archontis; ADAVANNE, Sharath; VIRTANEN, Tuomas. A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop* (*DCASE2020*). Tokyo, Japan, 2020, pp. 165–169. Available also from: `https://dcase.community/workshop2020/proceedings`.

41. POLITIS, Archontis; ADAVANNE, Sharath; KRAUSE, Daniel; DELEFORGE, Antoine; SRIVASTAVA, Prerak; VIRTANEN, Tuomas. A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection. In: *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop* (*DCASE2021*). Barcelona, Spain, 2021, pp. 125–129. ISBN 978-84-09-36072-7. Available also from: `https://dcase.community/workshop2021/proceedings`.

42. POLITIS, Archontis; SHIMADA, Kazuki; SUDARSANAM, Parthasaarathy; ADAVANNE, Sharath; KRAUSE, Daniel; KOYAMA, Yuichiro; TAKAHASHI, Naoya; TAKAHASHI, Shusuke; MITSUFUJI, Yuki; VIRTANEN, Tuomas. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In: *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop* (*DCASE2022*). Nancy, France, 2022, pp. 125–129. Available also from: `https://dcase.community/workshop2022/proceedings`.

43. RASCON, Caleb; MEZA, Ivan. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*. 2017, vol. 96, pp. 184–210. ISSN 0921-8890. Available from DOI: `https://doi.org/10.1016/j.robot.2017.07.011`.

44. HARTLEY, Richard I.; STURM, Peter. Triangulation. *Computer Vision and Image Understanding*. 1997, vol. 68, no. 2, pp. 146–157. ISSN 1077-3142. Available from DOI: `https://doi.org/10.1006/cviu.1997.0547`.

45. SEN, Satyabrata; NEHORAI, Arye. Performance Analysis of 3-D Direction Estimation Based on Head-Related Transfer Function. *IEEE Transactions on Audio, Speech, and Language Processing*. 2009,

vol. 17, no. 4, pp. 607–613. Available from DOI: 10.1109/TASL.2008.2008235.

46. ZHONG, Xiao-li; XIE, Bo-sun. Head-Related Transfer Functions and Virtual Auditory Display. In: GLOTIN, Herve (ed.). *Soundscape Semiotics*. Rijeka: IntechOpen, 2014, chap. 6. Available from DOI: 10.5772/56907.

47. MERINO-MARTÍNEZ, R.; SIJTSMA, P.; SNELLEN, M.; AHLEFELDT, T.; ANTONI, J.; BÄHR, C.J.; BLACODON, Daniel; ERNST, D.; FINEZ, A.; FUNKE, S.; GEYER, T.; HAXTER, S.; HEROLD, G.; HUANG, X.; HUMPHREYS, M.; LECLERE, Quentin; MALGOEZAR, A.; MICHEL, U.; PADOIS, T.; PEREIRA, A.; PICARD, C.; SARRADJ, E.; SILLER, H.; SIMONS, D.; SPEHR, C. A review of acoustic imaging methods using phased microphone arrays. *CEAS Aeronautical Journal*. 2019, vol. 10, no. 1, pp. 197–230. Available from DOI: 10.1007/s13272-019-00383-4.

48. DCASE. *Sound Event Localization and Detection with Directional Interference - Challenge results*. 2021. Available also from: https://dcase.community/challenge2021/task-sound-event-localization-and-detection-results. Accessed on 10. 4, 2023.

49. DCASE. *Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes - Challenge results*. 2022. Available also from: https://dcase.community/challenge2022/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes-results#Du_NERCSLIP_task3_report. Accessed on 10. 4, 2023.

50. NUGENT, James. *The 3 main types of microphones (with subtypes) and their best uses*. 2021. Available also from: https://higherhz.com/microphone-types/.

51. HENSHALL, Marc. *Microphone directionality and polar pattern basics*. [N.d.]. Available also from: https://www.shure.com/en-US/performance-production/louder/microphone-directionality-polar-pattern-basics. Accessed on 20. 4, 2023.

52. NAYLOR, G.M. ODEON—Another hybrid room acoustical model. *Applied Acoustics*. 1993, vol. 38, no. 2, pp. 131–143. ISSN 0003-682X. Available from DOI: https://doi.org/10.1016/0003-682X(93)90047-A.

53. PICAUT, Judicaël; FORTIN, Nicolas. I-Simpa, a graphical user interface devoted to host 3D sound propagation numerical codes. *Société Française d'Acoustique*. 2012.

54. SCHEIBLER, Robin; BEZZAM, Eric; DOKMANIC, Ivan. Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018. Available from DOI: 10.1109/icassp.2018.8461310.

55. ALLEN, Jont; BERKLEY, David. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*. 1979, vol. 65, pp. 943–950. Available from DOI: 10.1121/1.382599.

56. VORLÄNDER, Michael. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*. 1989, vol. 86, no. 1, pp. 172–178. Available from DOI: 10.1121/1.398336.

57. KNAPP, C.; CARTER, G. 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. Available from DOI: 10.1109/TASSP.1976.1162830.

58. BERG, Axel; O'CONNOR, Mark; ÅSTRÖM, Kalle; OSKARSSON, Magnus. Extending GCC-PHAT using Shift Equivariant Neural Networks. In: *Interspeech 2022*. ISCA, 2022. Available from DOI: 10.21437/interspeech.2022-524.

59. RAVANELLI, Mirco; BENGIO, Yoshua. *Speaker Recognition from Raw Waveform with SincNet*. 2019. Available from arXiv: 1808.00158 [eess.AS].

60. LATHOUD, Guillaume; ODOBEZ, Jean-Marc; GATICA-PEREZ, Daniel. AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking. In: BENGIO, Samy; BOURLARD, Hervé (eds.). *Machine Learning for Multimodal Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 182–195. ISBN 978-3-540-30568-2.

61. VELASCO, Jose; PIZARRO, Daniel; MACIAS-GUARASA, Javier. Source Localization with Acoustic Sensor Arrays Using Generative Model Based Fitting with Sparse Constraints. *Sensors*. 2012, vol. 12, no. 10, pp. 13781–13812. ISSN 1424-8220. Available from DOI: 10.3390/s121013781.

62. GAO, Weihua; KAMATH, Ganapathi; VEERAMACHANENI, Kalyan; OSADCIW, Lisa. A particle swarm optimization based multilateration algorithm for UWB sensor network. In: *2009 Canadian Conference on Electrical and Computer Engineering*. 2009, pp. 950–953. Available from DOI: 10.1109/CCECE.2009.5090268.

63. KABAL, Peter. *TSP Speech Database*. 2018. McGill.

64. BREDIN, Hervé; YIN, Ruiqing; CORIA, Juan Manuel; GELLY, Gregory; KORSHUNOV, Pavel; LAVECHIN, Marvin; FUSTES, Diego; TITEUX, Hadrien; BOUAZIZ, Wassim; GILL, Marie-Philippe. pyannote.audio: neural building blocks for speaker diarization. In: *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2020.

65. PASZKE, Adam; GROSS, Sam; MASSA, Francisco; LERER, Adam; BRADBURY, James; CHANAN, Gregory; KILLEEN, Trevor; LIN, Zeming; GIMELSHEIN, Natalia; ANTIGA, Luca; DESMAISON, Alban; KÖPF, Andreas; YANG, Edward; DEVITO, Zach; RAISON, Martin; TEJANI, Alykhan; CHILAMKURTHY, Sasank; STEINER, Benoit; FANG, Lu; BAI, Junjie; CHINTALA, Soumith. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. Available from arXiv: 1912.01703 [cs.LG].

66. PARK, Tae Jin; KANDA, Naoyuki; DIMITRIADIS, Dimitrios; HAN, Kyu J.; WATANABE, Shinji; NARAYANAN, Shrikanth. *A Review of Speaker Diarization: Recent Advances with Deep Learning*. 2021. Available from arXiv: 2101.09624 [eess.AS].

67. ROTH, Joseph; CHAUDHURI, Sourish; KLEJCH, Ondrej; MARVIN, Radhika; GALLAGHER, Andrew C.; KAVER, Liat; RAMASWAMY, Sharadh; STOPCZYNSKI, Arkadiusz; SCHMID, Cordelia; XI, Zhonghua; PANTOFARU, Caroline. AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection. *CoRR*. 2019, vol. abs/1901.01342. Available from arXiv: 1901.01342.

68. Google, [n.d.]. Available also from: `https://research.google.com/ava/challenge.html`. Accessed on 20. 2, 2023.

69. ZHANG, Yuanhang; LIANG, Susan; YANG, Shuang; SHAN, Shiguang. *UniCon+: ICTCAS-UCAS Submission to the AVA-ActiveSpeaker Task at ActivityNet Challenge 2022*. 2022. Available from arXiv: `2206.10861 [cs.CV]`.

70. KÖPÜKLÜ, Okan; TASESKA, Maja; RIGOLL, Gerhard. How to Design a Three-Stage Architecture for Audio-Visual Active Speaker Detection in the Wild. *arXiv preprint arXiv:2106.03932*. 2021.