THE UNIVERSITY *of* EDINBURGH

BEng Electronics and Electrical Engineering

Deep Neural Networks for Direction of Arrival Estimation using Circular Microphone Arrays

by

Ibrahim Aldarmaki
Matriculation Number : S2107086

May 2023

Mission Statement

Project Title: Deep Neural Networks for Microphone Array Beamforming

Student : Ibrahim Aldarmaki

Academic Supervisor : Dr Mehrdad Yaghoobi

Project Definition: An important problem facing engineers nowadays is to locate a person speaking in the background of many other people or with strong background noise. Hence, Direction of Arrival (DOA) estimation and beamforming are necessary. The aim of this project is to explore different methods used for DOA estimation and beamforming. Different deep learning/machine learning and signal processing methods will be inspected. For beamforming, more complex algorithms will be investigated. Simulated audio data and real audio data will be used to train different models/neural networks using Python. The target of this project is to accurately estimate the DOA using a circular microphone array and develop beamforming on the same microphone configuration.

Preparatory Tasks:

- Building a background in different methods used for DOA estimation and beamforming.
- Building a background in deep learning/machine learning.
- Exploring relevant python libraries.
- Exploring relevant data science/signal processing topics.

Main Tasks:

- Investigating deep learning/machine learning methods used in DOA estimation and beamforming.

- Investigating and comparing different models for DOA estimation and beamforming.
- Using simulated and real data for training the deep learning models.
- Collect data by an off the shelf circular microphone array for fine tuning the model.

Scope for Extension:

- Deploying the models on a microphone array connected to a Raspberry Pi Board for real-time DOA Estimation and Beamforming.

Background Knowledge:

- Deep Learning.
- Signal Processing.
- Python.

Resources:

- Graphics Processing Unit (GPU).
- Datasets.
- Simulators.

Abstract

This thesis introduces a Direction of Arrival (DOA) estimation algorithm using a Deep Neural Network (DNN) for localizing a sound source in complex environments. A comprehensive investigation was conducted, employing a circular microphone array and a python-based package to simulate various acoustic scenarios. Speech signals were recorded using the microphone array, and a Generalized Cross Correlation Phase Transform (GCC-PHAT) matrix was computed for all possible signal combinations. To enhance the GCC-PHAT matrix, spectral estimation techniques were employed. Subsequently, multiple Neural Networks (NNs) were trained using these GCC-PHAT matrices as input, with a Multilayer Perceptron (MLP) identified as the most efficient model for estimating the DOA. Comparative evaluations against the MUltiple SIgnal Classification (MUSIC) algorithm and the Steered-Response Power Phase Transform (SRP-PHAT) algorithm demonstrated superior performance by the MLP. Practical impacts of this project and potential strategies for its direct implementation were illustrated. Finally, avenues for future research in this field are outlined.

Declaration of Originality

I declare that this thesis is my

original work except where stated.

Ibrahim Aldarmaki

Statement of Achievement

Throughout the course of this project, I have achieved significant milestones in my understanding and application of diverse signal processing and deep learning techniques, encompassing both foundational and cutting-edge approaches. The project served as a rich learning experience, allowing me to explore various methods and algorithms that form the bedrock of signal processing. I gained comprehensive knowledge and proficiency in utilizing simulation tools, enabling accurate modeling of complex acoustic environments and providing valuable insights into sound propagation and recording phenomena. Moreover, this project enhanced my programming skills, as I successfully implemented and optimized data pipelines, ensuring efficient data processing and analysis. By connecting multiple pipelines and accommodating system changes, I gained a deeper understanding of the intricacies involved in structuring robust and adaptable systems.

In addition to advancing my technical skills, this project provided me with invaluable insights into the practical application of deep neural networks. Through hands-on experimentation and analysis, I gained a nuanced understanding of the strengths and limitations of different neural network architectures and learned to discern the most suitable network layers for specific tasks. This experience reinforced the importance of considering various factors, such as data characteristics, model complexity, and computational efficiency, when designing and implementing deep learning systems. Moreover, working on a real-world project allowed me to witness the tangible impact of deep neural networks in sound source localization, surpassing the performance of traditional algorithms. This project has significantly contributed to my knowledge and expertise in signal processing and deep learning, equipping me with valuable skills and insights for future research and development endeavors in this field.

Contents

List of Figures

List of Tables

Glossary

AI Artificial Intelligence.

ANN Artificial Neural Networks.

CC Cross Correlation.

CNN Convolutional Neural Network.

DL Deep Learning.

DNN Deep Neural Network.

DOA Direction of Arrival.

GCC Generalized Cross Correlation.

GCC-PHAT Generalized Cross Correlation Phase Transform.

HL Hidden Layer.

MLP Multilayer Perceptron.

MUSIC MUltiple SIgnal Classification.

NN Neural Network.

PHAT Phase Transform.

PRA PyRoomAcoustics.

ReLU Rectified Linear Unit.

RIR Room Impulse Response.

SNR Signal-to-Noise Ratio.

SRP-PHAT Steered-Response Power Phase Transform.

STFT Short-Time Fourier Transform.

TDOA Time Difference of Arrival.

Chapter 1

Introduction

Sensor arrays are an essential component in various technological applications and have witnessed widespread adoption across diverse fields. These arrays consist of multiple individual sensors, which work collectively to capture and analyze data from the surrounding environment. One prominent example is microphone arrays, where multiple microphones are strategically arranged to capture sound from different directions[1]. The utilization of sensor arrays offers numerous advantages, including improved spatial resolution, increased sensitivity, and enhanced signal processing capabilities. By leveraging the collective input from multiple sensors, these arrays enable accurate and reliable measurements, leading to more robust and efficient systems in various domains.

The problem of Direction of Arrival (DOA) estimation has been a topic of extensive research in signal processing, and it has found numerous applications in various fields, including acoustics[2], radar[3], and wireless communication systems[4]. DOA estimation is a signal processing technique that can be used in microphone arrays to determine the direction from which sound signals are arriving. It is an important tool for spatial audio analysis and has numerous applications in fields such as teleconferencing[5; 6], surveillance[7], speech recognition[8; 9], and automatic camera steering [10]. Humans can discern the direction of sound by utilizing both ears and instinctively combining the diverse signals they receive[11]. Similarly, DOA estimation algorithms examine signals from an array of microphones and analyze the discrepancies in the timing and amplitude of sound waves to determine the source of the sound. Accurate DOA estimation can greatly enhance the performance of audio processing systems, allowing for better noise reduction, signal enhancement, and source localization. Therefore, DOA estimation is a crucial technology

for improving the quality and effectiveness of audio-based applications.[2]

One of the key applications of DOA estimation is in selective auditory attention, which is the ability to focus on one sound source while suppressing all others in the surrounding environment. Selective auditory attention has important implications for a range of applications such as speech recognition[8], hearing aid design[12], and speaker localization[3]. In this context, DOA estimation plays a critical role in providing the necessary spatial information to enable selective auditory attention. Therefore, this thesis aims to investigate the state-of-the-art methods in DOA estimation[13].

1.1 Motivation

The estimation of DOA is an essential task in many signal processing applications. Nevertheless, numerous DOA estimation algorithms necessitate the inclusion of additional algorithms, thereby increasing the number of stages prior to obtaining a DOA estimate. Consequently, if an error occurs in one stage, it can greatly impact the accuracy of the final DOA estimation. Furthermore, traditional DOA estimation algorithms heavily rely on a multitude of assumptions that simplify the problem. However, these assumptions may not hold valid in various real-life scenarios, potentially limiting their applicability and accuracy[14; 15]. Therefore, the search for accurate DOA estimation algorithms continues. Fortunately, the emergence of Deep Learning (DL) has revolutionized the field of signal processing by allowing the extraction of complex and non-linear patterns directly from data. Deep Neural Networks (DNNs) can provide end-to-end models that map input data to the output target, eliminating the need for multiple intermediate estimation algorithms. As a result, researchers have turned to DNNs for DOA estimation, which promises a more robust and efficient solution to this problem.

The present study is driven by a genuine inquiry: Can DNNs be effectively employed for estimating the DOA of speech signals? Furthermore, if such utilization is feasible, what level of precision can be achieved? This research endeavor is specifically designed to comprehensively investigate these pivotal queries and provide well-substantiated responses.

1.2 Aim & Objective

This research endeavor aims to explore the feasibility of utilizing DNNs for the estimation of DOA in circular microphone arrays. This investigation will involve the use of simulation tools to acquire realistic data, as DNNs' learning processes heavily depend on data for training. Additionally, feature extraction techniques will be scrutinized to accentuate significant features in the data. Multiple DL models will be evaluated and compared to identify an optimal model. Ultimately, a data pipeline will be proposed for DOA estimation in circular microphone arrays via a DNN.

1.3 Thesis Overview

In chapter 2, a comprehensive background will be presented, covering the fundamental knowledge necessary for this thesis. Chapter 3 will delve into the employed methodology, detailing the step-by-step process that ultimately leads to the proposal of an efficient data pipeline for achieving optimal DOA estimation. The subsequent chapter, chapter 4, will focus on the results and analysis, extensively examining the conducted experiments and the associated trade-offs made to attain the most effective model. Chapter 5 will shed light on the impact and potential exploitation, providing an overview of the project's current stage within the development cycle and exploring its potential implications. Finally, chapter 6 will serve as the conclusion chapter, highlighting the key findings, summarizing the thesis, and discussing potential future avenues of research and development for this project.

Chapter 2

Background

This chapter sets the foundation for the forthcoming thesis by outlining the context of the primary research themes. Initially, an overview of the algorithms employed for DOA estimation is presented. The discussion also touches on power spectral estimation methods, with a particular emphasis on Bartlett's approach for spectral estimation. Following that, a background in DL is presented, with an emphasis on the DL methods deployed in previous studies for DOA estimation.

2.1 DOA Estimation Algorithms

There are numerous signal processing techniques available for estimating the direction of arrival of a speech signal from a microphone array. These methods have been studied and refined over the years, leading to significant advancements in the field of speech signal processing. Certain DOA estimation algorithms hinge on the detection of the Time Difference of Arrival (TDOA), which concentrates on the difference in signal arrival time between two receivers. Another popular method is beamforming, which involves applying filter weights to the signals captured by each microphone in a directional pattern. An alternative widely used technique is the sub-space method, which leverages covariance matrices to estimate the DOA.

2.1.1 Time Difference of Arrival Methods:

TDOA is a technique used to measure the difference in arrival time of the signal at two or more receivers. TDOA can be used to estimate the position of a sound source in a room or space. The equations below provide a model for signals received by two microphones, $x_1(t)$ and $x_2(t)$, from a

source transmitting a speech signal, $s(t)$, that travels through a noisy channel in free space, without taking multipath effect into account.

$$x_1(t) = A_1 s(t - t_1) + n_1(t) \tag{2.1}$$

$$x_2(t) = A_2 s(t - t_2) + n_2(t) \tag{2.2}$$

where A_1 and A_2 represent the amplitude attenuation, t_1 and t_2 represent the propagation delay, and n_1 and n_2 represent the noise. The noise is assumed to be stationary, random, and uncorrelated with the signal $s(t)$. The two noise terms, $n_1(t)$ and $n_2(t)$, are also assumed to be uncorrelated. Assuming that $t_1 < t_2$, eq. (2.3) can be used to represent the TDOA based on the delay terms.

$$TDOA = t_1 - t_2 \tag{2.3}$$

An alternative definition of the TDOA that expresses it in relation to the distance between the source and each microphone is demonstrated in eq. (2.4). This is also visually represented in fig. 2.1.

$$TDOA = \frac{d_1}{v} - \frac{d_2}{v} \tag{2.4}$$

where d_1 and d_2 represent the distance between the source and microphone, and v denotes the signal's speed in the give medium.

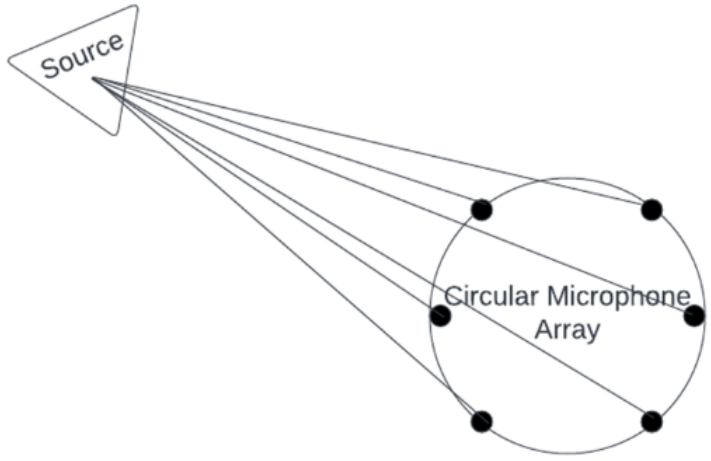


Figure 2.1: Illustration of the different line-of-sight paths of an acoustic signal travelling through a medium from a sound source to multiple microphones in a circular microphone array.

Generalized Cross Correlation:

In order to accurately estimate the TDOA, a method that is impervious to noise and interference must be employed. The Generalized Cross Correlation (GCC) method is a reliable choice, given that noise is typically assumed to be uncorrelated [16]. To further improve the Signal-to-Noise Ratio (SNR), a weighting function was introduced in the GCC calculation [17], thereby modifying the definition of GCC to conform to eq. (2.5).

$$R_{x_1,x_2} = \mathscr{F}^{-1}\Big\{\psi_g(f)G_{x_1,x_2}(f)\Big\} \tag{2.5}$$

where R_{x_1,x_2} represents the GCC between signal x_1 and signal x_2, $\mathscr{F}^{-1}$ represents the inverse Fourier transform, f represents the frequency, $\psi_g(f)$ represents the weighting function, and $G_{x_1,x_2}(f)$ represents the cross spectral density between signal x_1 and signal x_2.

The Phase Transform (PHAT) weighting function, as detailed in eq. (2.6), was identified as a robust method for estimating TDOA due to its peak value occurring at the TDOA value [17].

$$\psi_g(f) = \Big|\frac{1}{G_{x1,x2}(f)}\Big| \tag{2.6}$$

$$\hat{TDOA} = arg\,\underset{\tau}{max}\,\psi_{x_1,x_2}(\tau) \tag{2.7}$$

The estimation of TDOA for various combinations of microphones, in conjunction with the geometry of the microphone array, provides an opportunity to estimate the DOA of the signal.

Hyperbolic Position Location Estimation:

Hyperbolic Position Location Estimation is a technique that is usually used in wireless communications to determines the position of a mobile device. In order to perform Hyperbolic Position Location Estimation, a two-stage approach is utilized. Firstly, TDOA between transmitters is estimated via TDOA estimation techniques. These estimated TDOAs are then translated into range difference measurements between base stations, which results in a set of nonlinear hyperbolic equations. The second stage involves the utilization of efficient algorithms to achieve a clear solution to these nonlinear hyperbolic equations[18]. The two dimensional Euclidean distance

equation is used to define the distance, as shown in eq. (2.8), between the i-th source and receiver.

$$R_i = \sqrt{(X_i - x)^2 + (Y_i - y)^2} \tag{2.8}$$

where the source location is given by (x, y), and the coordinates of the i-th source is given by (X_i, Y_i). Compared to other methods, Chan's method [19] outperforms them significantly in terms of its ability to accurately locate multiple acoustic sources within a room[20]. Let

$$\begin{aligned} R_{i,1} &= R_i - R_1 \\ X_{i,1} &= X_i - X_1 \\ Y_{i,1} &= Y_i - Y_1 \end{aligned} \tag{2.9}$$

By applying Chan's method [19] to the scenario where three receivers are present, two TDOA values,x and y, are generated, which can be solved with respect to R1. The resulting solution is represented by eq. (2.10).

$$\begin{bmatrix} x \\ y \end{bmatrix} = -\begin{bmatrix} X_{2,1} & Y_{2,1} \\ X_{3,1} & Y_{3,1} \end{bmatrix}^{-1} \times \left(\begin{bmatrix} R_{2,1} \\ R_{3,1} \end{bmatrix} R_1 + \frac{1}{2} \begin{bmatrix} R_{2,1}^2 - K_2 + K_1 \\ R_{3,1}^2 - K_3 + K_1 \end{bmatrix} \right) \tag{2.10}$$

where

$$\begin{aligned} K_1 &= X_1^2 + Y_1^2 \\ K_2 &= X_2^2 + Y_2^2 \\ K_3 &= X_3^2 + Y_3^2 \end{aligned} \tag{2.11}$$

Substituting eq. (2.10) into eq. (2.8) with $i = 1$ results in a quadratic equation in terms of R_1. Solving the quadratic equation results in a positive root that when substituted back in eq. (2.10) gives the solution to the equation [19] which represents an estimation of the source location.

2.1.2 Beamforming Method:

Beamforming, also known as spatial filtering, is a widely-used signal processing method utilized in sensor arrays to facilitate directional transmission or reception of signals. The technique involves a combination of elements in an antenna array to achieve constructive interference for signals arriving at specific angles and destructive interference for signals arriving at other angles[21].

Beamforming can be used to estimate the DOA. It involves applying filter weights to the signals collected by each microphone in the array in a directional pattern. The DOA is determined based on the direction in which the filter weights are most effective in reconstructing the desired sound.

Delay-and-Sum Beamformer:

The Delay-and-Sum beamforming technique, as shown in fig. 2.2, involves the application of a delay and an amplitude weight to the signal captured by each sensor in an array, followed by summation of the resulting signals. This process enables steering the array's direction of observation towards the source by adjusting the delays. Additionally, the weights assigned to each sensor serve as gain factors that enhance the shape and diminish the sidelobe levels of the beam obtained from the received signals [22]. The output signal of the delay-and-sum beamformer can be computed using eq. (2.12).

$$z(t) = \sum_{m=1}^{M} w_m \cdot y_m(t - \Delta_m) \tag{2.12}$$

where $z(t)$ represents the output signal from the delay-and-sum beamformer, M represents the number of sensors, w_m represents the weight for the $m-th$ sensor, y_m represents the received signal by the $m-th$ sensor, and Δ_m represents the delay for the $m-th$ sensor. The delay-and-sum beamforming technique can effectively determine the DOA of a signal by measuring the strength of the signal at all possible arrival angles and selecting the angle that results in the highest power peak. This methodology is commonly known as the Steered-Response Power Phase Transform (SRP-PHAT) algorithm [23].

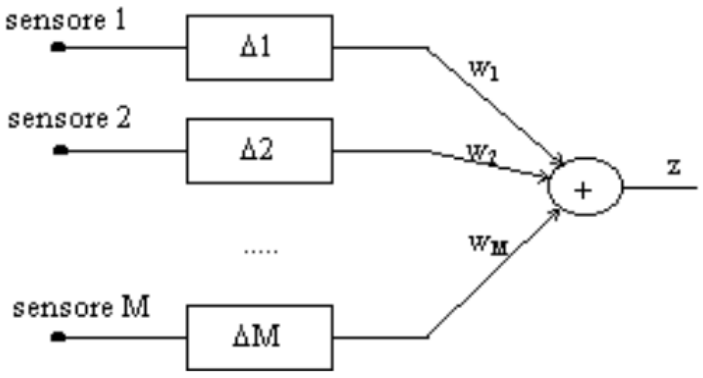


Figure 2.2: Visualization of the delay-and-sum beamformer.

Capon's Minimum Variance Beamformer:

The Capon's Minimum Variance method aims to reduce the effects of undesired interference in the output power eq. (2.14) of a sensor array, while simultaneously preserving a consistent gain, typically set to unity, in the desired direction[24]. The output of the sensor array can be expressed mathematically as shown in eq. (2.13).

$$y = w^H x \tag{2.13}$$

where y represents the output, w represents the weight matrix, H represents the Hermitian transpose, and x represents the signal received by the sensors.

$$P_{capon}(\theta) = \frac{1}{a^H(\theta)R^{-1}a(\theta)} \tag{2.14}$$

where P represents the output power, and R represents the covariance matrix.

The optimal weights are chosen with the goal of minimizing the output power of the system, while simultaneously preserving a unity gain in the desired direction. This relationship can be mathematically represented by eq. (2.15).

$$\min_w\{w^H R w\} \qquad subject \quad to \quad w^H a(\theta) = 1 \tag{2.15}$$

where $a(\theta)$ represents the steering vector.

The optimum weights are defined as eq. (2.16).

$$W = \frac{R^{-1}a(\theta)}{a^H(\theta)R^{-1}a(\theta)} \tag{2.16}$$

The DOA of the desired signal is estimated from the location of the maximum output power of the beamformer, which corresponds to the direction of the desired signal [22].

2.1.3 Subspace Methods:

Subspace-based techniques are commonly used to estimate the DOA of signals by exploiting the eigenspaces of the covariance matrix. These methods typically consist of three main steps: (1) estimation of the covariance matrix, (2) computation of the eigenvectors based on the covariance estimates, and (3) estimation of the DOA, which is dependent on the specific subspace method

used for DOA estimation [15].

MUSIC Algorithm:

MUltiple SIgnal Classification (MUSIC) is a type of subspace DOA estimation method that provides an estimate of both the number of signals and their DOA. It is essential that the number of signal sources is fewer than the number of receivers, which is a common assumption in many DOA estimation algorithms. The MUSIC algorithm operates by decomposing the covariance matrix into two mutually orthogonal subspaces: the signal subspace and the noise subspace [25]. The algorithm then generates the MUSIC pseudospectrum, which can be represented mathematically by eq. (2.17).

$$P_{MUSIC}(\theta) = \frac{1}{a(\theta)^H U_n U_n^H a(\theta)} \tag{2.17}$$

where $a(\theta)$ represents the steering vector, and U_n represents the noise eigenvectors matrix. The estimated DOA is indicated by the peak(s) observed in the pseudospectrum.

2.2 Power Spectral Estimation

In contemporary signal processing algorithms, the utilization of essential statistical metrics, such as probability density function, autocorrelation function, joint probability function, or power density function, is commonplace. Regrettably, these metrics often elude practical signals, posing a significant challenge. Consequently, spectral estimation emerges as a pivotal pursuit, aiming to derive an accurate estimation of a signal's power spectral density through a sequential analysis of time samples[26]. In this section, we focus our attention on the elucidation of Bartlett's method, alternatively known as the periodogram method, within the context of spectral estimation.

2.2.1 Bartlett's Method

Bartlett's method involves segmenting a lengthy signal sequence into K consecutive non-overlapping blocks of equal length M. Within each block, periodograms are computed, which are essentially the Fourier Transforms of the signal's correlation. This correlation-based Fourier Transform holds significance as it embodies the power spectrum, as stated by the Weiner-Khintchin theorem [27]. To obtain a comprehensive Bartlett power spectral estimate, the resulting periodograms are subsequently averaged together. This methodology provides a reliable means of characterizing the

power distribution across the signal's frequency domain[28].

The periodogram can be calculated using eq. (2.18)

$$P^{(i)}_{x_1,x_2}(f) = \frac{1}{M}\mathscr{F}\{r^{(i)}_{x_1,x_2}\} \tag{2.18}$$

where $P^{(i)}$ represents the periodogram of the i-th block, $\mathscr{F}$ represents the Fourier transform, and $r^{(i)}_{x_1,x_2}$ represents the cross-correlation function between signal x_1 and signal x_2. The Bartlett power spectral estimate can be calculated using eq. (2.19)

$$P^{B}_{x_1,x_2}(f) = \frac{1}{K}\sum_{i=0}^{K-1} P^{(i)}_{x_1,x_2}(f) \tag{2.19}$$

where $P^{B}_{x_1,x_2}(f)$ represents the Bartlett power spectral estimate, and K represents the number of blocks.

2.3 Deep Learning:

DL is a subset of machine learning that uses Artificial Neural Networkss (ANNs) to process complex datasets. The ANNs are dynamic systems that are nonlinear and adaptive in nature, comprising a vast number of interconnected neurons. In DL, the ANNs consist of multiple layers, allowing the network to learn complex features and patterns from the input data. The DL models are designed to learn from large datasets and can be used for a variety of applications, including speech recognition[29], natural language processing[30], image recognition[31], and image segmentation[32].

Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) are two popular DL architectures used in speech applications. MLPs are feedforward Neural Networks (NNs) that perform non-linear transformations on the input signal, and they are well-suited for solving classification problems. CNNs consist of multiple layers of convolutional filters and pooling layers that learn local patterns in the input signal. CNNs have shown excellent performance in various speech-related tasks, such as speech recognition, speaker identification, and voice activity detection. DL methods have the potential to improve the accuracy and efficiency of speech signal processing, leading to better performance in speech-related applications[33].

2.3.1 Deep Learning for Direction of Arrival Estimation:

Recent work has shown promising results in DOA estimation using DL architectures. For example, this paper [34] proposed a CNN-based method for DOA estimation. The authors showed that their method outperformed traditional DOA estimation methods with more robustness and lower computational cost. Similarly, this paper [35] presented an MLP-based method for DOA estimation by extracting features from GCC vectors. The authors showed that their method achieved state-of-the-art performance on both simulated and real data.

Chapter 3

Methodology:

In the preceding chapter, we provided an overview of the background and context of DOA estimation. In this current chapter, we delve into the methodology that has been employed and devised for this particular project. Our focus will be on elucidating the data collection approach, followed by an exploration of the feature extraction technique. Furthermore, we delve into the DL methods that have been employed in this study. To culminate, we propose a data pipeline that harnesses the power of a DNN for accurately estimating DOA.

3.1 Data Collection:

In order to effectively learn complex patterns, DL methods require a significant amount of data. To ensure that this data is relevant to our particular setting, it is important that it be collected in a manner similar to that of the ReSpeaker 6-Mic Circular Array, which is an off-the-shelf circular microphone array consisting of six microphones. The sampling frequency was chosen to be 16kHz, which is four times the frequency bandwidth of speech signals, in order to accurately sample speech signals and apply digital delays without sacrificing data [7]. Despite examining open-source data, we were unable to locate any datasets that meet these specifications, necessitating the collection of data through alternative means.

3.1.1 Sound Recording:

In order to collect a large number of sound samples, a software must be used for sound recording. PyRoomAcoustics (PRA) is a python-based open-source package that facilitates the simulation of

room-like environments, microphone arrays, and sound sources in order to record acoustic signals. The package offers efficient estimation of Room Impulse Response (RIR), as depicted in fig. 3.2, and enables the propagation of acoustic signals to be simulated between sources and receivers. The environment can be represented through the specification of various parameters [36]. To collect data, PRA was employed to simulate rooms according to the specifications outlined in table A.1, with the source location specified as shown in table A.2. Fig. 3.1 depicts the appearance of the simulated room.

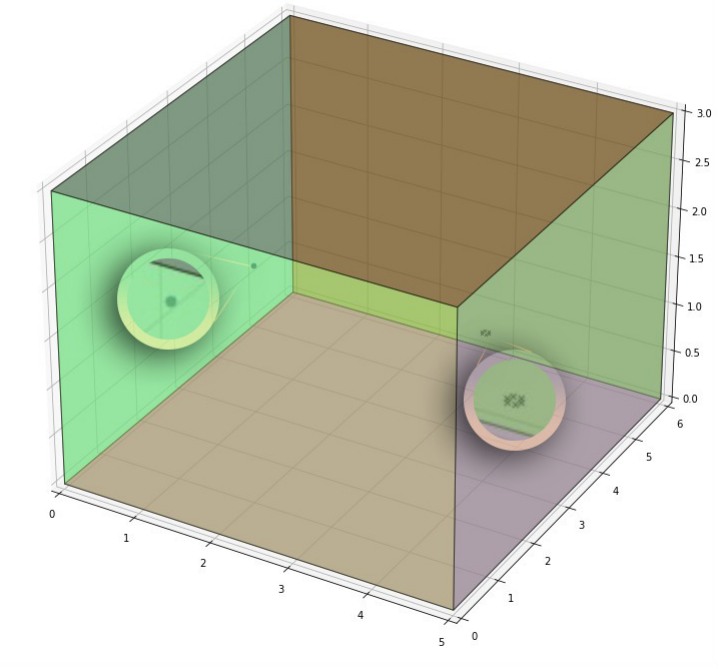


Figure 3.1: Illustration of a room that was created using PRA.

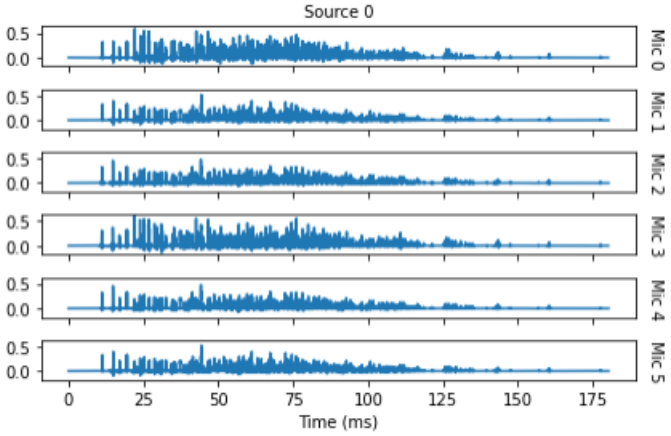


Figure 3.2: Visualization of a RIR that was estimated for an environment using PRA.

There are two common methods that are widely used to simulate acoustic signals, which are: (1) ray tracing method (2) image source model. Ray tracing assumes that the movement of sound within a given environment occurs in the form of ”rays”. This method can provide a detailed and accurate simulation of the acoustics of a space, taking into account factors such as reflections, diffraction, and absorption. On the other hand, the image source model is a simpler approach that assumes that sound waves propagate directly from a sound source to a listener, and that any reflections are accounted for by virtual ”image sources” that are created by the reflections of the original sound source. This method is less computationally intensive than ray tracing. The ray method provides a more realistic representation of acoustic signals at the cost of being more computationally expensive [37]. A hybrid method, that is available in PRA, was used to efficiently simulate the speech signals using PRA [38]. Fig 3.3 shows a speech signal that was recorded by one of the microphones in the array using the hybrid method.

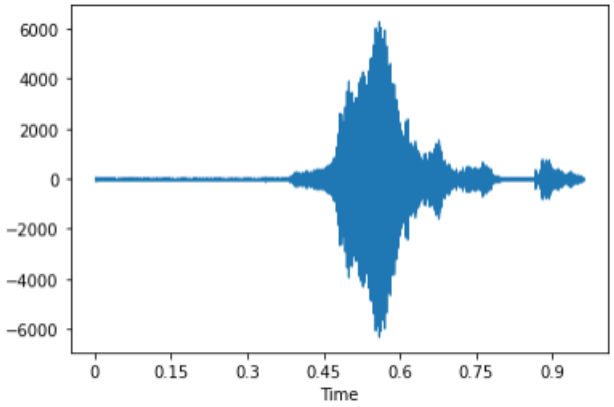


Figure 3.3: Representation of a speech signal that was recorded using PRA.

3.2 Feature Extraction:

In order to optimize DL models' ability to learn from data, it is crucial to represent the data in a manner that accentuates pertinent features. This approach can lead to enhanced performance without the need for deeper NNs. Two techniques for data representation include utilizing raw signals from microphone arrays, as demonstrated in [39], and utilizing Fourier transforms to compute the Short-Time Fourier Transform (STFT), as shown in [40], of time domain data as features. However, these techniques do not effectively spotlight critical information in the recorded signals for DOA estimation. An alternative approach, explained in chapter 1, is to employ the GCC be-

tween multiple microphones. This method emphasizes TDOA information, which is relevant to DOA estimation. In addition, this paper [17] proposes using GCC-PHAT, which is more robust and focuses primarily on TDOA. Consequently, the GCC-PHAT method will be utilized for feature extraction. Fig 3.4 shows a comparison between a Cross Correlation (CC), that has a its peak with a magnitude of 1.0, and GCC-PHAT for the same pair of signals. Notably, the GCC-PHAT exhibits a more precise representation of TDOA as evidenced by a distinct peak at the TDOA value. Conversely, while the CC also exhibits a peak at the same value, it is accompanied by multiple peaks throughout the sequence, which dilutes the emphasis on the TDOA value. This visual analysis underscores the superior accuracy of the GCC-PHAT method in accurately capturing the TDOA.

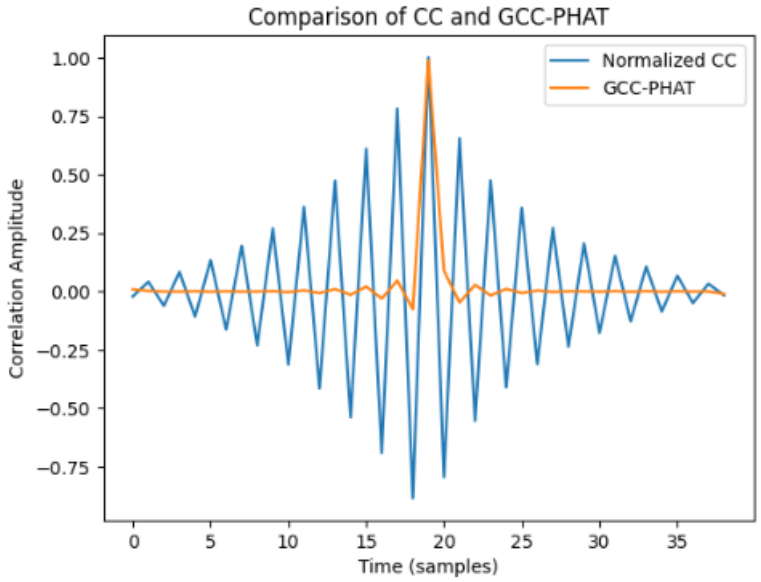


Figure 3.4: Visual representation of a comparison between normalized CC and GCC-PHAT.

3.2.1 GCC-PHAT Matrices:

The computation of the GCC-PHAT between a pair of microphones produces a correlation sequence, which can be represented as a vector. By computing all unique GCC-PHAT combinations between the six microphones, a total of 15 unique vectors can be obtained. Arranging these vectors in a matrix form results in a GCC-PHAT matrix, as shown in fig. 3.5, that can be interpreted as an image. This image, as illustrated in fig. 3.6, can be employed as a feature to represent the signals captured by the microphones. The length of the GCC-PHAT vector can be determined using eq. (3.1).

$$L_{GCC-PHAT} = 2 * N - 1 \tag{3.1}$$

where $L_{GCC-PHAT}$ represents the length of the GCC-PHAT sequence, and N represents the signal's length.

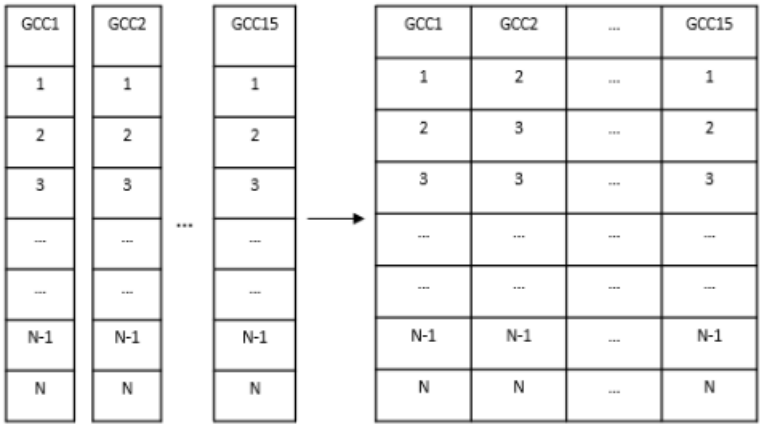


Figure 3.5: Illustration of stacking GCC-PHAT vectors to create a matrix.

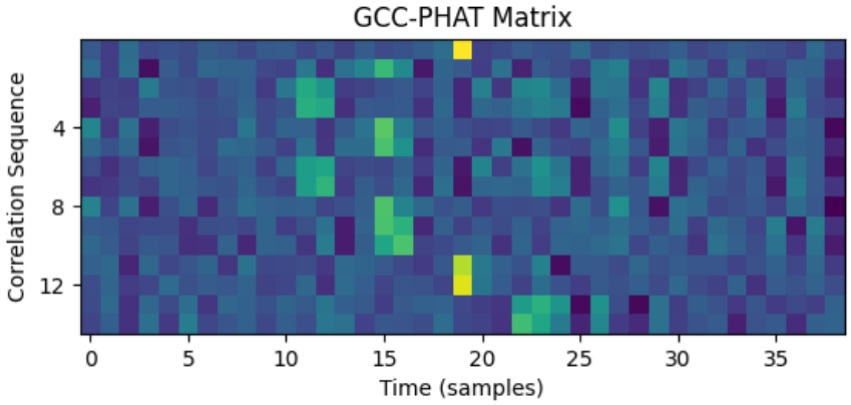


Figure 3.6: Illustration of representing the GCC-PHAT matrix as an image

3.2.2 Cross Spectral Density Estimation:

The GCC-PHAT method involves the calculation of cross spectral density between two signals. To enhance the accuracy of this estimation, cross spectral density estimation methods can be applied. However, these methods require the assumption that the process is ergodic. In the case of speech signals, this assumption holds true for time frames of up to 20ms [41]. As a result, spectral estimation methods can be used to improve the representation of the GCC-PHAT sequence. Bartlett's method, as discussed in section 2.2.1, can be employed to estimate the sequences of GCC-PHAT,

using the parameters specified in table 3.1.

Table 3.1: Parameters used for spectral estimation method.

Parameter	Value
Number of Blocks	10
Number of Samples per Block	20

3.3 Deep Learning:

To effectively harness the potential of data-driven learning, the utilization of DL techniques is crucial. However, the initial step necessitates framing the problem at hand within the context of DL. Subsequently, employing a DNN is imperative to approximate the DOA based on the provided input data. Moreover, establishing appropriate metrics to proficiently assess the performance of DL models holds significant importance.

3.3.1 Problem Definition:

This study examines the problem of DOA estimation using two distinct approaches. The first approach is to treat it as a classification problem, with the predicted output being one of a finite set of pre-defined classes. The second approach is to treat it as a regression problem, with the predicted output being a continuous value. The input for both methods is the GCC-PHAT Matrix discussed in section 3.1, and the data is labeled according to the chosen approach.

Classification:

In the classification task, the DOA is discretized into a predetermined number of classes, and input matrices are labeled accordingly. Sparse-categorical representation is utilized instead of one-hot encoded categories to minimize memory space requirements, as shown in table A.3. In the case of sparse encoding, categories are represented by a single numerical value, whereas one-hot encoding utilizes binary values where 0 indicates that the data does not belong to the category and 1 signifies that the data point belongs to the respective category. The output layer of the NN is composed of the Softmax function to represent the output as probabilities of the various DOA classes. The predicted class is the one with the highest probability. The Softmax function is described in eq. (3.2).

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3.2}$$

where σ represents the Softmax function, $\vec{z}$ represents the input vector to the Softmax function, z_i represents the i-th element of the input vector, and K represents the total number of elements in the input vector.

Regression:

In the regression method, the NN generates two outputs that represent a unit vector on the x-y (2 dimensional) plane pointing towards the source projection. The NN aims to estimate the two components of this unit vector, which is a continuous value ranging from -1 to 1. Notably, the network output is not constrained to have a magnitude of 1. The estimated vector components are subsequently utilized to compute the azimuth angle of arrival. To enable comparison with the classification method, a customized evaluation function is developed. This function divides the plane into classes analogous to the classification method and calculates the accuracy based on the computed angle. This regression approach is inspired by the hyperbolic position location estimation method that was discussed in section 2.1.1. The advantage of using the regression method is that the DOA is a continuous value, and representing it using continuous output is desirable. The use of unit vectors is preferred over the angle of arrival since the latter does not accurately represent angles. For instance, the difference between 0 degrees and 359 degrees is only 1 degree, but representing them as such yields a difference of 359, which may adversely affect the DNN's performance.

3.3.2 DNN Architictures:

Various DNN architectures are used to estimate the DOA. Firstly, a MLP architecture inspired by the methodology presented in [35] is utilized. The parameters of the selected MLP are detailed in table 3.2. Secondly, a CNN inspired by the approach adopted in [34] is implemented. The parameters of the selected CNN are summarized in table 3.3. The NN architectures employ the Rectified Linear Unit (ReLU) activation function for each Hidden Layer (HL), as it is both computationally efficient and robust, as demonstrated in [42].

Table 3.2: Parameters of the selected MLP-based DNN

Layer	Values
1^{st} HL	128 nodes

Table 3.3: Parameters of the selected CNN-based DNN

Layer	Value
Convolutional Layer	6 filters
Flatten Layer	-
Dense Layer	84 nodes

3.3.3 Models Evaluation:

Evaluating various DL models is essential to determine their relative performance. The DOA estimation task at hand can be categorized as a classification problem, and as mentioned earlier, the regression approach can also be utilized to predict categorical classes after computing the angle of arrival. Hence, evaluation of the regression approach can be carried out in a manner similar to classification tasks. Two methods for evaluating the DL models will be employed: accuracy and confusion matrix.

Accuracy:

In classification tasks, assessing the effectiveness of a DL model is commonly done by measuring its accuracy. This metric quantifies the percentage of accurate predictions made by the model. To be more specific, accuracy is computed by dividing the number of correctly classified instances by the total number of instances.

Confusion Matrix:

In order to assess the accuracy of a classification model, a confusion matrix can be used. This table compares the predicted and actual classifications of a model by mapping them onto rows and columns, respectively. Each cell of the matrix represents the number of instances predicted to belong to a particular class, with the actual class of those same instances given by the corresponding row. An illustration of a $2x2$ confusion matrix is depicted in fig. 3.7.

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Figure 3.7: Visualization of a 2x2 confusion matrix for a binary classification model

3.4 Data Pipeline:

Upon comprehensive evaluation of the methodology using diverse parameters and architectures, as extensively elaborated in chapter 4, a well-defined data pipeline emerges. Figure 3.8 showcases an optimal data pipeline, offering an efficient framework to estimate the DOA for speech signals using a circular microphone through the utilization of a MLP. Figure B.1 illustrates the identical data pipeline, albeit on a larger scale.

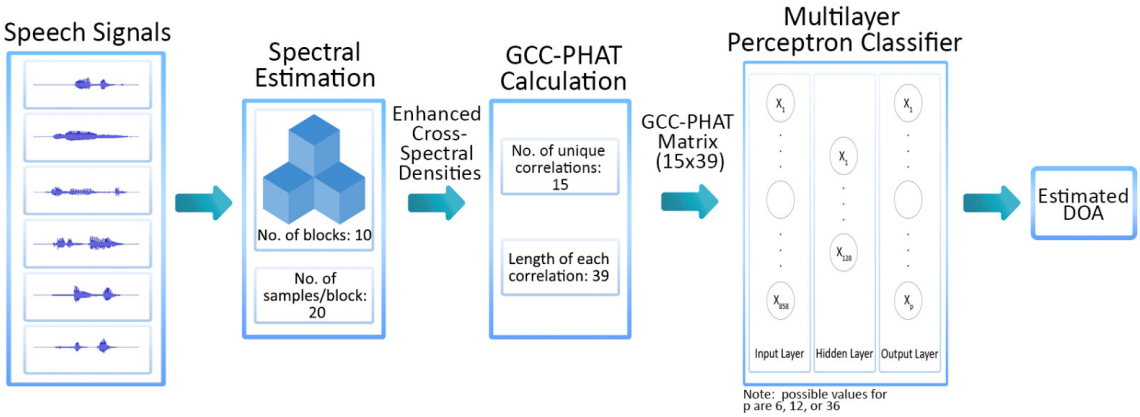


Figure 3.8: Illustration of an optimal data pipeline for DOA estimation using a MLP

Chapter 4

Results and Analysis:

The objective of this chapter is to present the outcomes of the conducted tests within this project and analyze them to elucidate the decision-making process. Initially, the chapter will delve into the procedure employed for selecting the optimal parameters for the feature extraction method. Subsequently, it will proceed to compare multiple DNNs in order to identify the most suitable one for accurately estimating the DOA. Furthermore, the performance of this chosen DNN in addressing both classification and regression tasks will be assessed. Finally, a comparative analysis will be conducted to evaluate the performance of the selected DNN in comparison to established DOA estimation algorithms.

4.1 Parameters Selection:

Before evaluating the DNN models, it is important to tune certain parameters. Two examples of such parameters are the length of the sampled signals used to compute the GCC-PHAT and the number of blocks used for spectral estimation.

In order to conduct a fair comparison, a two HL MLP with 128 nodes per HL for the 6 classes DOA estimation classification task was used. Early stopping, with parameters shown in table 4.1, was used to train the models. For this task, only accuracy will be analyzed to select the optimum parameters. Table 4.2 provides a summary of the number of samples used for training, testing, and validation sets. These samples were selected randomly from the simulated signals. The parameters used to train the models are demonstrated in table 4.3.

Table 4.1: Early stop parameters for parameters selection process.

Parameter	Value
Monitor	Validation Accuracy
Minimum Change	0.001 (0.1%)
Patience	2

Table 4.2: Number of samples used for different sets for parameter selection.

Set	Number of Samples	Percentage
Train	33800	80%
Validation	4225	10%
Test	4225	10%

Table 4.3: Parameters used for models training.

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Sparse Categorical Cross Entropy
Batch Size	32

4.1.1 GCC-PHAT Length:

In order to determine the optimal signal length for optimal performance, it is crucial to conduct testing. To achieve this goal, three different signal lengths, namely 10, 20, and 30, were tested. The results on the test set were then recorded in table 4.4. According to the table, signals of length 30 samples provided the most favorable outcomes. The rationale behind this finding is that when the sampled signal is longer, the GCC-PHAT values are more likely to accurately represent the TDOA, which, in turn, will help the DNN to more precisely estimate the DOA.

Table 4.4: Recorded accuracy on test set for different lengths of sampled signals.

Signal Length	Test Accuracy
10	79.74%
20	86.84%
30	88.38%

As previously noted, the length of the GCC-PHAT sequence is determined by eq. (3.1), which implies a trade-off, that can also be seen in table 4.4, between correlation length and accuracy. After careful evaluation, it was determined that a signal length of 20 samples would be more favorable than a signal length of 30 samples. This is because using a signal length of 20 samples

would decrease the GCC-PHAT sequence length by 20 samples with only a 1.54% decrease in performance.

4.1.2 Number of Blocks for Spectral Estimation:

Determining the optimal number of blocks for spectral estimation is crucial to accurately estimate TDOA, which ultimately affects the performance of the DNN. To identify the best value, we examined three block sizes: 1, 5, and 10. It's worth noting that the blocks are non-overlapping and do not exceed 20ms, implying that the signal can be assumed to be ergodic[41]. The outcomes on the test set are reported in table 4.5.

Table 4.5: Recorded accuracy on test set for different number of blocks used for spectral estimation.

Number of Blocks	Test Accuracy
1	86.84%
5	92.14%
10	93.33%

According to the results presented in table 4.5, the optimal number of blocks to use for spectral estimation is 10. It might be assumed by the reader that using 10 blocks for spectral estimation would increase the computational complexity, but this is not very accurate. This is because the 10 blocks will be used to estimate a single GCC-PHAT sequence, resulting in a similar length to using only one block. Additionally, these calculations will only be performed once and will not add more computations to later stages. Therefore, it has been decided that 10 blocks will be used for spectral estimation.

4.2 DNNs Selection:

There is a vast range of possibilities when it comes to structuring a deep neural network. Therefore, it is essential to explore various architectures with different depths and hyperparameters to identify a suitable and optimal model for the DOA estimation task. To achieve this, we will once again utilize the classification task to evaluate different models based on MLP and CNN. For this part, the only metric that was used to evaluate the models is accuracy. In this section, three values for the number of classes, and their respective angle ranges were considered. These values are shown

in table 4.6. Early stopping was employed using the same set of parameters outlined in table 4.1. The training parameters, as outlined in table 4.3, were also utilized in this task. Furthermore, table 4.7 details the number of data points used for training, testing and validation.

Table 4.6: Number of classes and angle ranges used for DNN selection process.

Number of Classes	Angle Range
6	60°
12	30°
36	10°

Table 4.7: Number of samples used for different sets for DNN selection.

Set	Number of Samples	Percentage
Train	114880	80%
Test	14360	10%
Validation	14360	10%

4.2.1 MLP-Based DNNs:

In order to determine the optimum MLP architecture, three kinds of MLPs with varying depths were tested, which are outlined in table 4.8. A visual depiction of the three MLP architectures is provided in fig. 4.1. The results on the test set are reported in table 4.9.

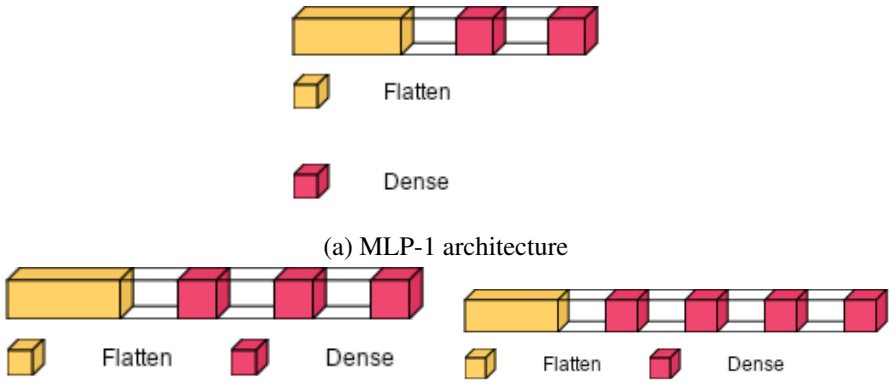
Table 4.8: Number of nodes used per HL in a MLP DNN, where - represents 0 nodes.

DNN	Nodes in HL 1	Nodes in HL 2	Nodes in HL 3
MLP-1	128	-	-
MLP-2	128	64	-
MLP-3	128	64	64

Table 4.9: Recorded accuracy for different MLP based architectures.

DNN	6 classes	12 classes	36 classes
MLP-1	95.81%	90.54%	78.29%
MLP-2	95.57%	87.98%	79.64%
MLP-3	95.77%	89.46%	78.93%

As observed from the results presented in table 4.9, the performance difference between various architectures for the same task is not significant. This indicates that in some applications, deeper NNs may not necessarily result in better performance. Therefore, a single HL MLP is selected as it performs well without the need for increased depth.



(a) MLP-1 architecture

(b) MLP-2 architecture

(c) MLP-3 architecture

Figure 4.1: Visualization of different MLP architectures used for DNNs selection process.

4.2.2 CNN-Based DNNs:

In order to determine the most effective CNN-based DNN architecture, three different architectures were tested. All architectures have convolutional layers at the beginning followed by fully connected layers. The specifications for the CNN-based DNNs are provided in section 4.2.2. Table 4.11a, table 4.11b, and table 4.11c show the specifications for the first, second, and third CNN-based DNNs, respectively. The three distinct architectures are depicted in fig. 4.2. The performance of the different models is summarized in table 4.12

Table 4.10: Parameters used for convolutional layers in CNN-based DNNs.

Parameter	Value
Convolution kernel size	3x3
Convolution strides size	1x1
Pooling Type	Max Pooling
Pooling kernel size	2x2
Pooling strides size	1x1

Table 4.11: Parameters used for CNN-based DNN architectures.

Layer Type	Value
Convolutional Layer	6 filters
Flatten Layer	-
Dense Layer	84 nodes

(a) CNN-1

Layer Type	Value
Convolutional Layer	6 filters
Pooling Layer	-
Convolutional Layer	16 filters
Pooling Layer	-
Convolutional Layer	120 filters
Flatten Layer	-
Dense Layer	84 nodes

(b) CNN-2

Layer Type	Value
Convolutional Layer	6 filters
Pooling Layer	-
Convolutional Layer	16 filters
Pooling Layer	-
Convolutional Layer	120 filters
Pooling Layer	-
Convolutional Layer	240 filters
Pooling Layer	-
Convolutional Layer	240 filters
Flatten Layer	-
Dense Layer	84 nodes

(c) CNN-3

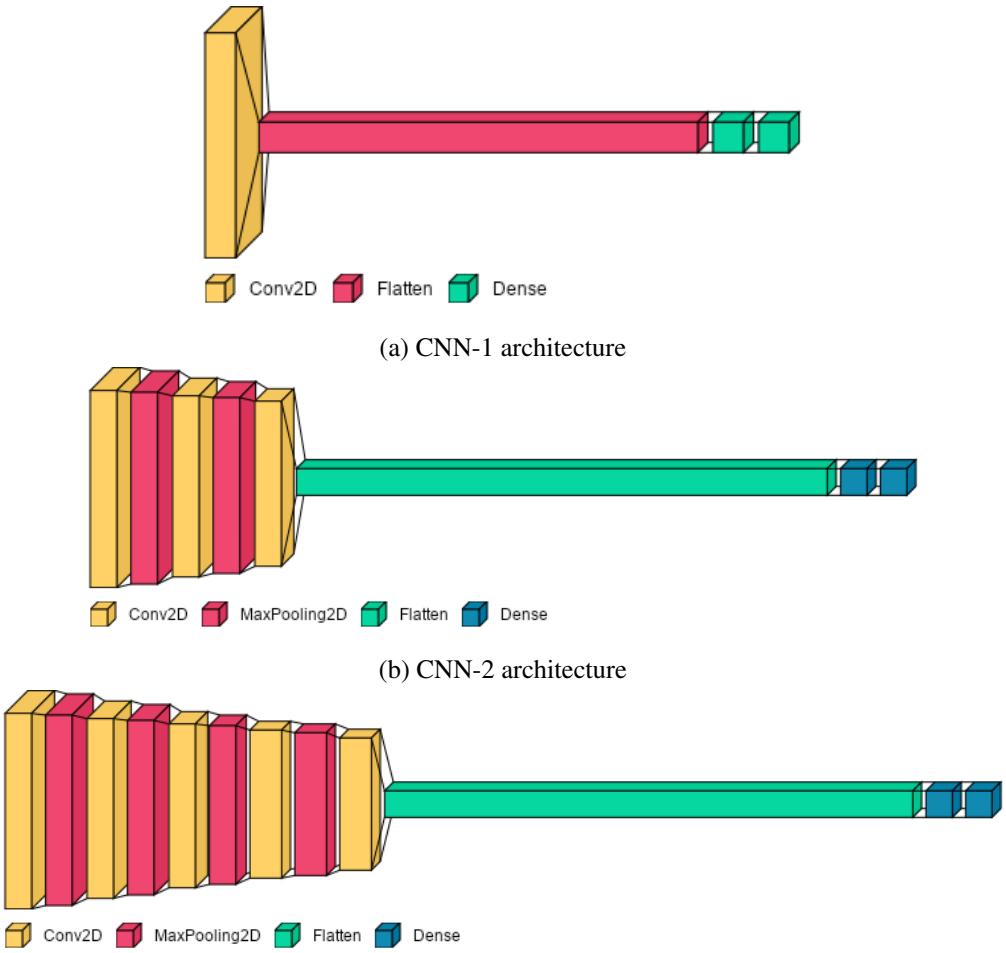
Table 4.12: Recorded accuracy for different CNN-based architectures.

DNN	6 classes	12 classes	36 classes
CNN-1	95.13%	90.68%	79.41%
CNN-2	94.46%	89.76%	77.99%
CNN-3	93,39%	86.09%	73.48%

According to the results presented in table 4.12, it is evident that the best performance is obtained using only one convolutional layer. Hence, the first CNN-based DNN (CNN-1) was selected for further analysis. The possible reasons behind this observation will be discussed in section 4.2.3.

4.2.3 Selected DNN:

Based on the results obtained from various tests conducted on different DNNs, it was surprising to observe that the single HL MLP exhibited satisfactory performance while utilizing significantly fewer parameters than the other DNNs. The reason for the better performance of the single HL MLP could be attributed to its comparison of the exact position in the input data, as opposed to the CNN that searches for particular patterns in the data using multiple filters without paying much



(a) CNN-1 architecture

(b) CNN-2 architecture

(c) CNN-3 architecture

Figure 4.2: Visualization of different CNN architectures used for DNNs selection process

attention to the specific location. As the input data in this study were correlation sequences, the exact positions contained information that facilitated better DOA estimation by the NNs. This may not be the case for very deep CNN-based DNNs due to receptive fields in CNNs, which is a topic that is beyond the scope of this study. Additionally, deeper NNs require more parameters, whereas our model needs to have fewer parameters. Therefore, the single HL MLP was selected for further analysis.

It is worth mentioning that when spectral estimation is not utilized , the CNN with five convolutional layers exhibits superior performance compared to other networks. This finding could be attributed to the fact that the absence of spectral estimation methods doesn't enhance the significant features of the GCC-PHAT sequences. This underscores the significance of employing convolutional layers, that utilize filters to extract features from the input data, especially when feature extraction methods employed are imprecise and obscure.

4.3 Classification vs. Regression

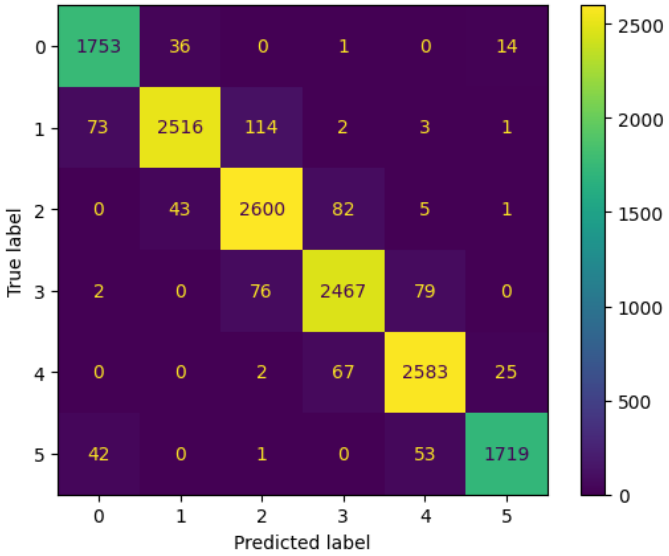
Once a model has been selected, it becomes necessary to compare its performance in solving different tasks, namely regression and classification, as detailed in section 3.3.1. To ensure a robust comparison, it is essential to utilize both accuracy and confusion matrix as evaluation metrics to gauge the model's efficacy across these tasks. As highlighted in section 3.3.3, it is important to note that the evaluation of the regression task can be approached in a similar manner to that of the classification task. The number of classes used for these tasks are displayed in table 4.6. Furthermore, the number of data points utilized for the train, test, and validation sets are presented in table 4.7.

4.3.1 Classification Task

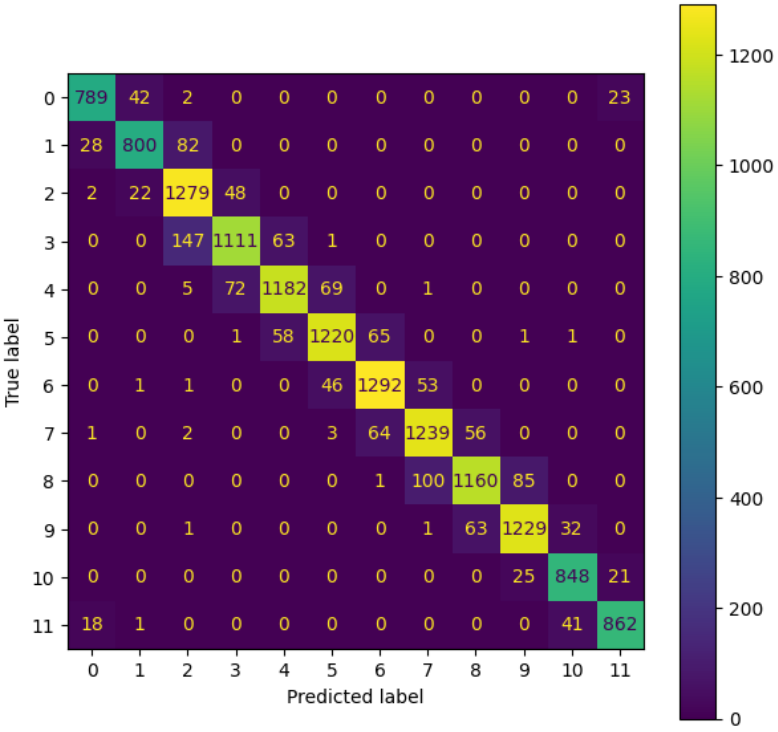
The classification task, as previously elucidated, entails the categorization of the input GCC-PHAT matrix into distinct DOA classes. In pursuit of this objective, the selected MLP, as detailed in section 4.2.3, was employed. Early stopping was used with the same set of parameters shown in table 4.1. Table 4.13 showcases the accuracy achieved by the MLP across diverse tasks, while fig. 4.3 depicts the corresponding confusion matrices pertaining to each classification task.

Table 4.13: Recorded accuracy for the classification task using the chosen MLP.

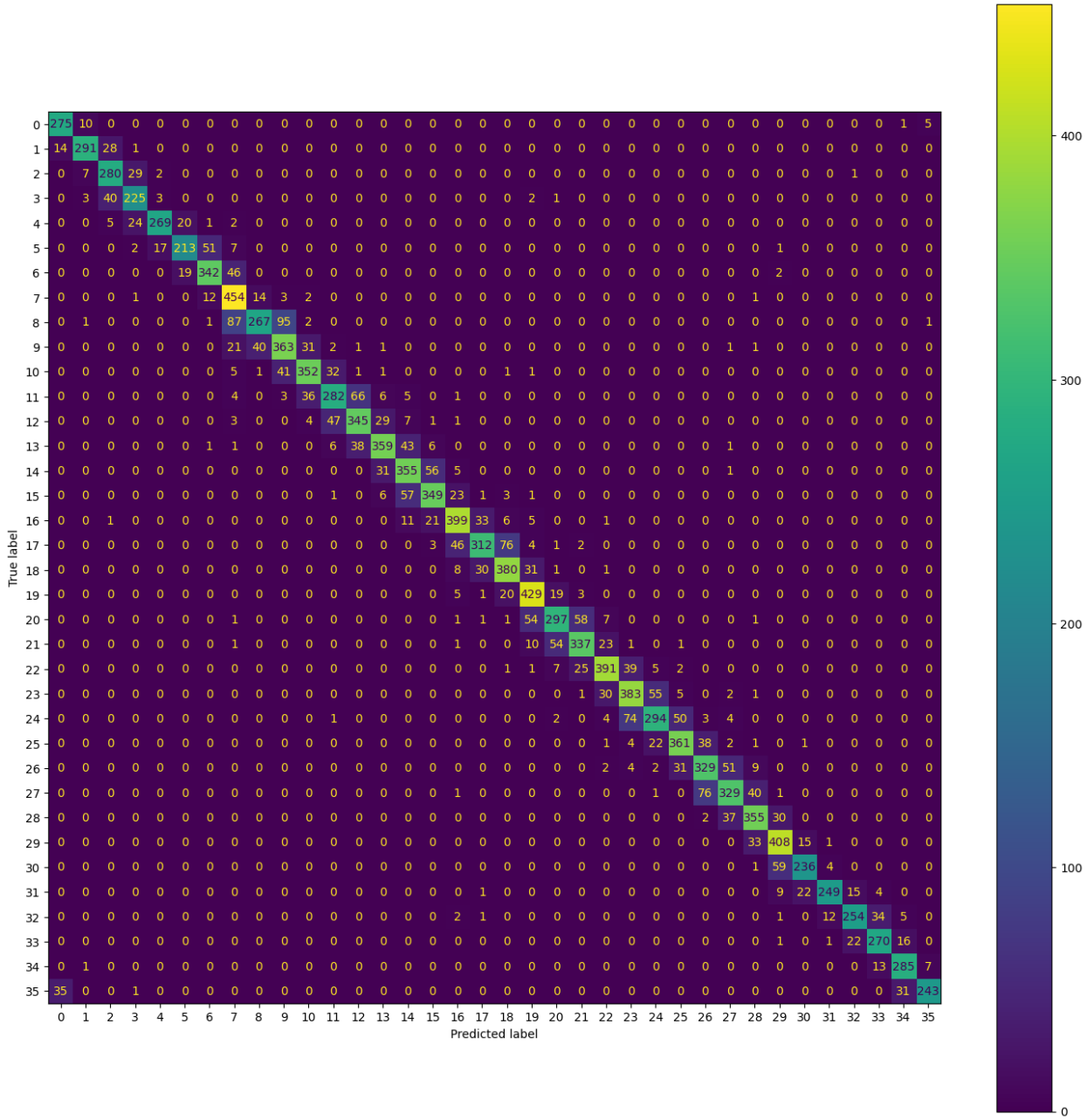
Number of Classes	Accuracy
6	95.27%
12	90.61%
36	80.52%



(a) 6 Classes



(b) 12 Classes



(c) 36 Classes

Figure 4.3: Confusion matrices for each classification task using the chosen MLP.

4.3.2 Regression Task

In the regression task, the focus lies on training the model to make predictions of the Cartesian unit vector that denotes the direction towards the source on the x-y plane. The specific training parameters employed for this purpose are presented in table 4.14. To ensure optimal model performance, early stopping is implemented and the relevant parameters are outlined in table 4.15.

Table 4.14: Parameters used to train models for the regression task.

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Mean Squared Error
Batch Size	32

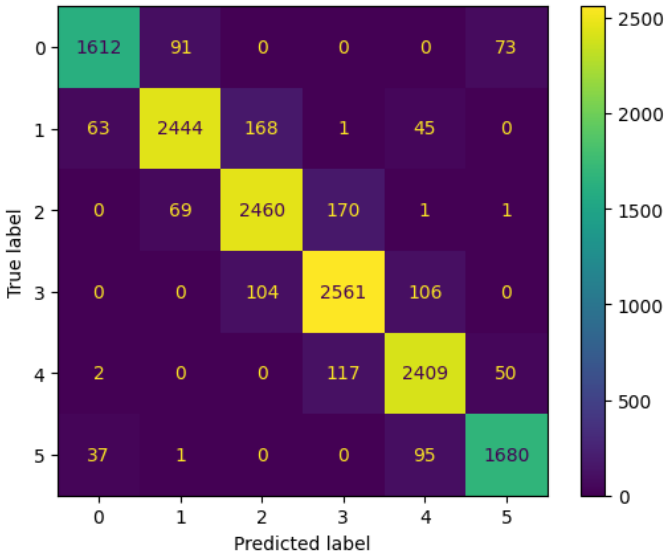
Table 4.15: Early stopping parameters used to train models for the regression task.

Parameter	Value
Monitor	Validation Loss
Minimum Change	0
Patience	3

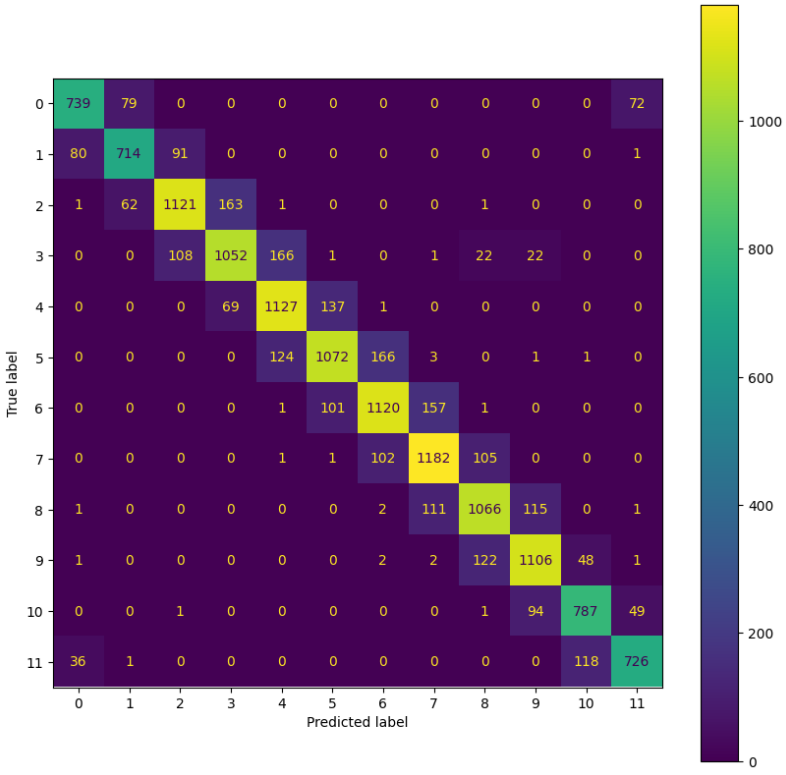
In order to facilitate a comparative analysis between the regression and classification models, the predicted unit vector obtained from the regression model serves as the basis for computing the angle of arrival. Subsequently, this angle is utilized to estimate the corresponding DOA class. It is important to note that a single trained model is employed for this purpose, and its outputs are utilized multiple times to compute distinct accuracies for each task. The resulting accuracies for the classification tasks are presented in table 4.16. Moreover, fig. 4.4 visualizes the associated confusion matrices pertaining to these classification tasks.

Table 4.16: Recorded accuracy for the regression task using the chosen MLP.

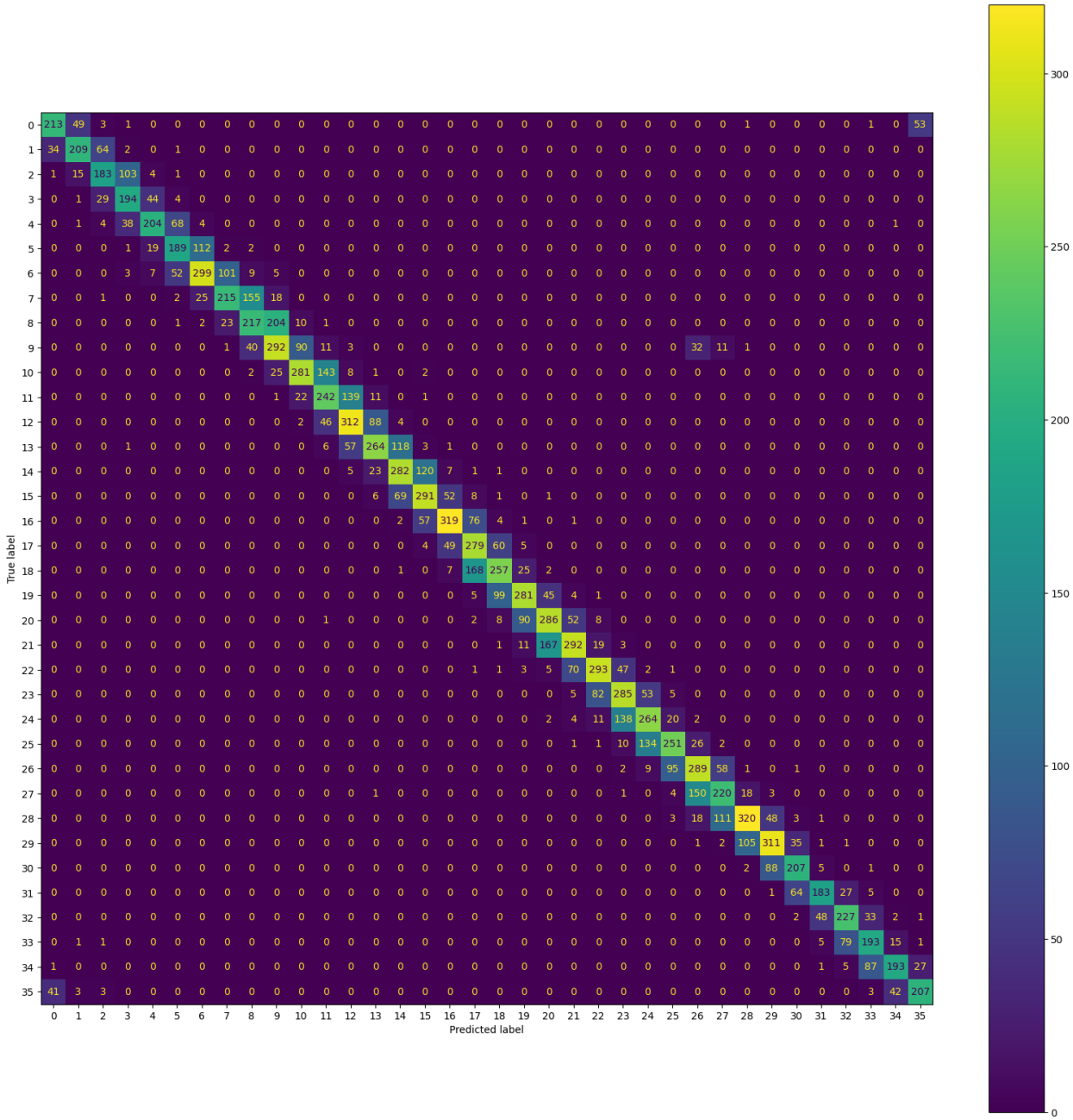
Number of Classes	Accuracy
6	91.68%
12	82.26%
36	63.21%



(a) 6 Classes



(b) 12 Classes



(c) 36 Classes

Figure 4.4: Confusion matrices for each classification task using the chosen MLP.

4.3.3 Discussion:

The comparative analysis presented in table 4.13 and table 4.16 reveals noticeable discrepancies in the performance of the regression MLP when faced with the scenarios involving 12 and 36 classes. Upon examination of the confusion matrices for both the regression and classification MLPs, it becomes evident that the models exhibit commendable performance, as the instances of

incorrect predicted DOA predominantly correspond to the adjacent class. This indicates that if an error occurs in the predicted DOA, it is likely to be in close proximity to the actual DOA of the signal.

While the classification MLP outperforms the regression MLP significantly for high resolution scenarios, it is important to emphasize that the regression model only necessitates a single training session as it predicts a unit vector. This distinctive characteristic empowers the system to adapt to changes and dynamically adjust the number of DOA estimate classes, all without the need for multiple DNNs or training multiple models.

4.4 Comparing DNN with DOA Algorithms:

In order to validate the selected DNN, it is imperative to compare its performance against established algorithms widely used for DOA estimation. The algorithms employed for this comparison encompass MUSIC and SRP-PHAT algorithms. To ensure a fair and unbiased assessment, the data utilized for evaluating these algorithms is identical to the the data used for the test set during DNN selection. For this task, accuracy will be the only evaluation metric utilized.

Table 4.17 provides a comprehensive overview of the DOA estimation accuracy achieved by the MUSIC algorithm, SRP-PHAT algorithm, and the chosen MLP for classification tasks involving 6 classes, 12 classes, and 36 classes. Remarkably, the MLP demonstrates superior performance compared to the algorithms employed. Furthermore, it is noteworthy that the MUSIC algorithm outperforms the SRP-PHAT algorithm in terms of accuracy.

Table 4.17: Recorded accuracy of the chosen MLP, MUSIC algorithm, and SRP-PHAT algorithm for different classification tasks.

Algorithm/Model \ Number of Classes	6	12	36
MLP	95.27%	90.61%	80.52%
MUSIC	92.31%	84.03%	61.66%
SRP-PHAT	73.67%	70.67%	44.43%

Chapter 5

Impact and Exploitation:

This chapter provides an overview of the current stage of the project's development cycle and explores its potential impact on a range of fields and applications, including speech processing, virtual reality, and antenna arrays. In addition, it examines various strategies that could be employed to maximize the project's impact in these areas.

5.1 Position in the Development Cycle:

Research on machine learning-based DOA estimation is a relatively recent development [43]. Traditionally, DOA estimation has relied on classical methods that involve multiple stages of estimation, such as determining the number of sources before estimating DOA. In these classical methods, inaccuracies in any of the intermediate stages can negatively impact the final DOA estimation. However, with the increasing prevalence of DL in signal processing, researchers have turned to DL methods for DOA estimation. These methods offer improved performance and generalization with fewer stages and less computational burden.

Currently, in the research and development cycle, this project is testing and developing different signal processing and DL methods to enhance and produce promising outcomes for the problem of DOA estimation. Furthermore, companies such as Amazon, who have virtual assistant products like Alexa, are already investing in research and development to improve their speaker localization using DL methods [44]. This project aligns well with the current demand for efficient and robust methods for speaker localization and DOA estimation, which is driving researchers to employ cutting-edge techniques, such as deep neural networks, to accelerate technological ad-

vancement.

5.2 Impacts:

The impact of the research presented in this thesis spans across several fields, with one key application being speech separation that relies on DOA estimation algorithms. Such algorithms can be used in a wide range of applications, including virtual assistants, hearing aids [12], surveillance [7], and more. The findings of this project also have significant implications for virtual and augmented reality, where DOA estimation algorithms are used to create immersive audio experiences. By utilizing accurate DOA estimation methods, users can have a more engaging experience with enhanced spatial awareness and a sense of presence [45]. The audio simulation methods used in this research can also aid in the development of these algorithms for improved user experience. Additionally, DOA estimation is not limited to acoustics and can be utilized in antenna systems to focus antenna arrays on specific directions, thereby improving signal quality and reducing interference [46]. DOA estimation methods are also essential in radar and sonar for tracking and navigation applications [3].

5.3 Methods to Influence:

This project enhances the awareness of governments and policymakers about the current state of technology and its potential uses, including both ethical and unethical applications. The project aims to encourage regulations that mitigate the risks of unethical technology use. The research presented in this thesis also underscores the significance of interdisciplinary research by utilizing signal processing and DL techniques to tackle current issues. The project has potential benefits for companies involved in producing virtual assistants, particularly Apple and their latest product, Homepod. The Homepod uses beamforming methods for speaker localization and environment sensing to optimize its tweeter array for superior sound production [47]. Therefore, our findings will assist these companies in advancing their technology further.

To translate the impacts into real-world applications, further research is required to enhance the algorithms and DNN architectures to consider a wide range of scenarios. I is believed that collaboration among experts from diverse sectors is crucial to advance the development to create more autonomous and intelligent products. Therefore, the allocation of additional funding to

facilitate these collaborations and to support the further development & testing of DOA estimation methods using DNNs is proposed. The expansion of the virtual assistant industry, which is expecting a market growth by USD 4.12 billion [48],coupled with the growing demand for more accurate DOA estimation techniques across various applications, increases the likelihood of securing funding for projects involving this technology. This collaboration will propel the product into the advanced stages of the development cycle, where it can undergo rigorous testing before being deployed to end-users.

Chapter 6

Conclusion:

This chapter marks the culmination of the thesis, providing conclusions on the key findings. Additionally, potential avenues for future research and extensions are discussed.

6.1 Conclusion:

In this thesis, the feasibility of DNNs for solving the problem of DOA estimation and source localization was thoroughly investigated. Simulation tools to collect data and evaluate the performance of DL models were utilized. Furthermore, the investigation encompassed an examination of feature extraction techniques, specifically utilizing spectral estimation with GCC-PHAT to improve significant features from the input data. The experiments showed that a single HL MLP achieved the best performance on the simulated data for multiple classification tasks. Furthermore, we compared the performance of the DNN with various DOA estimation algorithms and demonstrated that the DNN consistently outperforms them, particularly in high resolution scenarios. Notably, in the 36-classes scenario, the DNN exhibited significantly superior accuracy compared to the MUSIC algorithm by 18.86% and surpassed the SRP-PHAT algorithm by an impressive margin of 36.09%. These results underscore the remarkable performance of the DNN, particularly in more challenging DOA estimation tasks. Additionally, an alternative approach involving regression MLP was discussed, offering a fresh perspective on tackling the problem at hand. Overall, our findings demonstrate the potential of DNNs for solving the challenging problem of DOA estimation and source localization.

While the work presented in this thesis has added value to the field, it is acknowledged that the

project is still in its early stages. Further investigations are necessary to ensure that our findings are applicable to multiple source scenarios, which are more realistic, and to consider sources of noise and interference. Additionally, it is important to incorporate more room geometries and environments to enhance the generalizability of the DL models.

6.2 Future Work:

In future investigations, the potential for expanding the MLP by incorporating supplementary layers, or alternatively utilizing the MLP as an input to another DNN, arises. This can be leveraged to predict beamforming weights, with the aim of effectively isolating the signal of interest from interfering sources, thereby resulting in an enhanced SNR.The Capon's minimum variance beamformer, which has been previously discussed in the background chapter, stands out as a promising choice for utilizing a DNN to predict its beamforming weights as it can provide better noise reduction and interference rejection compared to the delay-and-sum beamformer.

In prospective investigations, there is potential for deploying the MLP within a real-time system to accurately predict the DOA of speech signals. However, to achieve this objective, it becomes necessary to fine-tune the original model, a process that entails recording real data for refinement purposes. The collection of a substantial amount of data across various environments assumes paramount importance, as it enables the model to exhibit efficient performance across diverse scenarios. This, in turn, facilitates the practical application of the model in real-life contexts, such as intelligent virtual assistants.

In this study, the primary focus on DNNs revolved around utilizing them for DOA estimation. However, it is worth noting the existence of numerous pre-trained DNNs in various speech applications, trained on extensive datasets comprising both real and simulated data. These pre-trained neural networks can be employed as feature extractors, wherein the last layer is removed, and the output of the penultimate layer serves as input features for the DOA estimation DNN. This approach not only aids in accentuating crucial features that traditional feature extraction methods may overlook but also serves as a dimensionality reduction technique by transforming a large amount of data into a more compact representation.

Acknowledgment

I dedicate this thesis to my parents, who have been my pillars of support since my childhood. Your unwavering belief in me, encouragement, and sacrifices have shaped me into the person I am today. I am grateful for your unconditional love and constant guidance.

I also dedicate this thesis to the unexpected paths in life that we have embarked on. The twists and turns, challenges, and triumphs have molded my perspective and enriched my understanding of the world. Each detour has led me to new insights and opportunities for personal and intellectual growth.

Furthermore, I extend my heartfelt gratitude to my family and friends who have stood by me throughout my journey. Your presence, encouragement, and unwavering support have provided the strength and motivation I needed to overcome obstacles and reach this milestone. I am truly fortunate to have you all in my life.

This thesis is a testament to the profound impact that my parents, my dear family, and my friends have had on my academic and personal development. Thank you for being my guiding lights, and for being with me every step of the way.

References

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2001.

[2] C. Junu Jahana, M. Sinith, and P. Lalu, "Direction of arrival estimation using microphone array," in *2021 Fourth International Conference on Microelectronics, Signals Systems (ICMSS)*, pp. 1–6, 2021.

[3] X. Lyu and J. Wang, "Direction of arrival estimation in passive radar based on deep neural network," *IET Signal Processing*, vol. 15, no. 9, pp. 612–621, 2021.

[4] M. Al-Nuaimi, R. Shubair, and K. Al-Midfa, "Direction of arrival estimation in wireless mobile communications using minimum variance distortionless response," in *The Second International Conference on Innovations in Information Technology (IIT'05)*, pp. 1–5, 2005.

[5] N. Dey, A. S. Ashour, N. Dey, and A. S. Ashour, "Challenges and future perspectives in speech-sources direction of arrival estimation and localization," *Direction of arrival estimation and localization of multi-speech sources*, pp. 49–52, 2018.

[6] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3d sound localization system with miniature microphone array for virtual reality," in *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1853–1857, IEEE, 2012.

[7] A. AlShehhi, M. L. Hammadih, M. S. Zitouni, S. AlKindi, N. Ali, and L. Weruaga, "Linear and circular microphone array for remote surveillance: simulated performance analysis," *arXiv preprint arXiv:1703.02318*, 2017.

[8] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[9] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E.-S. Chng, and H. Li, "The ntu-adsc systems for reverberation challenge 2014," in *Proc. REVERB challenge workshop*, p. o2, Spoken Language Systems MIT Computer Science and Artificial Intelligence ..., 2014.

[10] S. Zhao, E. S. Chng, N. T. Hieu, and H. Li, "A robust real-time sound source localization system for olivia robot," in *2010 APSIPA annual summit and conference*, 2010.

[11] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[12] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. R. Liu, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2010.

[13] A. Aroudi and S. Doclo, "Cognitive-driven binaural lcmv beamformer using eeg-based auditory attention decoding," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 406–410, 2019.

[14] X. Mestre and M. Á. Lagunas, "Modified subspace algorithms for doa estimation with large arrays," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598–614, 2008.

[15] B. D. Rao and K. Hari, "Weighted subspace methods and spatial smoothing: analysis and comparison," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 788–803, 1993.

[16] B. Kwon, Y. Park, and Y.-s. Park, "Analysis of the gcc-phat technique for multiple sources," in *ICCAS 2010*, pp. 2070–2073, IEEE, 2010.

[17] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[18] J. Stefański, "Hyperbolic position location estimation in the multipath propagation environment," in *Wireless and Mobile Networking: Second IFIP WG 6.8 Joint Conference, WMNC 2009, Gdańsk, Poland, September 9-11, 2009. Proceedings*, pp. 232–239, Springer, 2009.

[19] Y. T. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE transactions on signal processing*, vol. 42, no. 8, pp. 1905–1915, 1994.

[20] A. Singh, M. Yu, A. Gupta, and K. Bryden, "Localization of multiple acoustic sources in a room environment," *Applied energy*, vol. 109, pp. 171–181, 2013.

[21] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[22] V. Krishnaveni, T. Kesavamurthy, and B. Aparna, "Beamforming for direction-of-arrival (doa) estimation-a survey," *International Journal of Computer Applications*, vol. 61, no. 11, 2013.

[23] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University, 2000.

[24] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[25] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[26] P. Diggle and I. Al-Wasel, "On periodogram-based spectral estimation for replicated time series," *Developments in time series analysis*, pp. 341–354, 1993.

[27] N. Wiener, "Generalized harmonic analysis," *Acta mathematica*, vol. 55, no. 1, pp. 117–258, 1930.

[28] M. S. Bartlett, "Periodogram analysis and continuous spectra," *Biometrika*, vol. 37, no. 1/2, pp. 1–16, 1950.

[29] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8599–8603, IEEE, 2013.

[30] P. Goyal, S. Pandey, and K. Jain, "Deep learning for natural language processing," *New York: Apress*, 2018.

[31] M. Pak and S. Kim, "A review of deep learning in image recognition," in *2017 4th international conference on computer applications and information processing technology (CAIPT)*, pp. 1–3, IEEE, 2017.

[32] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[33] S. Ge, K. Li, and S. N. B. M. Rum, "Deep learning approach in doa estimation: a systematic literature review," *Mobile Information Systems*, vol. 2021, pp. 1–14, 2021.

[34] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral cnns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 313–321, 2022.

[35] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2814–2818, IEEE, 2015.

[36] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[37] M. Vorländer, "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*, vol. 86, no. 1, pp. 172–178, 1989.

[38] E. Panahi and D. Younesian, "Acoustic performance enhancement in a railway passenger carriage using hybrid ray-tracing and image-source method," *Applied Acoustics*, vol. 170, p. 107527, 2020.

[39] M. Wajid, B. Kumar, A. Goel, A. Kumar, and R. Bahl, "Direction of arrival estimation with uniform linear array based on recurrent neural network," in *2019 5th international conference on signal processing, computing and control (ISPCC)*, pp. 361–365, IEEE, 2019.

[40] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136–140, IEEE, 2017.

[41] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–15, 2006.

[42] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[43] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[44] J. Barber, Y. Fan, and T. Zhang, "End-to-end alexa device arbitration," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 926–930, IEEE, 2022.

[45] A. Tepljakov, S. Astapov, E. Petlenkov, K. Vassiljeva, and D. Draheim, "Sound localization and processing for inducing synesthetic experiences in virtual reality," in *2016 15th Biennial Baltic Electronics Conference (BEC)*, pp. 159–162, IEEE, 2016.

[46] M. Chen, Y. Gong, and X. Mao, "Deep neural network for estimation of direction of arrival with antenna array," *IEEE Access*, vol. 8, pp. 140688–140698, 2020.

[47] "A deep dive into homepod's adaptive audio, beamforming and why it needs an a8 processor." `https://appleinsider.com/articles/18/01/27/a-deep-dive-into-homepods-adaptive-audio-beamforming-and-why-it-needs-an-a8-processor` Accessed: 2023-04-28.

[48] "Virtual assistant market size to grow by usd 4.12 bn." `https://finance.yahoo.com/news/virtual-assistant-market-size-grow-080000234.html`. Accessed: 2023-04-28.

Appendix A

Tables Appendix

Table A.1: Simulation parameters used to create a room using PRA

Parameter	Value
Room Length	6 meters
Room Width	5 meters
Room Height	3 meters
Speed of Sound	343 meters per seconds
Reverberation Time	0.45 seconds
Maximum Images Order	1
Microphone Center's Coordinates	(4,3,1.5) meters

Table A.2: Simulation parameters used to create a sound source using PRA

Parameter	Value
Change in Distance	1 meter
Change in Azimuth Angle	1°

Table A.3: A comparison between sparse encoded categories and one-hot encoded categories.

Sparse Encoded Categories	One-Hot Encoded Categories		
Category	Category 0	Category 1	Category 2
2	0	0	1
0	1	0	0
1	0	1	0

Appendix B

Figures Appendix

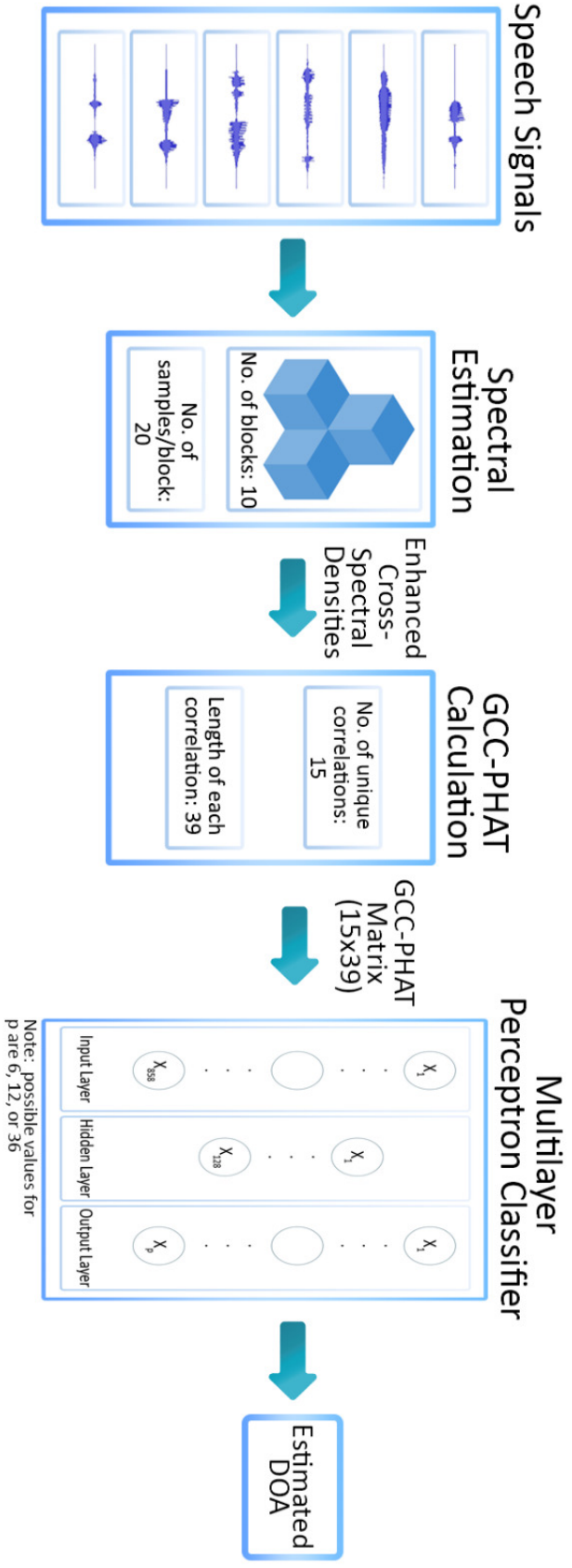


Figure B.1: Illustration of an optimal data pipeline for DOA estimation using a MLP