VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Saulius SAKAVIČIUS

IMPROVEMENT OF LEARNING-BASED METHODS FOR LOCALIZATION OF MULTIPLE SOUND SOURCES

DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
ELECTRICAL AND ELECTRONIC ENGINEERING (T 001)

Vilnius, 2021

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2016–2021.

Scientific supervisor

Prof. Dr Artūras SERACKIS (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

The Dissertation Defense Council of Scientific Field of Electrical and Electronic Engineering of Vilnius Gediminas Technical University:

Chairman

Prof. Dr Algirdas BAŠKYS (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

Members:

Dr Nicki HOLIGHAUS (Acoustics Research Institute of the Austrian Academy of Sciences, Electrical and Electronic Engineering – T 001),

Dr Tomas KRILAVIČIUS (Vytautas Magnus University, Informatics – N 009),

Prof. Dr Jurij NOVICKIJ (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001),

Assoc. Prof. Dr Raimondas POMARNACKI (Vilnius Gediminas Technical University, Electrical and Electronic Engineering – T 001).

The dissertation will be defended at the public meeting of the Dissertation Defense Council of Electrical and Electronic Engineering in the Senate Hall of Vilnius Gediminas Technical University at **10 a. m. on 10 December 2021**.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania.
Tel. +370 5 274 4956; fax +370 5 270 0112; e-mail: doktor@vilniustech.lt

A notification on the intend defending of the dissertation was send on 9 November 2021. A copy of the doctoral dissertation is available for review at Vilnius Gediminas Technical University repository http://dspace.vgtu.lt, at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and the Wroblewski Library of the Lithuanian Academy of Sciences (Žygimantų st. 1, LT-01102, Vilnius, Lithuania).

Vilnius Gediminas Technical University scientific book No. 2021-050-M

doi: 10.20334/2021-050-M

saulius.sakavicius@vilniustech.lt

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Saulius SAKAVIČIUS

MOKYMU GRĮSTŲ METODŲ KELIEMS GARSO ŠALTINIAMS LOKALIZUOTI TOBULINIMAS

DAKTARO DISERTACIJA

TECHNOLOGIJOS MOKSLAI,
ELEKTROS IR ELEKTRONIKOS INŽINERIJA (T 001)

Vilnius, 2021

Disertacija rengta 2016–2021 metais Vilniaus Gedimino technikos universitete.

Vadovas

prof. dr. Artūras SERACKIS (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001).

Vilniaus Gedimino technikos universiteto Elektros ir elektronikos inžinerijos mokslo krypties disertacijos gynimo taryba:

Pirmininkas

prof. dr. Algirdas BAŠKYS (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001).

Nariai:

dr. Nicki HOLIGHAUS (Austrijos mokslo akademijos akustikos tyrimų institutas, elektros ir elektronikos inžinerija – T 001),

dr. Tomas KRILAVIČIUS (Vytauto Didžiojo universitetas, informatika – N 009),

prof. dr. Jurij NOVICKIJ (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001),

doc. dr. Raimondas POMARNACKI (Vilniaus Gedimino technikos universitetas, elektros ir elektronikos inžinerija – T 001).

Disertacija bus ginama viešame Elektros ir elektronikos inžinerijos mokslo krypties disertacijos gynimo tarybos posėdyje **2021 m. gruodžio 10 d. 10 val.** Vilniaus Gedimino technikos universiteto Senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.
Tel.: (8 5) 274 4956; faksas (8 5) 270 0112; el. paštas: doktor@vilniustech.lt

Pranešimai apie numatomą ginti disertaciją išsiusti 2021 m. lapkričio 9 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto talpykloje http://dspace.vgtu.lt, Vilniaus Gedimino technikos universiteto bibliotekoje (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) bei Lietuvos mokslų akademijos Vrublevskių bibliotekoje (Žygimantų g. 1, LT-01102 Vilnius, Lietuva).

Abstract

Sound source localization is an important topic in human-machine interacting, teleconferencing, security systems, as well as autonomous driving and robotics. While current state-of-the-art sound source localization methods allow localization of a single or a small number of sound sources in moderately reverberant environments, it is known that their performance deteriorates when the reverberation time is increased. Moreover, the localization of multiple sound sources is an even more difficult task. Learning-based sound source localization methods recently gained interest as they tend to outperform the state-of-the-art methods in multiple source localization cases in reverberant environments. Nevertheless, this branch of sound source localization methods is not yet sufficiently investigated. Therefore, this thesis is aimed to the research of such methods. Both regression-based and classification-based methods for single and multiple sound source localization in two-dimensional and three-dimensional space are investigated. Supervised and semi-supervised training strategies are researched. A dataset of tetrahedral microphone array signals is collected for the evaluation of the performance of sound source localization methods. The dissertation consist of an introduction, three chapters and general conclusions. In the introduction, the dissertation problem is formulated, the object of the research is defined and the aim of the thesis is presented. Next, the objectives of the thesis are formulated. A brief presentation of the research methodology is provided, followed by the outline of the scientific novelty of the thesis and the practical value of the research findings. Finally, the defended statements are formulated. The first chapter reveals the state of the art of sound source localization using microphone arrays and networks. In the section, most important sound source localization methods are outlined, with an emphasis on learning-based source localization methods. In the second chapter presented are the learning-based sound source localization methods suggested by the author. Specifically, the multi-layer perceptron-based method for single sound source localization in two dimensions, the convolutional neural network-based methods for multiple sound source localization in two and three dimensions and the Graph-Regularized Neural Network-based single sound source localization method. In the third chapter, the experimental setups for evaluation of the performance of the original methods, presented in the second chapter, and the results of the experimentation are presented. In the final chapter, the discussion on the experimental results is presented and the conclusions are drawn. The results of the thesis were published in six scientific publications: three papers in the reviewed scientific journals and three papers in other journals. Additionally, the results of the research were presented in five conferences.

Reziumė

Garso šaltinio lokalizavimas yra svarbus elementas žmogaus ir kompiuterio sąveikos, telekonferencijų, apsaugos sistemų, taip pat autonominio vairavimo ir robotikos srityse. Nors dabartiniai moderniausi garso šaltinių lokalizavimo metodai leidžia lokalizuoti vieną ar nedidelį skaičių garso šaltinių vidutiniškai aidžioje aplinkoje, yra žinoma, kad padidėjus aidėjimo laikui jų veikimas blogėja. Be to, kelių garso šaltinių lokalizavimas yra dar sudėtingesnė užduotis. Mokymusi grįsti garso šaltinio lokalizavimo metodai pastaruoju metu sulaukia vis didesnio susidomėjimo, nes jų veikimo tikslumas pranoksta pažangiausius klasikinius šaltinių lokalizavimo metodus daugelio šaltinių lokalizavimo atvejų aidžioje aplinkoje. Nepaisant to, ši garso šaltinio lokalizavimo metodų šaka dar nėra pakankamai ištirta. Todėl ši disertacija skirta mokymusi grįstų metodų tyrimams. Tiriami tiek regresija, tiek klasifikavimu pagrįsti metodai, skirti vieno ir kelių garso šaltinių lokalizavimui dvimatėje ir trimatėje erdvėje. Tiriamos prižiūrimo ir pusiau prižiūrimo mokymo strategijos. Garso šaltinio lokalizavimo metodų veikimui įvertinti surinktas tetraedrinės mikrofonų gardelės signalų duomenų rinkinys. Disertaciją sudaro įvadas, trys skyriai ir bendros išvados. Įvade suformuluojama disertacijos problema, apibrėžiamas tyrimo objektas ir pateikiamas darbo tikslas. Toliau suformuluoti baigiamojo darbo tikslai. Pateikiamas trumpas tyrimo metodikos pristatymas, po kurio aprašoma baigiamojo darbo mokslinė naujovė ir praktinė tyrimo išvadų vertė. Galiausiai suformuluojami ginami teiginiai. Disertaciją sudaro įvadas, trys skyriai ir bendros išvados. Pirmame skyriuje aprašoma garso šaltinio lokalizacijos pažangiausi metodai, kurie naudoja mikrofonų gardeles ir dirbtinius neuronų tinklus. Skyriuje aprašomi svarbiausi garso šaltinio lokalizavimo metodai, akcentuojant mokymusi grįstus garso šaltinio lokalizavimo metodus. Antrame skyriuje pateikiami autoriaus siūlomi mokymu pagrįsti garso šaltinių lokalizavimo metodai: daugiasluoksniu perceptronu pagrįstas vieno garso šaltinio lokalizavimo dvimatėje erdvėje metodas, sąsūkos neuronų tinklu pagrįstas metodas, skirtas daugelio garso šaltinių lokalizavimui dvimatėje erdvėje, ir grafu reguliarizuotu dirbtiniu neuronų tinklu pagrįstas vieno garso šaltinio lokalizavimo dvimatėje erdvėje metodas. Aptariami garso šaltinio lokalizavimo tikslumą įtakojantys veiksniai. Pristatomi akustiniai požymiai, kurie gali būti naudojami su mokymu grįstais garso šaltinio lokalizavimo metodais. Trečiajame skyriuje pateikiami metodų, aprašytų antrajame skyriuje, eksperimentinių tyrimų aprašymai ir rezultatai. Pristatomas tetraedrinių gardelių signalų duomenų rinkinys. Paskutiniame skyriuje pateikiamos bendrosios disertacijos išvados. Darbo rezultatai buvo paskelbti šešiose mokslinėse publikacijose: trijuose recenzuojamuose mokslo žurnaluose ir trijuose kituose leidiniuose. Be to, tyrimo rezultatai buvo pristatyti penkiose konferencijose.

Notations

Abbreviations

ANN	–	Artificial Neural network
CNN	–	Convolutional Neural Network
DoA	–	Direction of Arrival
GRNN	–	Graph Regularized Neural Network
ILD	–	Inter-Aural Level Difference
IPD	–	Inter-Aural Phase Difference
ISOMAP	–	Isometric Mapping
MAE	–	Mean Average Error
MLP	–	Multi-Layer Perceptron
MSE	–	Mean Squared Error
NLDR	–	Non-Linear Dimensionality Reduction
PHAT	–	Phase Transform
RMS	–	Root Mean Square
RMSE	–	Root Mean Squared Error
SSL	–	Sound Source Localization
SRP	–	Steered Response Power
STD	–	Standard Deviation
STFT	–	Short-Time Fourier Transform
TDoA	–	Time Difference of Arrival

Contents

[1]The annexes are supplied in the enclosed compact disc.

Introduction

Problem Formulation

Sound source localization is an important topic in robotics, autonomous vehicles, public security, conferencing, sound engineering and other fields. Applications of sound source localization include speaker location discovering in a teleconference, event detection and tracking, robot movement in an unknown environment (Argentieri *et al.* 2015; Kotus 2013).

It is often needed to localize the sound source with accuracy that is close to or better than human sound source localization abilities, that is, the ability to determine the direction of arrival (DoA) of the sound source within accuracy of 15°. Moreover, there is often a need to determine not only the source DoA, but also the distance to the receiver. The task is becomes more complex when there is a need to localize multiple simultaneously active sound sources, by either selecting the most prominent sound source and suppressing others (one-vs-many scenario) or by localizing an arbitrary number of strongest sound sources.

Relevance of the Thesis

Current algorithms are not robust enough against the environmental noise, the adverse acoustical conditions (echoes, reverberation). The accuracy of existing sound source localization methods based on the determination of the propagation time

difference decreases when reverberation occurs in the environment and when there are noise sources in the environment, therefore, learning-based sound source localization methods have been actively researched recently. The growing interest in learning-based sound source localization methods can be illustrated by the number of articles published in the field, which was increasing exponentially between the year 2011 and 2019 (see Fig. 0.1). Currently available learning-based sound source localization methods can identify the direction of several sound sources, but not the distance to them or their coordinates in three-dimensional space. In 3 dimensions, the coordinates of several sound sources can be determined by formulating the problem as a regression problem, but for this the number of sources in the acoustic scene must be known in advance. A supervised training strategy requires a large number of labeled samples, and sample labeling is a complex and time-consuming task; strategies for unsupervised training or semi-supervised (hybrid) training would reduce or eliminate the labeling of training samples, thus reducing the time and cost of developing solutions.

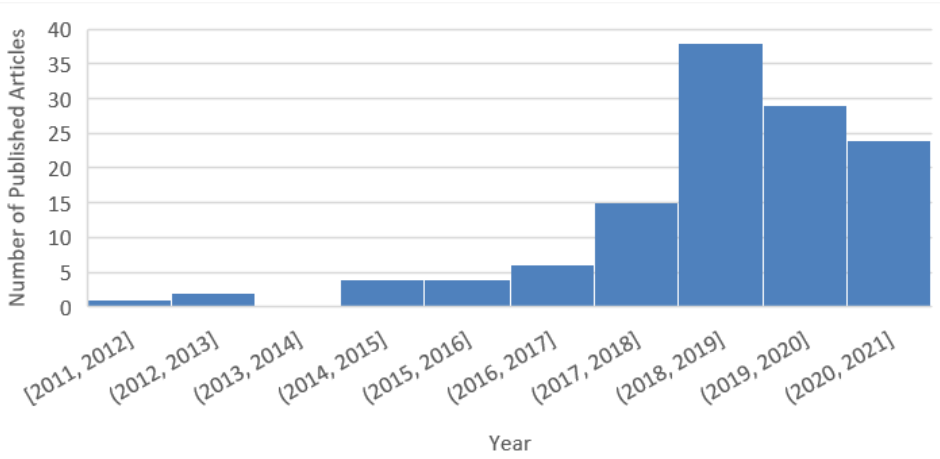


Fig. 0.1. Number of articles published in the field of learning-based sound source localization

Sound source localization is used for talker identification in teleconference scenarios, as an aid for robotic locomotion and orientation in an unknown environment (Kotus 2013) as well as in surveillance and security enforcement scenarios (Lopatka *et al.* 2011). Acoustic source localization is one of the key aspect of human-robot interaction (Athanasopoulos *et al.* 2015).

The Object of the Research

The object of the research of this thesis is learning-based methods for single and multiple acoustic source localization, considering the acoustic properties of the acoustic enclosure and the parameters of the sound source signals.

The Aim of the Thesis

The aim of the thesis is to propose original learning-based methods for acoustic source localization within reverberant enclosures.

The Objectives of the Thesis

In order to solve stated problem and reach the aim of the thesis the following main objectives are formulated:

1. To propose and investigate the supervised learning-based sound source localization approaches for multiple sound source two-dimensional localization within an acoustic enclosure.
2. To propose and investigate the supervised learning-based sound source localization approaches for multiple sound source three-dimensional localization within an acoustic enclosure.
3. To propose and investigate the semi-supervised and/or unsupervised sound source localization approaches for a single sound source localization within an acoustic enclosure.

Research Methodology

Experimental evaluation of the proposed methods for sound source localization involved computer simulation of microphone array signals using image-source method, as well as real-world microphone array signal acquisition and processing. Simulated microphone array signals were generated and artificial neural network models were trained using Python programming language with pyroomacoustics acoustic simulation package and TensorFlow machine-learning package. Other computer simulations and calculations were performed in either MATLAB or Python programming languages. Real-world microphone array signals were obtained using 4 element planar microphone arrays and 4 element tetrahedral arrays of various apertures.

Scientific Novelty of the Thesis

1. A dataset of tetrahedral microphone array signals for the localization of one and two sound sources in a reverberant environment has been prepared and made public, allowing to evaluate the performance of sound source localization methods with real data and to compare simulated microphone array signals to real signals.
2. Novel methods, based on convolutional artificial neural networks for determining the two-dimensional direction of arrival of several sound sources using microphone array signals' cross-correlation in frequency bands and spectrum phase component as an input features are presented.
3. A novel convolutional artificial neural network-based method for the localization of multiple sound sources in three-dimensional space involving a three-dimensional neural network output structure is presented.
4. A method based on hybrid learning and graph-regularized artificial neural network (GRNN) for localization of a single source in two-dimensional space is presented, allowing the training of a sound source localization system using a limited set of labeled samples.

Practical Value of the Research Findings

A data set of tetrahedral microphone grid signals with one and two sound sources in the echo environment has been collected and made public. The dataset has not only the positions of the sources and microphones marked, but also the geometry and acoustic parameters of the room measured and presented, allowing to compare real and simulated microphone array signals and to determine the accuracy of the signal simulation.

A methodology for the creation of simulated acoustic signal datasets for the study of graph-regularized neural networks is presented. A method for increasing the accuracy of sound source coordinates in three-dimensional space using grouping is investigated.

The Defended Statements

1. Using a graph-regularized artificial neural network trained with a semi-supervised training strategy and SRP-PHAT spatial spectra input features, the mean average error of the localization of a single sound source in two

dimensions can be reduced up to 4% compared to the SRP-PHAT intensity map peak determination method.

2. Cross-correlation in frequency bands features can be used to determine the direction of arrival of a sound source in two-dimensional space and achieve a mean localization error of 23 degrees for one sound source and a mean localization error of 26 degrees for two sound sources.
3. Using the spectral phase component as a feature, using convolutional artificial neural network can achieve up to 36% smaller localization error of three audio sources in two-dimensional space than with the widely used SRP-PHAT algorithm.
4. Using the proposed modification of the convolutional artificial neural network output layer and the spectral phase component as the input feature, it is possible to achieve a localization error as small as 1.08 m for two speaker localization in three-dimensional space in a reverberant environment.

Approval of the Research Findings

The results of the research were published in 6 scientific publications, 3 in peer-reviewed scientific papers, 3 in conference proceedings. Additionally, the results of the research were presented in following conferences:

- Two Young Scientist Conferences Science – Future of Lithuania, 2017, 2019, Vilnius, Lithuania.
- Two Open Conferences of Electrical, Electronic and Information Sciences (eStream), 2017, 2019, Vilnius, Lithuania.
- IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), 2017, Riga, Latvia.

Structure of the Dissertation

The dissertation consist of introduction, three chapters and general conclusions. The volume of the dissertation is 150 pages, in which are given: 97 equations, 69 figures and 19 tables. Additionally, in the dissertation 101 items are cited.

1

Review of Methods for Sound Source Localization

In this chapter, a review of sound source localization state-of-art is presented. First, various possible acoustic scenarios in which the sound source localization may take place are presented. Then the sound signal acquisition and processing systems (microphone arrays or networks) that are commonly used for the sound source localization task are presented. The theoretical limits of the sound source localization capabilities that depend on the parameters of such systems are discussed. Lastly, state-of-art methods for sound source localization will be presented, with an emphasis on learning-based acoustic source localization methods. The review, presented in this chapter is published in three scientific papers (Sakavičius, Serackis 2019, Sakavičius 2021, Sakavičius, Serackis 2021).

1.1. Acoustic Scenarios of Sound Source Localization

In this section, key concepts regarding acoustics that are used throughout the thesis are defined.

Acoustic scenario referred here to is a collection of parameters of the acoustic medium, acoustic space, the sound field, the sound sources present within the acoustic space, the receivers (microphones) and the associated processing system, as well as the task that is intended to be performed. These parameters will be discussed in detail in this section.

Acoustic medium is the volume filled with particles through which the acoustic waves propagate. Acoustic waves can propagate in any material - gases such as air, liquids such as water, and solid materials such as wood or concrete. The speed of sound within the acoustic medium depends on the stiffness factor and the density of the medium. These measures, especially the density in turn depends on the material composition of the medium and its temperature. In the scope of this thesis, the acoustic medium is dry air unless noted otherwise. All acoustic simulations were carried at the speed of sound $c_s = 340\,\mathrm{m\,s^{-1}}$, corresponding to a temperature of $14.7\ ^\circ\mathrm{C}$. It should be noted, that the speed of sound changes less than $10\,\%$ in a temperature range of $-30\ ^\circ\mathrm{C}$ to $30\ ^\circ\mathrm{C}$. This can have some effect on the performance of sound source localization algorithms (Rabenstein, Annibale 2017), but the effects of variable speed of sound were not investigated in the scope of this thesis.

Acoustic space is any space filled with an acoustic medium in which the acoustic waves might propagate. An ideal acoustic space is isotropic and infinite. In such space, acoustic waves travel without change of their velocity or direction (that is, without refraction or reflection). Also, since there are no acoustic boundaries, other wave propagation phenomena are also absent: scattering, diffraction, and diffusion.

Such infinite and isotropic acoustic spaces do not exist in the real world. In the real world, acoustic spaces inevitably contain boundaries - rooms have walls and furniture, and open spaces are limited at least by the ground. Acoustic boundary is considered to be a limit of space where the velocity of sound changes considerably because of the change in the density of the acoustic medium (i.e., due to change of the material of the acoustic space or the temperature gradient). Acoustic boundaries can refract or reflect the acoustic waves. Acoustic boundaries that occur at the change of the acoustic medium material are called walls or ground in the scope of this thesis, while the acoustic boundaries that are present due to the change of the density of the same material (i.e., air) due to the temperature gradient are not discussed further in this thesis. In the scope of this thesis, acoustic spaces that are only limited by the ground are considered an acoustically open spaces, while the acoustically closed spaces are limited by solid boundaries (walls).

In acoustically open spaces, the sound field is considered a *free field*, which is free from reflected sound waves – only the direct sound energy from the sound source arrives at the receiver. In practice, free field can only be achieved in anechoic rooms. In all other cases, at least the ground reflections exist. Nevertheless, the almost all open spaces – unlimited by walls, and limited only by the ground, can be approximately analyzed as having a free field. Absence of reflections means that there are no phantom, or image sources present in the acoustic scenario, which allows even the most simplistic sound source localization algorithms perform well and localize the sound source with only a few percent mean squared error.

In other hand, acoustic spaces that are limited by walls are considered acoustic enclosures. Sound source localization in acoustic enclosures is much more complex when compared to the localization in open spaces. In acoustic enclosures, sound waves emitted by the sound source tend to reflect from the boundaries of the enclosure creating image sources (see Fig. 1.1).

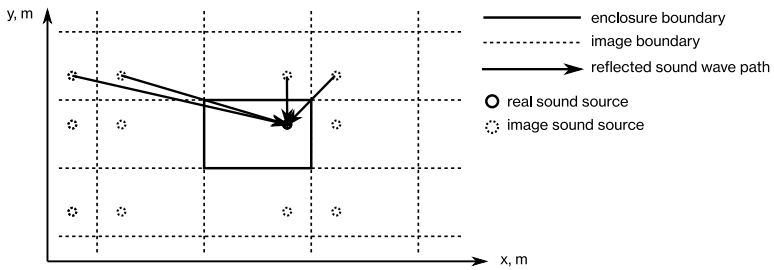


Fig. 1.1. Source images produced by sound wave reflections from the boundaries of the acoustic enclosure

When there are reflections within the acoustic enclosure, sound waves travel between the source and the receiver in multiple paths (direct and reflected), thus the multipath wave propagation occur. These paths are of different lengths, and so is the time that the sound waves take to arrive from the source to the receiver. In other words, the sound waves arriving at the source are delayed proportionally to the length of the path taken by the sound waves. Furthermore, the sound waves are filtered at the boundary, and the transfer function depends on the acoustic properties of the material that forms the boundary. For example, the sound energy can be attenuated more in the high frequency range, while the low-frequency content can be reflected back into the enclosure virtually without much attenuation if the wall is composed of a porous material. Such material may absorb acoustic waves of short wavelength (of comparable lengths to the pores of the material) while the longer acoustic waves are reflected without attenuation. Acoustic waves can reflect many times until their energy is completely absorbed (that is, their energy is converted to heat) by the enclosure surfaces. The number of times the acoustic wave reflects before arriving at the receiver is called the reflection order. For example, if the sound wave is reflected 3 times before reaching the receiver, such sound wave is called a third order reflection.

The most prominent product of acoustic reflections is the reverberation. Reverberation might be viewed as the sum of many delayed acoustic waves arriving at the receiver, each with different delay duration and signal attenuation due to the different sound energy absorption properties of the enclosure boundaries.

Inside the acoustic enclosure, the sound field is not a free field anymore, Depending on the ratio between the direct acoustic wave energy and the energy of the

reflections, also known as direct-to-diffuse ratio (DDR) of acoustic wave energy, the sound field is either of mixed or diffuse type.

Conventional time-delay estimators exhibit dramatic performance deterioration in the presence of multipath signals. This limits their application in reverberant enclosures, particularly when the signal of interest is speech and it may not possible to estimate and compensate for channel effects prior to time-delay estimation (Brandstein, Silverman 1997). Time-delay estimation based locators might not be able to determine the correct time-lag of cross-correlation maximum, because the reflected and delayed signal might introduce another peak in the cross-correlation result, which can sometimes be higher than the peak at the real time-lag.

Generally the receivers obtain the delayed source signal via direct path, but also delayed and filtered copies of the source signals via indirect paths. When the reverberation is strong (the sound field inside the enclosure is the reverberant, or diffuse field), the reflected signals are as strong as the direct signal. This is the main cause of the sound source localization algorithm performance deterioration.

In this thesis, only the sound propagation in enclosures is analysed as it has greater potential to practical use, i.e. in enhancing robotic hearing, teleconferencing, ambient intelligence and security systems, which are mostly used inside rooms and buildings.

An acoustic source is an emitter that creates the vibration of acoustic medium, thus creating acoustic waves of the medium particles. An ideal acoustic source is a point at which the acoustic waves start to emerge in all directions, that is, the ideal source is of negligible dimensions and unidirectional. This source is called point source. The wavefront of such source is spherical. A complex acoustic source might be of non-negligible dimensions. In this case, the source is defined by a collection of points on the surface of the vibrating body of the source. Each point of such complex acoustic source might emit different acoustic waves. The waves emitted by these points are subject to wave superposition and the resulting radiation pattern might not be uniform in every direction. In real world situations ideal acoustic sources do not exist – every source has certain defined dimensions. Yet it is possible to approximate any acoustic as an ideal point source if the dimensions of the source are much smaller (by at least an order of magnitude) than the distance between the source and the receiver – in this case, the wavefront of the emitted wave is almost spherical. Take a guitar for example: each point of guitar's body vibrate when a string is strung. Different points might vibrate in different phases, creating a complex vibration pattern that in turn creates a complex sound field around the guitar. It is possible to analyze every point of a guitar as an independent point source, but this would be a very complex and resource intensive approach. But as we move further from the guitar, the wavefront of all summed point sources approaches a sphere, which is the radiation pattern of an ideal point source. Thus, if we move away from the guitar far enough, we might analyze the guitar as a

point source. Acoustic source might be described by these parameters: source signal spectrum, source signal dynamic range, radiation directivity pattern.

In an ideal unlimited isotropic acoustic space, it is possible to have only one acoustic source. In an acoustic enclosure, though, reflections and thus image sources are always present, and even for a single active sound source within an enclosure, virtually infinite image sources exist. This means that the localization of an acoustic source is more complex in the acoustic enclosure than in an open acoustic space. It is worth noting that the image source waveforms are correlated with the real acoustic source waveform. However, in real world situations, almost always there are more than one acoustic sources that emit uncorrelated waves. Acoustic noise that is present within an acoustic scene can also be considered an acoustic source. For example, environmental noise is almost always present in any acoustic enclosure – more than $26\,\%$ of Lithuanian population is constantly exposed to noise sound pressure levels (SPL) of $55\,\mathrm{dB_{SPL}}$ to $59\,\mathrm{dB_{SPL}}$.

In this thesis, acoustic sources are analysed as point sources unless noted otherwise. The emphasis is put on speech source localization. Speech bandwidth is considered to be in range of $300\,\mathrm{Hz}$ to $4000\,\mathrm{Hz}$, that is, the lengths of speech waves are in range $1.13\,\mathrm{m}$ to $0.085\,\mathrm{m}$. Considering that these lengths are virtually always smaller than the dimensions of the rooms in which the speakers are active, the premise of a speaker being a point source is considered true.

In case of sound source localization, one might wish to localize a single sound source either in a noiseless or noisy environment – in this case, the localization scenario is single source localization or single-versus-many source localization. In cases where more than one simultaneously active sound sources need to be localized, the localization scenario is multiple-source or multiple-versus-many source localization. "Versus-many" indicates that not all sources that are present within the acoustic scene are to be localized, but only one or a few, as opposed to noise sources that are also present in the scene and are often numerous, or spatially distributed. For example, one might wish to localize two speakers within a room, whose mouths (acoustic sources) can be modelled as point sources, and do not need to localize the noise that comes from the windows of the room, which might also not be possible to approximate as point sources due to their dimensions.

The acoustic waves emitted by sound sources in an acoustic scenario are captured by acoustic receivers. Acoustic receivers are devices that convert acoustic energy into other energy form for further processing. Generally, the resulting energy form is no longer influenced by the acoustic events. Most prominent examples of acoustic receivers are ears and microphones. In either case, the acoustic signal is converted to an electric signal (nerve signals or electric current) and can no longer be modified by the acoustics of the room or other acoustic signals.

Acoustic receivers might be broadly described by these parameters: sensitivity, frequency response, directional pattern, Signal-to-Noise ratio (SNR). The

frequency response might be considered a spectral sensitivity measure, while the directional pattern might be considered a spatial sensitivity measure. Together, they define a spectral-spatial response of the receiver. For example, the human ear has a complex spectral-spatial response pattern due to the shape of the cochlea, which help human localize the sound source (Argentieri *et al.* 2015). In this thesis, omni-directional receivers with flat frequency response within the frequency range of the analysed source signals are used unless noted otherwise. This is to not over-complicate the analysis.

Smallest number of receiver array that might be used for acoustic source localization is one (El Badawy, Dokmanic 2018). In such scenario, a complex acoustic structure is used to create a complex spectral-directional sensitivity pattern of a single microphone.

Acoustic receivers are often operated in arrays. This allows the use of array processing algorithms to be used, which will be discussed later in greater detail. Acoustic receivers contained within the array are referred to as array elements. A most simple receiver array consist of two elements. An example of such array might be the human ears. Two element acoustic arrays are referred to as binaural arrays (Löllmann *et al.* 2018). Such array allow to obtain most simple acoustic features: interaural level difference (ILD), interaural time difference (ITD) and interaural phase difference (IPD).

Receiver arrays that have more than two elements might be arranged spatially in a single line (linear array) or in circle (circular array). Such arrays can have any number of receivers. Both linear and circular arrays are considered planar arrays, that is, their elements are positioned on a plane. Array elements can be arranged on a two-dimensional grid, and in this case the array is considered a two-dimensional planar array. An example of such array is Pepper robot microphone system (He *et al.* 2018a). Increasing the number of elements even more also increases the options of spatial arrangement of the elements. 4 element array can be arranged in a tetrahedron shape, with receivers positioned at the nodes of the tetrahedron. This is the simplest non-planar array. More complex non-planar arrays include spherical arrays, for example, the Eigenmike (Löllmann *et al.* 2018). The way the array elements are spatially arranged is referred to as the array geometry.

Generally, array elements are placed at constant distances. The distance between adjacent array elements is called the spatial sampling period. The inverse of spatial sampling period is the spatial sampling rate, and denotes the spatial frequency of array elements.

The largest distance between array the elements is called the array aperture. For binaural arrays, the array sampling period is equal to the array aperture. For all other receiver array geometries, the array aperture is always larger than the spatial sampling period. Uniformly sampled linear arrays are called Uniform Linear Arrays (ULA). Some other cases include non-linear spatial sampling, such as ge-

ometric sampling used in DICIT array (Löllmann *et al.* 2018). Increased array aperture offers better TDoA estimation resolution. The time of flight between the edges of the array translates to number of temporal samples, which in turn defines the minimum time difference between two neighboring TDoA values, and thus the DoA angles. Increased spatial sampling rate offers better frequency response of the array. If the elements of the array are spaced more than a single wavelength apart, it is impossible to unambiguously determine the TDoA of the source at the microphone elements, since it is unknown, how many wavelengths actually fit between the elements. For non-periodic signals, ambiguity might mitigated by selecting a longer analysis window; this will be discussed further in greater detail. To achieve high spatial sampling rate with large array aperture, a large number of receiver elements is needed, raising computational and monetary cost of the array, which might be the limiting factor. By aperture size, receiver arrays can be categorized into compact arrays, where the distance between edge elements is smaller than the shortest wavelength of an incoming acoustic signal, and large aperture arrays, with aperture of comparable size or larger than the shortest received wavelength.

Receiver elements might not be uniformly spaced. In this case in the scope of this thesis the collection of receivers is called not an array, but a distributed sensor network. Sensor networks were investigated by Astapov *et al.* (2015). When a set of distributed receiver arrays are used instead of singular microphones, a configuration is called distributed sensor array.

In all above cases except the single microphone setup, the receiver elements are considered to be omnidirectional. For the sake of completeness, another class of acoustic receiver arrays must be also presented: the ambisonics, or soundfield microphones. Such arrays or array networks can also be used for acoustic source localization, for example, Soundfield SPS200 or Oktava 4D-Ambient (Hack 2015). In this thesis, such soundfield microphones are not further analysed, and generally the focus is on omnidirectional microphone arrays.

In all practical cases, there is more than one active sound source within the considered environment. In real world situation, there can me many simultaneous talkers, noise sources and background ambience. The number of active sound sources can vary throughout the time. Main reasons for this change of active sound sources are:

1. A sound source becomes silent at one moment and becomes active at another.
2. A sound source enters or leaves the considered environment.

A localization algorithm can be designed to perform the following:

1. Localize a single, most prominent sound source within an environment, disregarding any other weaker sound sources.

2. Localize a defined number (known a priori) of most prominent sound sources.

Most single sound source localization algorithms rely on the fact that there are only one prominent of desired sound source within the acoustic scene. In practice, there are always more than one sound source in the acoustic scene, but it possible to select only the most prominent sound source and determine its DoA or position.

One versus many scenario occurs when there are many active sound sources within the acoustic scene, but the localization algorithm is designed to discern one sound source against background noise (which can be a sum of multiple sound sources) and estimate its position.

The main challenge for the multiple sound source localization using features that are based on TDoA estimation (trilateration, GCC-PHAT, CCFB) is that there are no means to determine a priori which source is active. When calculating the cross-correlation function, there might be several peaks in the cross-correlation result (when the signals of the sound sources are similar), and there are no way to determine which of the peaks correspond to real TDoAs (of the same sound source), and which are erroneous, manifesting due to the similarity of the signals of the sound sources.

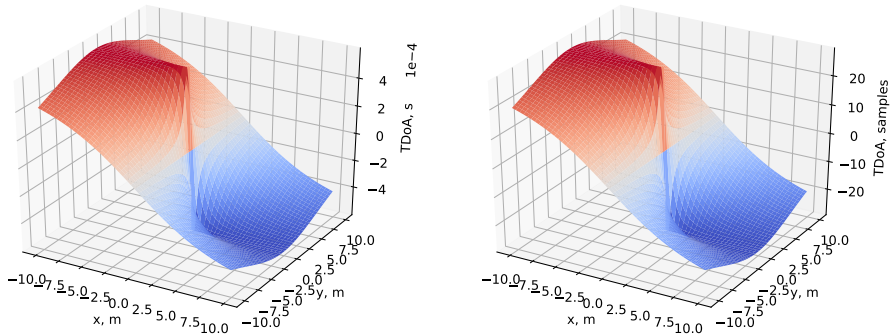
1.2. Theoretical Limits of the Accuracy of the Sound Source Localization

There are an exact theoretical limitations on the resolution of the DoA estimation imposed by the geometry of the microphone array and the sampling rate of the audio signal.

Maximal resolution of the localization is limited by the parameters of the audio system – most importantly – the sampling rate, f_s, and the quantization resolution, Q, which affects the Signal-to-Quantization Noise Ratio, SQNR.

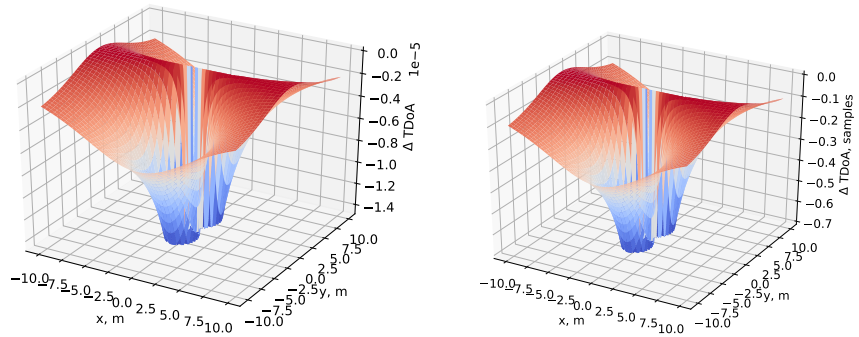
If the difference of TDoA is very small, and resolves to one or few audio samples, it might be difficult to discern estimate the TDoA and in turn the DoA or position of the sound source.

Considering a case where two microphones $\mathbf{m}_1$ and $\mathbf{m}_2$ are placed on the x axis, at $10\,\mathrm{cm}$ to opposite directions from the y axis ($\mathbf{m}_1 = [0; 0.1]\ \mathrm{m}$; $\mathbf{m}_2 = [0; -0.1]\ \mathrm{m}$), and the source is $\mathbf{s} = [(-10 \ldots 10); (-10 \ldots 10)]\ \mathrm{m}$, assuming the speed of sound (group velocity) $v_\mathrm{S} = 340\,\mathrm{m\,s^{-1}}$, the maximal TDoA is $\Delta T_\mathrm{A} = 0.59\,\mathrm{ms}$ If the sampling rate is $f_\mathrm{s} = 48\,\mathrm{kHz}$, this corresponds to $28.2\,\mathrm{samples}$. This is illustrated by Fig. 1.2.



a) **b)**

Fig. 1.2. TDoA as a function of the distance between the source and two listening points; a) z axis in seconds; b) z axis in samples, $f_s = 48\,\mathrm{kHz}$



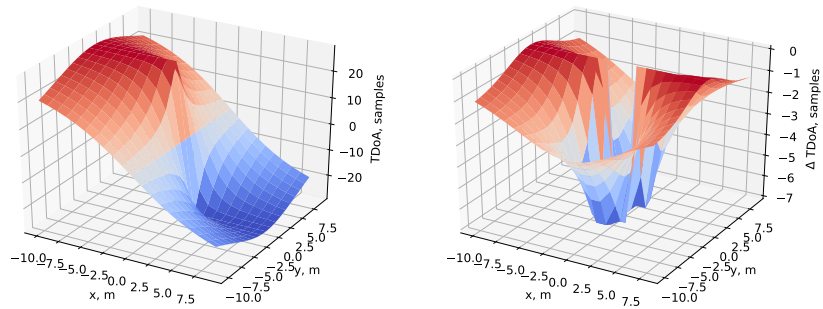
a) **b)**

Fig. 1.3. Difference of TDoA between adjacent grid points (clipped); a) z axis in seconds; b) z axis in samples, $f_s = 48\,\mathrm{kHz}$

The change in TDoA ($\Delta\Delta T_\mathrm{A}$) when the source position is changed by a constant value alongside any of the axes, is different and depends on the angle and the distance between the source and the center of the listening array.

By calculating the derivative, we find that the difference between adjacency points. The grid points onto which the sound source is located, are separated in both x and y directions by $0.1\,\mathrm{m}$.

It can be seen in Fig. 1.3 that the difference between two grid points is less than 1 sample, when the grid is spaced by $0.1\,\mathrm{m}$. Increasing the spacing leads to greater differences in TDoA between adjacent points. This means that the resolu-



a) **b)**

Fig. 1.4. TDoA and the difference of the TDoA between adjacent grid points (clipped) when grid spacing is 1 m, z axis in samples; a) TDoA; b) difference of the TDoA

tion of localization based on TDoA is limited by sampling rate. For example, when the spacing of the grid is increased to 1 m, difference of TDoA between adjacent grid points is in range of single samples (see Fig. 1.4). It can be shown that this difference of TDoA is directly proportional to the grid step and to the sampling ratio.

1.3. Categorization of Sound Source Localization Methods

The selection of the most suitable sound source localization method in a particular situation is dependent on the acoustic properties of the environment where the localization takes place, where the sound sources and receivers are active. Depending on the type of the sound field – be it a free or diffuse field, some SSL algorithms might perform better than others.

Existing source localization procedures may also be loosely divided into three general categories: those based upon maximizing the steered response power (SRP) of a beamformer, techniques adopting high-resolution spectral estimation concepts, and approaches employing time-difference of arrival (TDoA) information (DiBiase *et al.* 2001).

Steered-Beamformer-Based Locators work on the principle of computationally steering the directivity of a microphone array via beamforming, that is, via delay-and-sum or filter-and-sum of the microphone signals. The most simple such locator can be realized by cross-correlating two microphone signals and searching for the time-lag at which the maximum of the cross-correlation product occurs.

Such is the Generalized Cross-Correlation (GCC) locator. By applying the phase transform (PHAT), a widely known GCC-PHAT locator is obtained. Going further, more microphone pairs can be employed, and by calculating the time-lags for all pairs to maximize the response of the microphone array at a certain DoA, a Steered Response Power locator is constructed. Again, by applying the phase transform, SRP-PHAT locator is achieved. Overall, the computational requirements of the focalization-based ML estimator, namely the complexity of the objective function itself as well as the relative inefficiency of an appropriate optimization procedure, prohibit its use in the majority of practical, real-time source locators. Furthermore the steered response of a conventional beamformer is highly dependent on the spectral content of the source signal. Many optimal derivations are based on a priori knowledge of the spectral content of the background noise, as well as the source signal. In the presence of significant reverberation, the noise and source signals are highly correlated, making accurate estimation of the noise infeasible. Furthermore, in nearly all array applications, little or nothing is known about the source signal. Hence, such optimal estimators are not very practical in realistic speech-array environments.

Another class of sound source location methods is high resolution spectral-estimation-based locators. These algorithms tend to be significantly less robust to source and sensor modeling errors than conventional beamforming methods. incorporated models typically assume ideal source radiators, uniform sensor channel characteristics, and exact knowledge of the sensor positions. Such conditions are impossible to obtain in real-world environments. While the sensitivity of these high-resolution methods to the modeling assumptions may be reduced, it is at the cost of performance. Additionally, signal coherence, such as that created by the reverberation conditions of primary concern here, is detrimental to algorithmic performance, particularly that of the eigenanalysis approaches. This situation may be improved via signal processing resources, but again at the cost of decreased resolution.

The third class of sound source locators is TDoA-based locators. Primarily because of their computational practicality and reasonable performance under amicable conditions, the bulk of passive talker localization systems in use today are TDoA-based.

The two major sources of signal degradation which complicate this time delay estimation problem are background noise and channel multi-path due to room reverberations. The noise-alone case has been addressed at length and is well understood. Assuming uncorrelated, stationary Gaussian signal and noise sources with known statistics and no multi-path, the ML time-delay estimate is derived from a SNR-weighted version of the Generalized Cross-Correlation (GCC) function.

In the presence of single-path propagation, maximum likelihood (ML) versions of the GCC- and LS-based time delay estimators have been well studied

and shown to be practical and to obtain theoretical bounds. However, these methods exhibit dramatic performance degradation in the presence of simple multipath channels (a few echoes) as well as the more complex scenario of a reverberant room (a very large number of closely spaced echoes, the equivalent of nearly-flat multiplicative noise in the spectral domain). These shortcomings limit their applicability for time-delay estimation in realistic enclosures.

In the past, some studies have investigated the time delay estimation problem in the presence of a few correlated additive echoes. However, the results obtained cannot be used to predict the effects of reverberation on time delay estimation performance since reverberation consists in the superposition of a very large number of closely spaced echoes, a phenomenon that is more adequately modeled as multiplicative noise in the frequency domain (i.e., convolutional smearing) (Champagne *et al.* 1996).

1.3.1. Generalized Cross Correlation

Given the signals acquired by a couple of microphones, a coherence measure can be defined as a function that indicates the similarity degree between the two signals realigned according to a given time lag. Coherence measures can hence be used to estimate the time delay between two signals. For example, Cross-Correlation is the most straightforward coherence measure (Brutti *et al.* 2008).

The most common approach adopted in the sound source localization community to compute a coherence measure is the use of GCC-PHAT Knapp, Carter (1976). Let us consider two digital signals $x_1(n)$ and $x_2(n)$ acquired by a couple of microphones, GCC-PHAT is defined as follows:

$$\text{GCC-PHAT}(d) = \text{IFFT}(\frac{X_1 \cdot X_2^*}{|X_1||X_2|}), \tag{1.1}$$

where d is a time lag, subject to $|d| < \tau_{\max}$, while X_1 and X_2 are the discrete Fourier transforms (DFT) of x_1 and x_2 respectively and IFFT denotes the inverse fast Fourier transform. The inter-microphone distance determines the maximum valid time delay $\tau_{\max}$. It has been shown that, in ideal conditions, GCC-PHAT presents a prominent peak in correspondence of the actual TDoA. On the other hand, reverberation introduces spurious peaks which may lead to wrong TDoA estimates (Champagne *et al.* 1996).

Our environment is an acoustic enclosure (room), hence the propagation of sound waves is interfered by objects, such as: walls, furniture and people. This interference creates reverberation, or multi-path propagation of the waves. Reverberation could severely effect the performance of many processes done on the

microphone array. Therefore it must be incorporated into the acoustic model to best cope with the realistic conditions.

Let $h'(\vec{d}_m, \vec{d}_s, t)$ denote the room impulse response for both direct-path and reflected paths from the sound source s at location $\vec{d}_s$ to microphone m at location $\vec{d}_m$. Let $v(\vec{d}_s, t)$ be the response describing the characteristics of the microphone m. Since the position and orientation of microphone m are known and fixed, this response function only depends on the source location $\vec{d}_s$. The microphone signal at microphone m can be modelled as follows:

$$x_m(t) = s(t) * h'(\vec{d}_m, \vec{d}_s, t) * v(\vec{d}_s, t) + n_m(t), \tag{1.2}$$

where $s(t)$ is the source signal, $n_m(t)$ is the noise corresponding to the m^{th} channel and $*$ denotes linear convolution.

The impulse response from the source-output to the microphone-output is the convolution of $h'(\vec{d}_m, \vec{d}_s, t)$ and $v(\vec{d}_s, t)$. This impulse response only depends on the source location if we [assume that the] the microphone m is located at a fixed known position forever. Denote the impulse response by $h(\vec{d}_s, t)$. Equation (1.2) becomes

$$x_m(t) = s(t) * h(\vec{d}_s, t) + n_m(t). \tag{1.3}$$

The equation completely describes the signal received at the microphone m where the reverberant channel's impulse response and uncorrelated noise are taken into account (Do 2009: p. 8).

GCC has been a popular method to determine the time-difference of arrival (TDoA) between two microphones in a pair. Then for multiple TDoA values, one can estimate the source location. Take a 4-element microphone array for example (Do 2009: p. 12). If the distance from microphone m to the source $r_m (m = 1, 2, 3, 4)$, the time delay (traveling time) of the signal from the source to that microphone is $\tau_m = \frac{r_m}{c}$ Then the time difference of arrival, TDoA between two microphones m and n can be defined as

$$\tau_{mn} = \tau_m - \tau_n = \frac{r_m - r_n}{c}. \tag{1.4}$$

From this relation one can estimate the source location using several techniques: linear intersection, spherical interpolation, etc.

Recall Equation (1.3) for a microphone k:

$$x_k = s(t) * h(\vec{d}_s, t) + n_k(t). \tag{1.5}$$

Consider a signal at another microphone l:

$$x_l = s(t - \tau kl) * h(\vec{d_s}, t) + n_k(t). \tag{1.6}$$

Note that to be accurate, we would have to include the time delay τk into the source signal $s(t)$, i.e. $s(t - \tau k)$ in Equation (1.5) to show the signal received at microphone k. However, for simplicity, here we normalized so that the time delay from the source to microphone k, τk is 0. In other words, we are only concerned with the relative TDoA, τkl between these two microphones k and l.

The cross-correlation of thee two microphone signals will show a peak at the time-lag where these two shifted signals are aligned, corresponding to the TDoA τkl. The cross-correlation of $x_k(t)$ and $x_l(t)$ is defined as

$$c_{kl} = \int_{-\infty}^{\infty} x_k(t) x_l(t + \tau) \, \mathrm{d}t. \tag{1.7}$$

Taking the Fourier Transform of the cross-correlation results in a *cross-power spectrum*

$$C_{kl}(\omega) = \int_{-\infty}^{\infty} c_{kl}(t) e^{j\omega t} \, \mathrm{d}t. \tag{1.8}$$

Applying convolution properties of the Fourier Transform for (1.7) when transforming it into (1.8), we have

$$C_{kl} = X_k(\omega) X_l^*(\omega), \tag{1.9}$$

where $X_i(\omega)$ is the Fourier Transform of the signal $x_i(t)$ and $\cdot^*$ denotes complex conjugate.

The inverse Fourier Transform of (1.9) gives us the cross-correlation function of the microphone signals:

$$c_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_k(\omega) X_l^*(\omega) e^{j\omega\tau} \, \mathrm{d}\omega. \tag{1.10}$$

The generalized cross-correlation (GCC) of $x_k(t)$ and $x_l(t)$ is the cross-correlation of their two filtered versions. Denoting the Fourier Transforms of these two filters as $W_k(\omega)$ and $W_l(\omega)$, we have the GCC defined as

$$R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (W_k(\omega) X_k(\omega))(W_l(\omega) X_l(\omega))^* e^{j\omega\tau} \, \mathrm{d}\omega. \tag{1.11}$$

We define a combined weighting function $\Psi_{kl}(\omega)$ as

$$\Psi_{kl}(\omega) = W_k(\omega)W_l^*(\omega). \tag{1.12}$$

Substituting (1.12) into (1.11), the GCC becomes

$$R_{kl}(\tau) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{kl}(\omega)X_k(\omega)X_l^*(\omega)e^{j\omega\tau}\,\mathrm{d}\omega. \tag{1.13}$$

The TDoA between two microphones k and l is the time lag τ that maximizes the GCC in the real range limited by the distance between the microphones:

$$\hat{\tau} = \arg\max_{\tau} R_{kl}(\tau). \tag{1.14}$$

In reality, $R_{kl}(\tau)$ has many local maxima thus making it harder to detect the global maximum. The choice of weighting functions $\Psi_{kl}(\omega)$ affect the performance of the GCC.

It has been shown that the phase transform (PHAT) weighting function is robust in realistic environments (Elko, Anh-Tho Nguyen Pong 1997; Silverman *et al.* 2005) even though it is sub-optimal to the maximum likelihood (ML) weighting function which was studied in under reverberant-free conditions.

PHAT is defined as follows:

$$\Psi_{kl} \equiv \frac{1}{|X_k(\omega)X_l^*(\omega)|}. \tag{1.15}$$

Applying the weighting function PHAT from Equation (1.15) into the expression for GCC in Equation (1.13), the Generalized Cross-Correlation using the Phase Transform (GCC-PHAT) for two microphones k and l is defined

$$R_{kl}(\tau) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|X_k(\omega)X_l^*(\omega)|}X_k(\omega)X_l^*(\omega)e^{j\omega\tau}\,\mathrm{d}\omega. \tag{1.16}$$

In an M-microphone array there are $\binom{M}{2}$ or $\frac{M\times(M-1)}{2}$ pairs of microphones. Using GCC-PHAT on any subset Q of these pairings to estimate the TDoA of each pair creates Q TDoA estimates. For each hypothesized point $\vec{x}$ in 3D space of the room containing the sound source, true TDoAs can be calculated for that Q pairs. From the estimated TDoAs $\hat{\tau}_Q(\vec{x})$ and the true TDoAs $\tau_Q(\vec{x})$ one can establish the root mean square (RMS) error as follows

$$E_{\mathrm{RMS}}(\vec{x}) = \sqrt{(\hat{\tau}_Q(\vec{x}) - \tau_Q(\vec{x}))^2}. \tag{1.17}$$

And the source location estimate $\vec{x}_s$ is

$$\vec{x}_s = \arg\min_{\vec{x}} E_{\mathrm{RMS}}(\vec{x}). \tag{1.18}$$

The peak in GCC-PHAT is used to estimate the TDoA. However, under real conditions, the GCC-PHAT is corrupted by reverberation and noise (He *et al.* 2018a). The GCC-PHAT is not optimal for TDoA estimation of multiple source signals since it equally sums over all frequency bins disregarding the "sparsity" of speech signals in the time-frequency (TF) domain and the randomly distributed noise which may be stronger than the signal in some time-frequency bins (He *et al.* 2018a).

1.3.2. Steered Response Power

Steered Response Power (SRP) and SRP with Phase Transform (SRP-PHAT) vectors can be considered the middle ground between the trivial acoustic features like TDoA and ideal features, like room impulse response (RIR) or relative transform function (RTF). SRP-PHAT features are obtainable in real world, are relatively high-dimensional and contain information about sound reflections within the room.

The signal $x_m(t)$ at microphone m is (see (1.3)):

$$x_m(t) = s(t) * h(\vec{d}_s, t) + n_m(t). \tag{1.19}$$

In an M-microphone array system, the unitarily weighted *delay-and-sum* beamformer which has been briefly introduced in Chapter 1 can be created by delaying the microphone signals $x_m(t)$ with appropriate *steering delays*, δ_m with $m = 1, 2, .., M$ to make them aligned in time, and then summing all these time-aligned signals together. Mathematically, it is defined as follows

$$y(t, \delta_1, \delta_2, \ldots, \delta_M) \equiv \sum_{m=1}^{m=M} x_m(t - \delta_m). \tag{1.20}$$

To make the microphone signals time-aligned, the steering delays δ_m can be set to

$$\delta_m = \tau_m - \tau_0, \tag{1.21}$$

where τ_m is the time delay from the source to microphone m and τ_0 is set to the minimum of all time delays τ_i, $i = [1, 2, \ldots, M]$ to make δ_m non-negative and hence system is causal.

Now we can express the output of a delay-and-sum beamformer in terms of the source signal, the channel's impulse response and the noise as follows

$$y(t,\delta_1,\delta_2,\ldots,\delta_M) \equiv s(t) * \sum_{m=1}^{m=M} h(\vec{d}_s, t-\tau_m+\tau_0) + \sum_{m=1}^{M} n_m(t-\tau_m+\tau_0). \quad (1.22)$$

When an adaptive filter is applied to the delay-and-sum beamformer, a *filter-and-sum* beamformer is achieved. In the frequency domain, a filter-and-sum beamformer output is

$$Y(\omega,\delta_1,\delta_2,\ldots,\delta_M) = \sum_{m=1}^{m=M} G_m(\omega) X_m(\omega) e^{-j\omega\delta_m}. \quad (1.23)$$

where $X_m(\omega)$ is the Fourier Transform of the microphone signal $x_m(t)$ and $G_m(\omega)$ is the Fourier transform of the filter.

In general, the steered response power (SRP) is the output power of a filter-and-sum beamformer over all points $\vec{x}$ in a predefined region. For each point $\vec{x}$ it is a function of steering delays, and in the frequency domain is defined as

$$P(\delta_1,\ldots,\delta_M) \equiv \int_{-\infty}^{\infty} Y(\omega,\delta_1,\delta_2,\ldots,\delta_M) Y^*(\omega,\delta_1,\delta_2,\ldots,\delta_M)\,\mathrm{d}\omega. \quad (1.24)$$

Substituting equation (1.23) into equation (1.24), we have

$$P(\delta_1,\ldots,\delta_M) \equiv \int_{-\infty}^{\infty} \left(\sum_{k=1}^{k=M} G_k(\omega) X_k(\omega) e^{-j\omega\delta_k} \right) \left(\sum_{l=1}^{l=M} G_l^*(\omega) X_l^*(\omega) e^{j\omega\delta_l} \right) \mathrm{d}\omega. \quad (1.25)$$

Rearranging the expression yields

$$P(\delta_1,\ldots,\delta_M) \equiv \int_{-\infty}^{\infty} \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} G_k(\omega) G_l^*(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega(\delta_l-\delta_k)}\,\mathrm{d}\omega. \quad (1.26)$$

From the equation (1.21) it is easy to see that

$$\delta_l - \delta_k = \tau_l - \tau_k. \quad (1.27)$$

Inserting (1.27) into (1.26) we obtain

$$P(\delta_1, \ldots, \delta_M) \equiv \int_{-\infty}^{\infty} \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} G_k(\omega) G_l^*(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega(\tau_l - \tau_k)} \, \mathrm{d}\omega. \tag{1.28}$$

Note that the integral converges because in practice the microphone signals and the filters have finite energy. Hence, the summations can be interchanged with the integral and moved outside the integral as follows:

$$P(\delta_1, \ldots, \delta_M) \equiv \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{\infty} G_k(\omega) G_l^*(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega(\tau_l - \tau_k)} \, \mathrm{d}\omega. \tag{1.29}$$

Define the combined weighting function

$$\Psi_{kl} \equiv G_k(\omega) G_l^*(\omega). \tag{1.30}$$

Recall that (1.4) gives us

$$\tau_l - \tau_k = \tau_{kl}. \tag{1.31}$$

Substituting the expressions in (1.30) and (1.31) back into (1.29) gives us the expression for the SRP:

$$P(\delta_1, \ldots, \delta_M) \equiv \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{\infty} \Psi_{kl} X_k(\omega) X_l^*(\omega) e^{j\omega\tau_{kl}} \, \mathrm{d}\omega. \tag{1.32}$$

Now we recall the GCC from (1.13):

$$R_{kl}(\tau) \equiv \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{kl}(\omega) X_k(\omega) X_l^*(\omega) e^{j\omega\tau} \, \mathrm{d}\omega. \tag{1.33}$$

It can easily be seen that the SRP and the GCC have almost identical expressions except that the SRP is summed over all pairs of microphones and there is a constant offset of 2π. Therefore, this provides us a means to calculate the steered response power (SRP) of a microphone array by summing the generalized cross-correlation (GCC) of all pairs of microphones in the array (here the constant offset is ignored since it is just a scalar).

Similar to the idea of GCC-PHAT, when the weighting function *phase transform*, PHAT, is applied to the steered response power (SRP), we obtain the SRP-

PHAT. The SRP-PHAT for each point $\vec{x}$ n the space is defined as follows

$$P(\delta_1, \ldots, \delta_M) \equiv \sum_{k=1}^{k=M} \sum_{l=1}^{l=M} \int_{-\infty}^{\infty} \frac{1}{|X_k(\omega) X_l^*(\omega)|} X_k(\omega) X_l^*(\omega) e^{j\omega\tau_{kl}} \, \mathrm{d}\omega. \quad (1.34)$$

Since the GCC between microphone k and microphone l is the same as between microphone l and k, the elements summing to form the above SRP-PHAT functional form a symmetric matrix with fixed energy terms on the diagonal. Therefore, the art of the SRP-PHAT that changes with $\vec{x}$ is either the upper-part or the lower part of the matrix. In other words, for a particular point $\vec{x}$ in the space, the part of the SRP-PHAT in Equation (1.34) that changes with $\vec{x}$ can be computed by summing the GCC of not all pairs of the M-microphone array, but only a subset Q of the pairs where $Q = [k, l], \forall\ k \in [1, \ldots, M-1], M \geq l > k$:

$$P'(\delta_1, \ldots, \delta_M) \equiv \sum_{k=1}^{k=M} \sum_{l=k+1}^{l=M} \int_{-\infty}^{\infty} \frac{1}{|X_k(\omega) X_l^*(\omega)|} X_k(\omega) X_l^*(\omega) e^{j\omega\tau_{kl}} \, \mathrm{d}\omega. \quad (1.35)$$

To find the source locations, we steer the beamformer over all possible points in a focal volume containing the source. The points that give the maximum weighted output power (SRP-PHAT) of the beamformer will be the source locations. For a single source, the location estimate $\vec{x}_s$ is

$$\vec{x}_s = \arg\max P'(\vec{x}), \quad (1.36)$$

where $P'(\vec{x})$ is the SRP-PHAT at point $\vec{x}$ and is defined in equation (1.35). Note that the calculation of any particular point $P'(\vec{x})$ will be called a functional evaluation (fe).

The hypothesis is that the SRP-PHAT will peak at the actual source location even under very noisy and highly reverberant conditions. However, the problem with SRP-PHAT is its expensive computational cost because the search space has many local maxima, and thus computationally intensive grid-search methods have been required to find the global maximum.

1.4. Assumptions about W-disjoint Orthogonality of the Sound Sources

Assuming the signal to be W-disjoint orthogonal allow to perform signal demixing and blind source separation. Given a windowing function $W(t)$, we call two functions $s_i(t)$ and $s_j(t)$ W-disjoint orthogonal if the supports of the windowed Fourier

transforms of $s_i(t)$ and $s_j(t)$ are disjoint. The windowed Fourier transform of $s_i(t)$ is defined,

$$\mathcal{F}^W(s_i(\cdot))(\omega,\tau) = \int_{-\infty}^{\infty} W(t-\tau)s_i(t)e^{-i\omega t}\,\mathrm{d}t, \tag{1.37}$$

which will be referred to as $S_i^W(\omega,\tau)$ where appropriate. The W-disjoint orthogonality assumption can be stated concisely,

$$S_i^W(\omega,\tau)S_j^W(\omega,\tau) = 0, \forall\, i \neq j, \forall\, \omega, \tau. \tag{1.38}$$

Note that, if $W(t) = 1$, $S_i^W(\omega,\tau)$ becomes the Fourier transform of $s_i(t)$, which we will denote $S_i(\omega)$. In this case, W-disjoint orthogonality can be expressed,

$$S_i(\omega,\tau)S_j(\omega,\tau) = 0, \forall\, i \neq j, \forall\, \omega, \tag{1.39}$$

which is called disjoint orthogonality (Jourjine *et al.* 2000).

Human speech is considered to be approximately W-disjoint orthogonal (Rickard 2002).

1.5. Learning-Based Sound Source Localization

Due to the ability to approximate complex functions that define the relationship between the microphone array signals and features extracted from those signals and the positions or DoAs of sound sources, learning-based sound source localization methods might be further advantageous in such circumstances.

Multiple investigations of application of ANNs for SSL were presented recently (Argentieri *et al.* 2015; Grumiaux *et al.* 2021b). Sound source localization using ANN is commonly formulated either as a classification (Bohlender *et al.* 2021; Chakrabarty, Habets 2019a; Grumiaux *et al.* 2021a; Hao *et al.* 2020; Hirvonen 2015; Hubner *et al.* 2021; Ma, Liu 2018; Roden *et al.* 2015; Vargas *et al.* 2021), or a regression problem (Cao *et al.* 2019; Grondin *et al.* 2019; Kim 2014; Pertila, Cakir 2017; Youssef *et al.* 2013).

In case of the regression problem, the output of the ANN is a one-dimensional, two-dimensional or three-dimensional vector (in case of a single sound source (Huang *et al.* 2020; Park *et al.* 2020)) or a set of vectors (in case of a multiple source localization (Kapka, Lewandowski 2019; Kim, Ling 2011; Yasuda *et al.* 2020)).

In case of the classification problem, the input features are classified to an array of spatial classes, representing the source coordinates in 1, 2 or 3 dimensions.

In case the source is localized in a single dimension, this dimension is most often the azimuth of the sound source (DoA, (Chakrabarty, Habets 2019b; He *et al.*

2018a; Subramanian *et al.* 2021; Takeda, Komatani 2016a; Zermini *et al.* 2016)). In case of 2 dimensions, source azimuth and elevation (2D DoA (Lin, Wang 2019; Noh *et al.* 2019; Perotin *et al.* 2018)). Some authors provide a method to localize a sound source 3 dimensions, either in polar (Roden *et al.* 2015) or Cartesian (Adavanne *et al.* 2019a; Phan *et al.* 2020; Ronchini *et al.* 2020; Singla *et al.* 2020) coordinates. It must be noted that source localization in 3 dimensions is usually approached as a regression problem, and was only investigated as a classification problem by Roden *et al.* (2015) and Takeda, Komatani (2016b).

Methods for the classification-based localization of single (Kucuk *et al.* 2019; Yalta *et al.* 2017) as well as multiple (Perotin *et al.* 2018; Subramanian *et al.* 2021) simultaneously active sound sources were proposed.

A variety of input features were proposed to use with an ANN-based sound source localization methods, ranging from ILD and ITD (Roden *et al.* 2015; Youssef *et al.* 2013) and IPD (Pak, Shin 2019; Sivasankaran *et al.* 2018) to GCC-PHAT (He *et al.* 2018a; Lu 2019; Vesperini *et al.* 2016; Xiao *et al.* 2015) or SRP-PHAT power maps (Diaz-Guerra *et al.* 2021) to magnitude and phase spectrograms of the array signals (Adavanne *et al.* 2018; Kapka, Lewandowski 2019; Lin, Wang 2019; Schymura *et al.* 2021; Zhang *et al.* 2019) and even unprocessed audio waveforms (Chytas, Potamianos 2019; Huang *et al.* 2018; Jenrungrot *et al.* 2020; Pujol *et al.* 2019; Sundar *et al.* 2020; Suvorov *et al.* 2018; Vecchiotti *et al.* 2019; Vera-Diaz *et al.* 2018).

While there are many available ANN types, most commonly, the CNN (Salvati *et al.* 2018; Vargas *et al.* 2021) and RNN (Wang *et al.* 2019) or a combination of both are used (Huang *et al.* 2018; Ma *et al.* 2015; Roden *et al.* 2015; Takeda, Komatani 2016a; Youssef *et al.* 2013) and Autoencoders (Huang *et al.* 2020; Wu *et al.* 2021; Zermini *et al.* 2016) are also investigated.

Since the sound source localization needs to map the input features to a metric coordinates or spatial classes that represent the metric coordinates of the sounds source(s), unsupervised learning methods are not commonly used, as it is impossible for a machine learning algorithm to learn the accurate mapping between the feature space and the physical space without any supervision. Although there are some investigations in semi-supervised (Bianco *et al.* 2020; Moing *et al.* 2021; Takeda, Komatani 2017) or weakly-supervised learning strategies (Opochinsky *et al.* 2019), nevertheless, most often, a supervised learning strategy is employed.

The performance of the mentioned proposed learning-based SSL methods were evaluated using either simulated or real-world sound source datasets, with sound sources captured in either anechoic or reverberant environments.

Signals of the localized sound sources range from narrowband or harmonic signals (Ding *et al.*) to broadband noise signals (Huang *et al.* 2021) to human speech (Diaz-Guerra *et al.* 2021; Subramanian *et al.* 2021) or a variety of

real-world signals and noises (Adavanne *et al.* 2019a; Huang *et al.* 2021; Kapka, Lewandowski 2019).

Generally, for supervised training of an ANN, a large dataset of labeled samples is needed, which is costly to acquire and the case of SSL, it needs to contain a number of multichannel array signal recordings with the positions of the sound source labeled for each frame of the recording. A few of such datasets exist (Guizzo *et al.* 2021; Löllmann *et al.* 2018) and were used extensively by many authors. Another way of obtaining a SSL dataset is to simulate the array signals using acoustic models. Most commonly, an image source method (Allen, Berkley 1976) is used for RIR simulation, which can then be convoluted with dry signals to obtain a rendition of the sound source signals propagated within an acoustic enclosure, with introduced multi-path propagation and associated effects. While the localization of a sound source in a reflection-free environment is often an easier task, most authors investigated sound source localization in a reverberant environments because such scenario is more common in real-life situations.

Most often, the number of the sound sources desired to be localized is known or set in advance, although localization of an a-priori unknown number of sources was also demonstrated (Cao *et al.* 2021; Chazan *et al.* 2019; He *et al.* 2018b).

Although great majority of the reviewed research focused on localization of stationary sound sources, advances in moving source localization were made (Adavanne *et al.* 2019b).

1.5.1. Acoustic Features for Sound Source Localization

Several types of acoustic features that can be used for acoustic sound source localization are discussed further.

Features can be grouped by their dimensionality into low-dimensional and high-dimensional feature groups. The criteria for determining whether the feature is low- or high-dimensional is relative to the number of feature observation. If the the dimensionality of features p is much larger than the number of observations N, often written $p >> N$, then the feature is considered to be high-dimensional.

Time Difference of Arrival (TDoA) is a trivial acoustic feature, that can be estimated using cross-correlation of the signals of the pairs of the microphones within a microphone array, for example, the GCC-PHAT. Knowing the TDoA for several non-co-linear (or non-parallel) microphone pairs, it is possible to estimate the position of the sound source using triangulation (trilateration).

While this would be a simple and straightforward method, the accurate TDoA estimation becomes very tricky in reverberant or noisy environments. Moreover, the TDoA contains only very little information about the distance between the sound source and the microphone pair (just one value per pair). For a microphone array with 4 elements, that's only 6 values. TDoA does not explicitly contain

any information about the structure of reflections withing the enclosure, nor the geometry or acoustic properties of the enclosure.

TDoA is also commonly known as interaural time difference, ITD (in context of binaural or multiaural hearing).

In contrast to the ITD, the interaural phase difference, or IPD for short, operates on much smaller time difference than ITD. IPD is most useful when the receiver array is compact and no spatial aliasing occurs.

High dimensional acoustic features are represented by a vector that has a number of elements that is comparable to a number of observed samples of the feature.

It is assumed that high-dimensional acoustic features, such as room impulse response (RIR) or room transfer function (RTF) contain a unique fingerprint of sound source and microphone positions within an enclosure. This is because the structure of room reflections is unique for every source position and every microphone position (theoretically, there might be some cases when same RIR is obtained for more than one combination of microphone and sound source positions, but this is probably possible in ideal room, which exhibit point symmetry around the center of the room; in real rooms this is impossible; also the microphones must be also placed symmetrically in the enclosure for this effect to occur).

While the RIRs and RTFs contain enough information to uniquely determine the position of the source within an enclosure, in practice it is impossible to obtain RIR without knowing the positions of the sound source and the microphone within the room beforehand. RTFs are a viable option, utilized by Laufer-Goldshtein *et al.* (2016).

1.6. Conclusions of the First Chapter and Formulation of Dissertation Tasks

1. The Steered Response Power with Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. The performance of SRP-PHAT-based source localization algorithms deteriorate considerable when compact microphone arrays are used. Learning-based sound source localization methods might be further advantageous in such circumstances.
2. There are several learning-based source localization approaches, based on either semi-supervised or supervised learning paradigms. In both of these approaches to work, a set of acoustic features from known sound source positions (the labeled dataset) is needed.

3. Labeled feature acquisition is very costly, although is relatively easy to obtain a large dataset of unlabeled acoustic features. Unlabeled acoustic features could be leveraged to improve the performance of learning-based sound source localization methods.
4. High-dimensional acoustic features, such as room impulse response (RIR) or room transfer function (RTF) contain a unique fingerprint of sound source and microphone positions within an enclosure.

Three hypotheses were formulated as a result of the performed literature survey:

1. Learning-based sound source localization approaches perform better in terms of source position estimation error than geometrical or subspace-based methods in adverse acoustic conditions.
2. High-dimensional acoustic features are more suitable to use in learning-based scenarios than the low-dimensional acoustic features because they contain more information about the acoustic scene and acoustic parameters of the enclosure and thus such features are better suited when the acoustic conditions are adverse.
3. It is possible to simultaneously localize more than one active sound source in two and three spatial dimensions in the acoustic scene if the sources exhibit W-disjoint orthogonality.

To test the formulated hypotheses, following tasks should be solved:

1. To propose and investigate the supervised learning-based sound source localization approaches for multiple sound source two-dimensional localization within an acoustic enclosure.
2. To propose and investigate the supervised learning-based sound source localization approaches for multiple sound source three-dimensional localization within an acoustic enclosure.
3. To propose and investigate the semi-supervised and/or unsupervised sound source localization approaches for a single sound source localization within an acoustic enclosure.

2

Theoretical Research of Learning Based Sound Source Localization Methods

In this chapter, the methods and results of the theoretical research of learning-based sound source localization are presented. Firstly, an investigation on the effects of signal thresholding on the accuracy of sound source localization is presented. Secondly, a simple supervised learning method for sound source location estimation using sound intensity features and a wide aperture microphone array is presented. Then an unsupervised learning method for acoustic feature embedding and mapping to the metric coordinate system is presented. Finally, a supervised learning methods for sound source localization in 2D (azimuth and elevation) and 3D (Cartesian coordinates) localization using a CNN trained on noise signal STFTs are presented.

The research results presented in this chapter are published in four papers (Sakavičius *et al.* 2017; Sakavičius, Serackis 2019; Sakavičius 2021; Sakavičius, Serackis 2021) and announced at the international "AIEEE" (Riga, 2017), "eSTREAM" (Vilnius, 2017, 2019) and national "Science – Future of Lithuania" (Vilnius, 2017, 2019) scientific conferences.

2.1. Analysis of Effects of Signal Thresholding

Even though the source localization is well understood and is comparatively reliable in a reflection-less environment, such as acoustically free field, the performance of time difference of arrival (TDoA) based methods, such as generalized

cross-correlation with phase transform (GCC-PHAT) or steered response power with phase transform (SRP-PHAT) deteriorates considerably when the multipath wave propagation due to the reflections of the sound waves at the boundaries of the acoustic enclosure (Brandstein, Silverman 1997; Datum *et al.* 1996; DiBiase *et al.* 2001). When the acoustic wave propagation is analyzed in a real-world environment, additional factors such as: acoustic properties of the enclosure, the size and geometry of the microphone array, the signal-to-noise ratio (SNR) of the microphone array and the associated acquisition system, sampling rate and quantization resolution, duration of the analysis window (in case of a digital signal processing system), must be considered (Xiao *et al.* 2016).

Experimental studies with real sound data (Löllmann *et al.* 2018) yielded dependencies on the accuracy of sound source direction determination, which aimed to evaluate the influence of the following optional parameters: the sampling rate, the oversampling rate, the type of voice activity detector and its parameters. Presented research shows that in order to increase the accuracy of audio source localization, it is necessary to separate the segments of signals received by microphones with and without a useful signal (speech), thus avoiding incorrect determination of the direction of the source.

Audio source localization is performed by performing a cross-correlation of the signals received in the two synchronized microphones, obtaining the peak of the time lag function and the TDoA estimate, from which the angle of the source with the section connecting the pair of microphones is then calculated. To evaluate the accuracy of audio source localization, we have used an audio signals with labeled source coordinates provided in the publicly available database LOCATA (Löllmann *et al.* 2018); source coordinates were labeled with at a sampling frequency of 120 Hz. We have used only one pair of microphones from the 32 microphone array (Eigenmike). The distance between the microphones was 8 cm.

In the following section is presented the investigation of the factors that have an impact of the accuracy of sound source direction of arrival (DoA) estimation via microphone signals cross-correlation, namely, the length and the SNR of the analysis frame. A measure is proposed for cross-correlation time-lag estimate reliability, called signal-to-minimum-error-amplitude ratio, SMEAR.

2.1.1. Application of Cross-Correlation of Two Microphones

The first phase of the study aimed to compare the location of the sound source obtained from the displacement of the cross-correlation peak (that is, the time lag) with the actual location of the sound source calculated from the change in the ground truth coordinates of the speaker as labeled in the dataset. The coordinates of each of the microphones and the coordinates of the sound source in three-dimensional space were used to determine the actual location of the sound source.

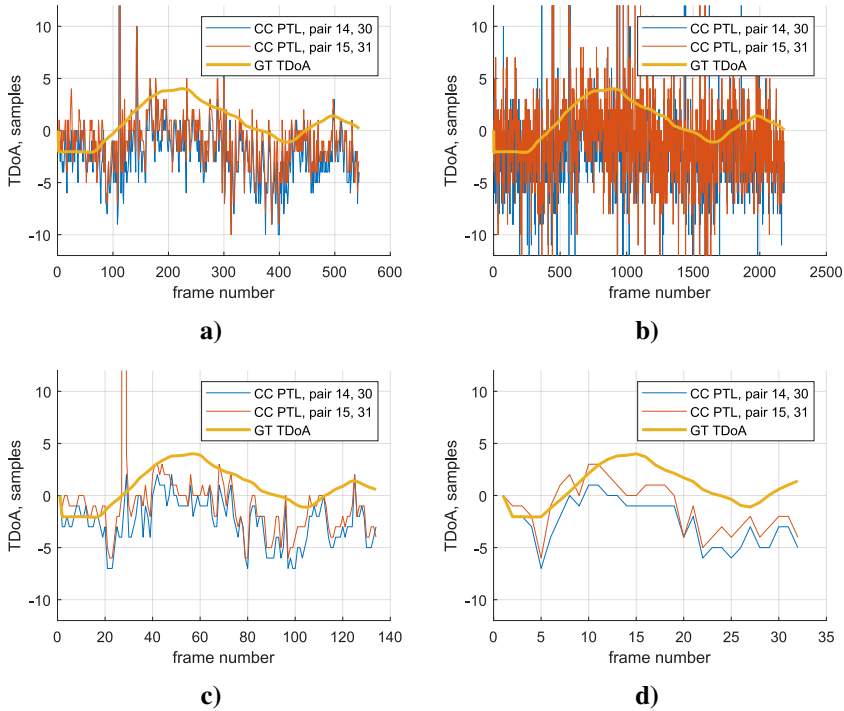


Fig. 2.1. Audio source localization results using different frame length; LOCATA dataset Eigenmike signals, speech source; CC PTL is cross-correlation peak time lag; GT TDoA is the ground truth TDoA; a) 1024 samples; b) 4096 samples; c) 16384 samples; d) 65536 samples

Since the speaker (sound source) was moving within the enclosure on a horizontal plane and the microphone grid was not moving during the recording of the data set, it was rational to use only two spatial coordinates (x, y), ignoring the vertical axis (z). The sound wave time of arrival (ToA) between the sound source and the i-th microphone ToA_i is calculated for each microphone according to the Pythagorean theorem:

$$\mathrm{ToA}_i = \sqrt{(x_s - x_{mi})^2 - (y_s - y_{mi})^2}/v_s, \tag{2.1}$$

where x_{mi}, y_{mi} are the coordinates of the i-th microphone, x_s, y_s are the coordinates of the sound source, v_s – speed of sound in air. After calculating the ToA time for both microphones, we can calculate the time difference of arrival TDoA_{ij}, which should coincide with the displacement of the correlation peak:

$$\mathrm{TDoA}_{ij} = \mathrm{ToA}_i - \mathrm{ToA}_j. \tag{2.2}$$

The results of the study (Fig. 2.1) were obtained using signal analysis frames of different lengths. No frame selection algorithm was used in the study (i.e., based on the signal Zero Crossing Rate (ZCR) or Short-Time Energy (STE)).

As can be seen from the presented trajectories of the sound source motion (Fig. 2.1), the noise of the source DoA estimation from the TDoA obtained from cross-correlation peak time lag decreases with increasing analysis frame length. Nevertheless, the error remains considerable.

Longer frames of analysis can give a more accurate estimate of the delay time difference compared to shorter frames. This can be attributed to the consideration that the speech signal contains both periodic portions (Fig. 2.2a) and also has expressed transient envelopes (Fig. 2.3). For shorter analysis frames, a portion of a signal might not contain a transient and only contain the periodic signal. If the wavelength of such signal within the analysis frame is shorter than the distance between the microphones, the TDoA estimation from the cross-correlation time lag becomes ambiguous, as one can not certainly determine whether the time lag was obtained for the same period of the wave or if the time lag contained more than one periods (Fig. 2.2b). If the analysis window is longer, there is a higher probability that a non-periodic, transient envelope of the signal is contained within the frame, for which the cross-correlation time lag estimate is robust (Fig. 2.4).

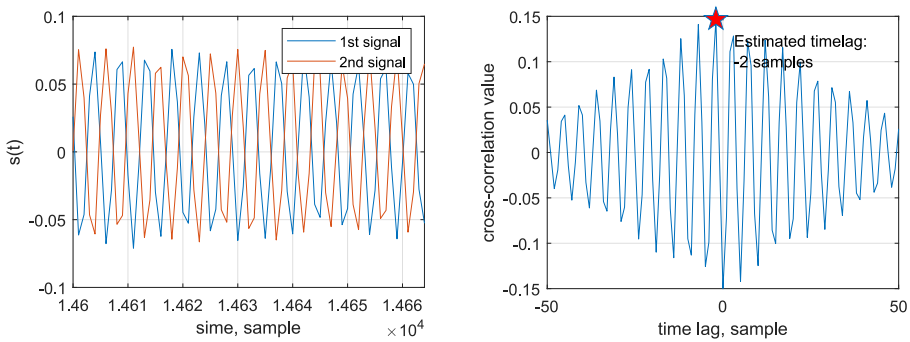


Fig. 2.2. Comparison of synthetic signals; a) received by microphones; b) their correlation result using 64 sample analysis frame; time lag estimation is ambiguous

By calculating the correlation for the 64 sample analysis frame, we obtain a clear correlation maximum corresponding to the signal delay (7 samples) (Fig. 2.4). However, such transient might not be included in a frame of the same length; in this case, an incorrect the correlation peak (incorrect time lag) (Fig. 2.2). The length of the analysis frame is considered too small to calculate the correct difference

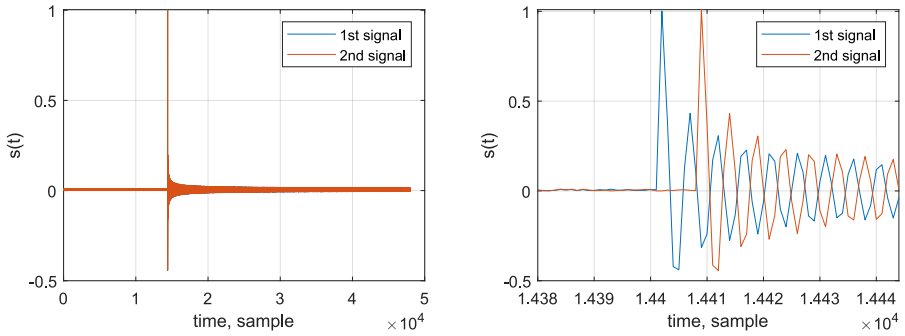


Fig. 2.3. Comparison of synthetic audio signals received by two microphones: a) entire signal; b) transient portion of the signal

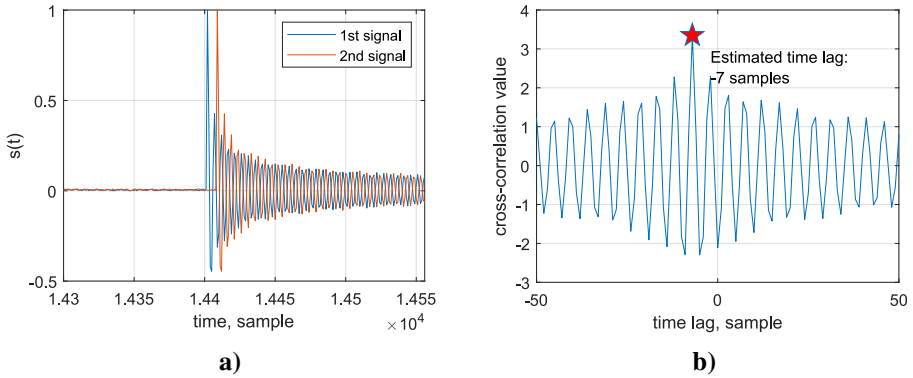


Fig. 2.4. Comparison of synthetic signals; a) received by microphones; b) their correlation result using 256 sample analysis frame; signal contains a transient envelope

in delays from the envelope variation because the signal noise amplitude is larger than the envelope variation in the analysis window (Fig. 2.3). The signal graph shows that the time lag between the signals relative to each other is about 180°, and indeed more (several periods); in this situation it is impossible to determine the correct time lag.

If the selected analysis frame is longer, there is an increased probability of a transient occurring in such window, and the change in the signal envelope will be greater than the amplitude of noise. The influence of the envelope of the signal on the correlation result is greater than the influence of noise.

2.1.2. Signal Amplitude to Minimum Error Amplitude Ratio

The speech signal has exhibits time-varying properties. Some phonemes in some words are very similar to random noise while others exhibit signal periodicity. Moreover, the amplitude of the speech signal is also inconsistent and contains transients. By analyzing and comparing the similarities and differences in the amplitudes of the audio signals recorded by the two microphones, it was observed that the influence of noise on the low amplitude signal can significantly affect the calculated correlation result and lead to incorrect estimation of the cross-correlation time lag. For this reason, it was decided to investigate how the accuracy of sound source localization changes in correlation with selecting only those audio signal frames in which the ratio between the signal amplitude range and the noise amplitude range exceeds a certain threshold.

We speculate that selection of frames based on such threshold would increase the accuracy of the source DoA estimation as the frames which produce unreliable TDoA estimates would be filtered out. Rationale for this would be that some audio frames would contain noisy audio signals. For such frames the cross-correlation time lag can not be reliably obtained and such frame is unusable for DoA calculation.

We speculate that the cross-correlation time lag can be considered reliable for a frame that exhibits a high coherence of signals at a single time lag. We select amplitude of the difference of the signals (the error amplitude) as the coherence measurement. Lower error amplitude of the signals within a frame indicates high coherence of the signals. Since the real TDoA of the signals is unknown, we measure the signal coherence at every time lag within the limits set by the microphone array geometry: the maximum time lag must be lower than the time a sound wave takes to propagate between the microphones. We select the lowest error amplitude and hereafter call it the minimum error amplitude (MEA).

We also consider that a reliable cross-correlation time lag can be obtained for a frame which exhibits a large SNR, since, as it was shown previously, noisy signals can lead to incorrect time lag estimates. We calculate the SNR of the frame as the ratio of the signal amplitude to the MEA. We call this ratio hereafter the SMEA ratio or SMEAR. We investigate the influence of SMEAR thresholding in the following section.

2.1.3. Influence of Thresholding on the Localization Accuracy

The accuracy of the source DoA estimation using SMEAR thresholding with various threshold values: 2, 3 and 5 has been estimated.

The DoA estimates were compared with the ground truth DoA obtained from the dataset source position labels. The results of this investigation are presented in Fig. 2.5.

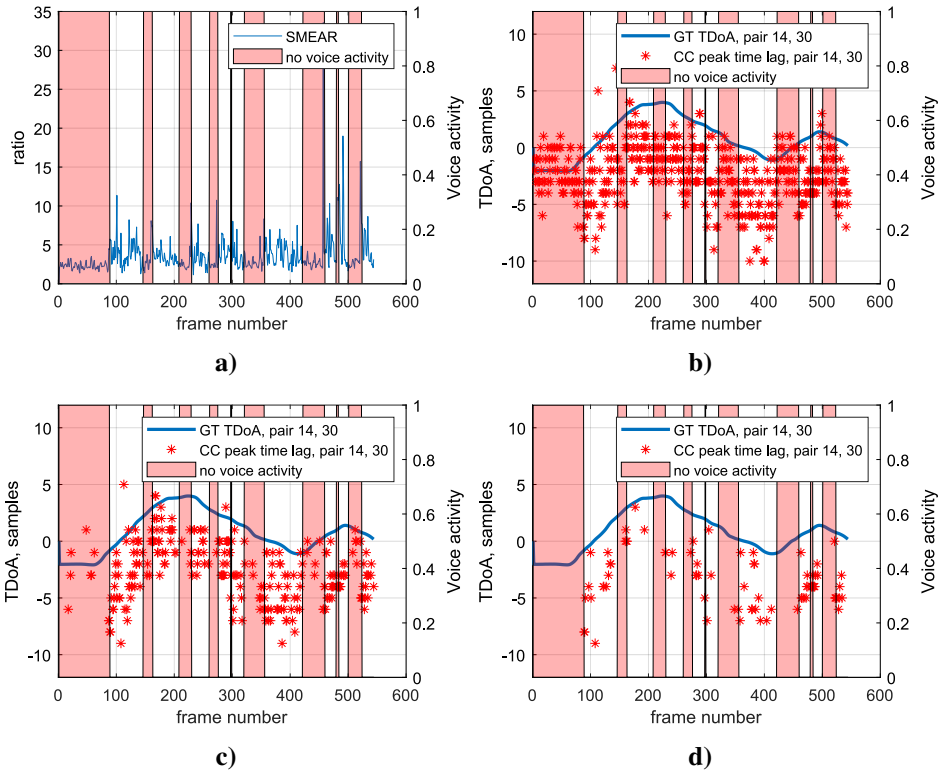


Fig. 2.5. Illustrations of: a) SMEAR calculated for audio signal frames; audio source positioning deviations by selecting different SMEAR threshold values: b) two; c) three; d) five; LOCATA dataset Eigenmike signals, speech source; frame size 4096 samples

Despite the fact that the introduction of thresholding of audio signal frames has slightly improved the localization of the audio source, there is still too much uncertainty (from one to several tens of degrees). The reason for this situation is illustrated by the comparison of multiple signal analysis frames at different recording locations (Fig. 2.6). It can be seen from the figure that even high-amplitude signals, with low amplitude noise levels, can lead to an incorrect setting of the time lag estimation, depending on the time of the signal it will be calculated.

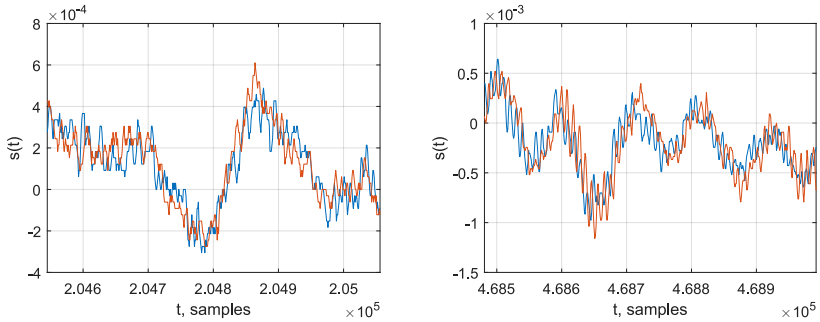


Fig. 2.6. Comparison of signals recorded by two microphones at different times; frame numbers: a) 800; b) 1831; LOCATA dataset Eigenmike signals, speech source

Analysis of speech signal recordings received by a pair of microphones shows that even at a small distance of 8 cm between two microphones, due to room acoustics, time-varying signal characteristics and other distortions, accurate localization of the sound source is not possible with real world signals using signal cross-correlation without additional processing. The study showed that by considering the signal amplitude to noise amplitude ratio, some of the erroneous results of the sound source localization can be eliminated, but other types of noise remain, making signal analysis ineffective on the time axis.

2.2. Single Sound Source Localization Using Multilayer Perceptron

In this section, a sound source localization system based on a MLP is investigated. The proposed system uses four microphones array for sound recording, a feature extraction stage, and an MLP for estimation (prediction) of sound source location.

Sound source propagation is subject to inverse-square decay of the sound intensity with respect to the distance between the sound source and the sound receiver (microphone). This is true for point sources, which emit acoustic waves that have a spherical wavefront. All acoustic sources can be approximated as point sources if the distance between the receiver and the source D is much greater than the dimensions of the sound source d: $D >> d$. For a spherical wavefront, the area of the wavefront A is related to the distance from the source via the spherical equation:

$$A = 4\pi r^2. \tag{2.3}$$

The sound intensity I that is propagated through an unit area of the wavefront is inversely proportional to the area of the wavefront sphere. Therefore, $I \propto \frac{1}{r^2}$. This implies that sound intensity differences are expected at receivers that are placed at different positions with respect to the sound source.

Since the distances between the sound source and each of the microphones are different, it is theoretically possible to uniquely assign one position to one set of amplitudes' ratios. Neither TDoA nor signal phase information is not taken into account.

Sound source localization by evaluating the amplitude ratios between signals of spatially distant microphones is a multidimensional approximation problem. A multidimensional function that is being approximated maps the microphones' signal amplitude values to sound source coordinates.

This multidimensional mapping problem can be viewed as a regression problem and solved using an ANN. In this section, an evaluation of the suitability of MLP as the regressor for the ILD feature regression to acoustic source coordinates is presented.

It can be shown that the sound intensity decreases by approximately $6\,\mathrm{dB}$ for each doubling of distance from the sound source. Therefore, the ILD obtained by the microphone array depends on the distances between the microphones (the array aperture, a) in the array as well as the distance between the sound source and the microphone array. If $a = D$, $\mathrm{ILD} = 6\,\mathrm{dB}$. It is intuitive that by increasing D, ILD becomes smaller:

$$\lim_{(D/d)\to\infty} \mathrm{ILD} = 0. \tag{2.4}$$

Therefore, the microphone array aperture size should not be $a << D$. Practically, the minimum array aperture shall not be more than 10 times smaller than the maximum D. In the case of this investigation, a room of dimensions A square array of four microphones was used in this investigation, situated on the same plane with 0.6 m distance between microphones on a perimeter of the square. The differences in sound signal amplitude, commonly known as Interaural Level Difference, ILD, between separate microphones were used as the main feature for sound source localization.

The amplitudes of the microphone array signals vary in time. Therefore, a Root-Mean Square (RMS) estimates of the recorded signal frames have been used as an input feature that were presented to the MLP. The only two outputs of MLP were used as two estimated coordinates of the sound source on a plane.

2.2.1. Microphone Array Signal Modeling

Every i-th virtual sound source ($i \in (1, 2 \ldots, I)$, I is the number of sound sources) was modeled as a discrete sine signal $s_i(n)$ with frequency f_i, amplitude A_i and

sampling rate F_s ($F_\mathrm{s} = 48$ kHz was used throughout this investigation):

$$s_i(n) = \begin{cases} A_i \sin(2\pi f_i \frac{n}{F_s}), & 0 < n < N_i; \\ 0, & n < 0, N_i < n, \end{cases} \tag{2.5}$$

where N_i is the duration of the signal (number of samples).

The position of the i-th sound source is defined by vector $\mathbf{S}_i = (S_{i_1}, S_{i_2}, \ldots, S_{i_D})$. D is number of spatial dimensions. In this investigation, $D = 2$.

Signal of the j-th virtual microphone ($j \in (1, 2 \ldots, J)$, J is the number of microphones) $m_j(n)$ is created by scaling the signal if the i-th sound source $s_i(n)$ by value $\Delta_{i,j}$, proportional to distance between the j-th microphone and the i-th sound source (to account for the sound level to distance inverse proportionality):

$$m_j(n) = \Delta_{i,j} \cdot s_i(n). \tag{2.6}$$

Microphone position is defined by a vector of its coordinates $\mathbf{M}_j = (M_{j_1}, M_{j_2}, \ldots, M_{j_D})$. Decay value for each pair of i-th sound source and j-th microphone $\Delta_{i,j}$ is

$$\Delta_{i,j} = \frac{1}{d(\mathbf{S}_i, \mathbf{M}_j)}, \tag{2.7}$$

where $d(\cdot)$ denotes the Euclidean distance between the microphone and the sound source position. Distances between the i-th sound source and the microphone array and the corresponding decay values are shown in Fig. 2.7. In this figure, $\mathbf{O} = (0, 0, 0)$ denotes the origin point of the coordinate system, which corresponds to the center of the microphone array. $\Delta_{i,\mathrm{O}}$ denotes distance from the i-th sound source to the origin of the coordinate system.

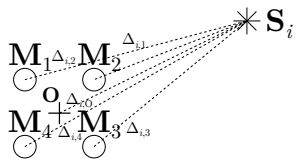


Fig. 2.7. Distances between the sound source and the microphone array

In this investigation, the localization of non-static (moving or irregularly active) sound sources is discussed. Thus, a static sound source position can only be estimated for a short time T_τ, which we call a time frame duration. There are $N_\tau = T_\tau \cdot F_\mathrm{s}$ signal samples in one time frame.

Each microphone's signal is divided into non-overlapping time frames $\tau_{j,k}(n_k) = m_j(n_k); n_k \in (k \cdot N_\tau, \ldots (k+1) \cdot N_\tau)$ here k is time frame number, $k \in (1, 2, \ldots, N/T_\tau)$, here N is the total duration of the measurement. Incomplete time frames are not discarded. For the experiment, a time frame duration T_τ of 0.2 s was chosen.

RMS value $\tau_{\mathrm{RMS}_{j,k}}$ of time frame k of the signal of the j-th microphone is calculated by

$$\tau_{\mathrm{RMS}_{j,k}} = \sqrt{\frac{1}{N_\tau} \sum_{n=k \cdot N_\tau}^{(k+1) \cdot N_\tau} |\tau_{j,k}(n)|^2}. \tag{2.8}$$

A set of 4 RMS values (for each microphone) of k-th time frame $\mathbf{A}_k = (\tau_{\mathrm{RMS}_{1,k}}, \tau_{\mathrm{RMS}_{2,k}}, \ldots, \tau_{\mathrm{RMS}_{4,k}})$ comprises one input sample to the MLP (MLP one input per microphone).

2.2.2. Selection of the Neural Network Structure

The structure of the MLP is presented in Fig 2.8. Two hidden MLP layers were selected because the previous research with only one hidden MLP layer did not provide acceptable source localization results.

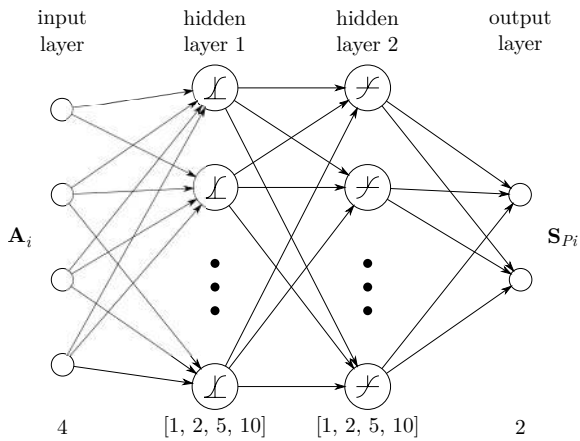


Fig. 2.8. Structure of the MLP

The input feature of the MLP are the 4 amplitudes obtained in the previous step.

MLP has 2 outputs, at which the predicted coordinates $\mathbf{S}_{\mathrm{P}i} = (x_{\mathrm{P}i}, y_{\mathrm{P}i})$ are presented. The signal flow is illustrated in Fig. 2.9.

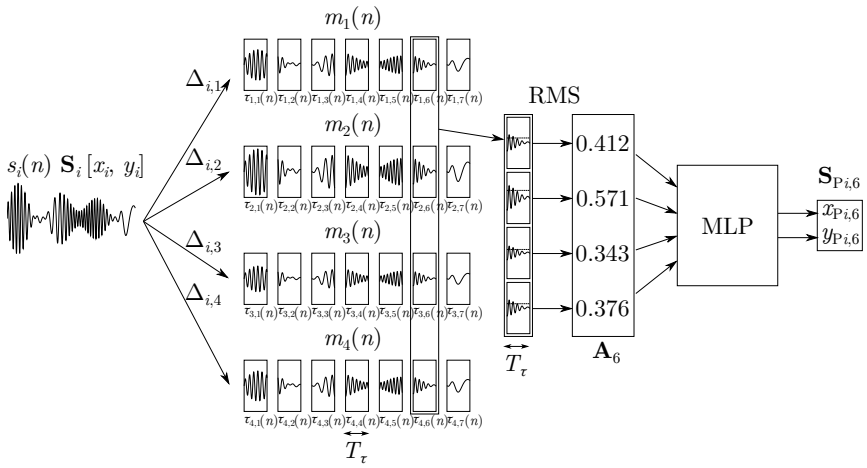


Fig. 2.9. Signal flow diagram of the simulation (6-th time frame is selected)

Positions of the source, predicted by the MLP, were compared to the ground truth sound source positions. A relative error of the i-th sound source position prediction was calculated by

$$e_{\mathrm{rel}_i} = \left(\frac{d(\mathbf{S}_{\mathrm{P}_i}, \mathbf{O})}{d(\mathbf{S}_i, \mathbf{O})} - 1 \right) \cdot 100\%. \tag{2.9}$$

To separately evaluate the ability of the MLP to predict the direction of the sound source relative to the orientation of the microphone array, and the distance from the center of the array to the sound source, the angular error e_{ang_i} and the distance error e_{dist_i} were calculated:

$$e_{\mathrm{ang}_i} = \left| \frac{\theta_{\mathrm{P}_i}}{\theta_i} \right|; \quad e_{\mathrm{dist}_i} = \frac{r_{\mathrm{P}_i}}{r_i}, \tag{2.10}$$

where $\theta_{\mathrm{P}_i} = \tan^{-1}(y_{\mathrm{P}_i}/x_{\mathrm{P}_i}) \cdot 180/\pi$ and $\theta_i = \tan^{-1}(y_i/x_i) \cdot 180/\pi$ (angles expressed in degrees); $r_{\mathrm{P}_i} = \sqrt{x_{\mathrm{P}_i}^2 + y_{\mathrm{P}_i}^2}$ and $r_i = \sqrt{x_i^2 + y_i^2}$. These error values were calculated for every of the testing samples.

In order to test the concept of the proposed system, computer-based simulations were performed at first and experiments with real-world microphone array recordings were performed afterwards, that are presented at length in Section 3.1.

2.3. Sound Source Localization Using Graph Regularized Neural Network

Labeled feature acquisition is considered to be very costly. It is relatively easy to obtain a large dataset of unlabeled audio features. Considering our setting, it is relatively easy to collect a large amount of acoustic features without labels (in our case – coordinates of the sound source), and it is very tedious to provide labels for such data. As for the unsupervised learning approaches, while some proposed algorithms are able to find the relative source distance, they are not bound to any physical dimensions. Therefore, it is desired to leverage the unlabeled data as much as possible. Thus, it is here focused on semi-supervised learning approaches. In this section, a theoretical background for this investigation is provided.

As stated in the previous section, it is assumed, that high dimensional acoustic features lie on a low-dimensional manifold, embedded in a high-dimensional feature space. Since the acoustic features are only dependent on the coordinates of the sound source, it is expected that the manifold would represent the spatial relations between the acoustic features. Acoustic features that may be used for sound source localization are reviewed in Section 1.5.1. In this investigation, SRP-PHAT feature was chosen because of its high dimensionality and exhibited spatial smoothness.

2.3.1. Acoustic Feature Acquisition

The most important property of all acoustic features in this investigation is the spatial smoothness of feature space. In other words, acoustic features are similar to each other for sound source positions that are close together. In this investigation, SRP-PHAT spatial spectra are used as acoustic features. Acoustic features were obtained within an acoustic enclosure using a single sound source, z coordinate was fixed at height $\mathrm{m_s}$. $\mathrm{N_M}$ circular microphone arrays were used for acoustic signal acquisition, each with $\mathrm{N_m}$ microphone elements and radius $\mathrm{m_M}$. Planes of the microphone arrays were parallel to the ground. Both arrays were held at a fixed height $\mathrm{m_M}$. Signals of the microphones are recorded at a fixed sampling frequency f_s and a fixed resolution Q.

The unlabeled dataset may be obtained from an array audio recording where the sound source is slowly moving inside the acoustic enclosure. The maximal speed of the sound source movement $v_{\mathrm{s\,max}}$ should be lower than the maximum expected localization error distance $e_{\max}$ per frame duration T_τ:

$$v_{\mathrm{s\,max}} = \frac{e_{\max}}{T_\tau}. \tag{2.11}$$

The labeled dataset may be obtained from an array audio recordings where the sound source is stationed at a known position $\mathbf{s}_{(x,y,z)}$, described by coordinates (x, y, z) in Cartesian coordinate system within the acoustic enclosure and is producing signal (speech or noise) for a period of $T_{\mathbf{s}}$ seconds. A collection of $n \in \mathrm{N}_{\mathbf{s}}$ recordings at fixed source positions may be obtained.

Audio signals obtained from the microphone arrays are split into frames of duration T_τ seconds to obtain N_τ frames.

For each audio frame $j \in N_\tau$ and for each microphone array $i \in \mathrm{N_M}$, a set of time-frequency representations of the microphone signals is calculated with N_{STFT} FFT points, without frame overlap and no windowing function.

A SRP-PHAT spatial spectrum $\mathrm{SRP}_{\text{SRP-PHAT}(j,i)}$ is obtained for each frame and for each array. $\mathrm{SRP}_{\text{SRP-PHAT}(j,i)}$ is a vector with $\mathrm{N_{SRP}}$ elements, representing the received acoustic power at a particular DoA and covering an azimuth angle $\theta_{\mathrm{M}} \in [0°; 360°]$. SRP-PHAT spectra of all arrays are then concatenated per frame to obtain the acoustic feature SRP_j of $\mathrm{N_M} \cdot \mathrm{N_{SRP}}$ elements.

If the audio recording has an associated location label (known coordinates), a frame is assigned the position label $\mathbf{s}_{(x,y,z)}$.

It is considered that the sound source might not be active at all times, and that the signal is non-stationary (in the case of a speech signal, it might be considered quasi-stationary for frames that contain only one phoneme or a part of a phoneme). Thus, in case of an audio frame where the source is not active, the DoA of a sound source can not be determined, and the acoustic feature is considered to contain only noise. Such frames are to be discarded. For the selection of the audio frames in which the acoustic feature is usable, a thresholding algorithm was used. A metric $p_{i,j} = f(\mathrm{SRP}_{\text{SRP-PHAT}(j,i)})$ is calculated for and compared to the threshold level $L_{\mathrm{thr.}}$ which is the scaled mean of the metric of all obtained frames $L_{\mathrm{thr.}} = \mathrm{k}_p \frac{1}{N_\tau} \sum_{j \in N_\tau} p_j$, where k_p is the scaling coefficient used to control the threshold value. Metric $p_{i,j}$ is calculated per array to address the fact that the arrays might be not identical in terms of audio signal gain, signal-to-noise ratio, and frequency response. The metrics used to evaluate the fitness of the acoustic feature of a particular audio frame are:

1. Root-mean-square value of the SRP-PHAT spectrum:

$$p_{i,j}^{\mathrm{RMS,}}(\mathrm{SRP}_{\text{SRP-PHAT}(j,i)}) = \sqrt{\langle \mathrm{SRP}^2 \rangle}. \tag{2.12}$$

2. Crest factor of the SRP-PHAT spectrum:

$$p_{i,j}^{\mathrm{CF}}(\mathrm{SRP}_{\text{SRP-PHAT}(j,i)}) = \frac{|\max(\mathrm{SRP}_{\text{SRP-PHAT}(j,i)})|}{p_{i,j}^{\mathrm{RMS,}}(\mathrm{SRP}_{\text{SRP-PHAT}(j,i)})}. \tag{2.13}$$

After determining $p_{i,j}$ of the $\mathrm{SRP}_{\text{SRP-PHAT}(j,i)}$ per array per frame, feature vectors SRP_j are selected of those frames j for which $p_{i,j} > L_{\text{thr.}}$ for all microphone arrays $i \in \mathrm{N_M}$.

The labeled dataset is split into training and testing subsets by randomly selecting samples from $\mathrm{N_{ts.}}$ source positions for training and the rest of the source positions $\mathrm{N_{tr.}} = \mathrm{N_s} - \mathrm{N_{ts.}}$ for testing from the entire set of labeled source positions. The following operations are performed separately for training and testing labeled datasets.

2.3.2. Acoustic Manifold Embedding Learning

Manifold embedding can be learned using a Nonlinear Dimensionality Reduction (NLDR) algorithm, such as isometric mapping (ISOMAP), t-distributed stochastic neighbor embedding (t-SNE) or locally linear embedding (LLE), among others. ISOMAP NLRD algorithm was employed to obtain the high-dimensional feature embeddings in low-dimensional space, that is, learn the acoustic feature manifold.

One of the earliest approaches to manifold learning is the ISOMAP algorithm. ISOMAP can be viewed as an extension of Multi-dimensional Scaling (MDS) or Kernel Principal Component Analysis (PCA). ISOMAP seeks a lower-dimensional embedding which maintains the geodesic distances between all points Pedregosa *et al.* (2011).

SRP-PHAT features from both labeled and unlabeled training datasets are embedded into $\mathrm{D_{emb.}}$-dimensional embedded space using ISOMAP, with $\mathrm{k_{emb.}}$ nearest neighbors considered. For each SRP_j feature, an embedding $\mathbf{Z}_j = [z_{d_1}, z_{d_2}, \ldots, z_{d_{\mathrm{D_{emb.}}}}]$. This way, a low-dimensional representations of the high-dimensional acoustic features are obtained. Moreover, the learned manifold corresponds to the spatial structure of the acoustic feature space. Thus, the relative distances in the embedded space of unlabeled features to labeled features are known.

2.3.3. Preparation of the Graph dataset

The combined dataset for training the neural network is comprised of two datasets: $\mathrm{N_u}$ acoustic feature samples without source position labels (the unlabeled dataset) and $\mathrm{N_l}$ acoustic feature samples with source position labels (the labeled training dataset). Each sample feature $\mathrm{SRP}_j^{\mathrm{u,l}}$ in the combined dataset also has a corresponding ISOMAP embedding $\mathbf{Z}_j^{\mathrm{u,l}}$. In order to train the GRNN with graph regularization, the dataset must be preprocessed: for each sample, regardless of whether it is a labeled or an unlabeled sample, alongside the main feature, neighbor features SRP_j^n and their weights a^n where $n \in \mathbf{n}$ must be introduced. $\mathbf{n}$

denotes the neighborhood of the sample feature in the embedded space. This is done by first determining the $\mathrm{k_G}$ nearest neighbors of a particular sample in the embedded feature space and then appending those features as well as their weight coefficients to the training sample.

The dataset for training the neural network is comprised of N source position and SRP-PHAT feature vector pairs.

In the embedded space, Euclidean distances are calculated between every point. The distances between each data sample constitute the distance matrix, which is in turn used to calculate the affinity matrix.

In the embedded space, Euclidean distances are calculated between every feature. The distances between each data sample constitute the distance matrix $\mathbf{D}$, which is in turn used to calculate the affinity matrix. Affinity matrix $\mathbf{A}$ is calculated by subtracting $\mathbf{D}$ from 1: $\mathbf{A} = 1 - \mathbf{D}$. The distance matrix contains the Euclidean distances between each sample in the low-dimensional embedded space:

$$\mathbf{D} = (d_{ij}); \tag{2.14}$$

$$d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2^2, \tag{2.15}$$

where $\mathbf{p}_i = (\alpha_i, \beta_i)$ is the point coordinate vector in the embedded space (in case of $N_{\mathrm{ISO}} = 2$), α and β are the Cartesian coordinates in the embedded space.

Neighbor weights are inversely proportional to the Euclidean distances between the main feature and the neighbor features in the low-dimensional embedded space. Affinity matrix is scaled to the range of $[0; 1]$. An example of an affinity matrix is presented in Fig. 2.10.

For the training of the GRNN, each training sample must contain the main SRP-PHAT feature and $\mathrm{k_G}$ neighbor SRP-PHAT features (used for calculating the graph loss). Additionally, each neighbor feature is associated with its weight a, which is the corresponding element in the affinity matrix. To obtain the $\mathrm{k_G}$ neighbors of each sample, each row of the affinity matrix is thresholded so that only the $\mathrm{k_G}$ highest-valued elements remain their values, while the other row elements are set to zero. The dataset is then expanded so that each sample now has associated neighbor SRP-PHAT features (the indices of which are the non-zero elements in the rows of the affinity matrix).

For the training dataset, a flag m denoting whether the sample is labeled or unlabeled is introduced. This flag holds the value of either "True" or "False" (1 or 0). The content of this field is interpreted by the GRNN during the calculation of the loss function. Effectively, the supervised loss component is multiplied by the flag. In the case of an unlabeled sample, the supervised loss is ignored, and only the graph loss is considered. In real-world scenarios, GRNN expects all fields, including the target feature (the label, the coordinates of the source) to be passed

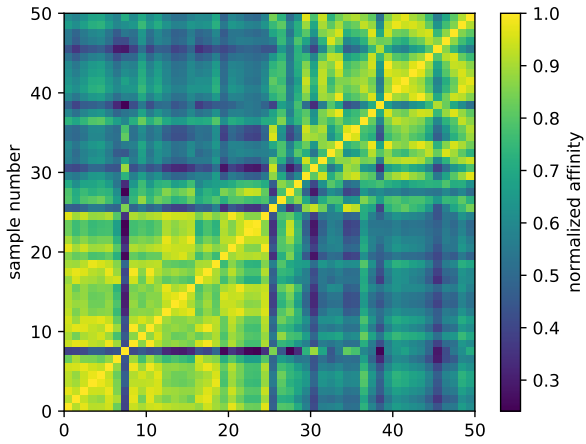


Fig. 2.10. Example of an affinity matrix for 50 samples

during training. In the case of the unlabeled sample (whether during the training phase or during the prediction phase), the supervised loss is not calculated, the label is ignored, and thus it can be set to random values or to zero.

One might wish to train the GRNN using as few as possible labeled samples. It was found that the network is trained more effectively when the labeled samples are introduced more times (more often) than the unlabeled samples. It might be called "dataset balancing" . Labeled samples (those with $m = 1$) are repeated N_R times ($N_R \in [1, \ldots, 199]$) and appended to the training data subset.

2.3.4. Graph-Regularized Neural Network

Proposed here is a neural network that is trained considering not only the labeled samples, but also neighboring labeled and unlabeled samples.

Any neural network can be converted to graph-regularized neural network (GRNN) by introducing additional inputs for neighboring features as well as modifying the loss function to accommodate the graph loss.

A general architecture (one of possibilities) of a GRNN model is provided in Fig. 2.11. In this figure, dotted lines encompasses the input vectors. Dashed lines inside the GRNN block denote prediction (a forward pass). The loss function is given by $L = m(\hat{y}_0 - y) + \sum_{i \in k_g} a_i(\hat{y}_0 - \hat{y}_1)$. The loss function is discussed further in more detail. x_0 is the main input feature, $x_{1..4}$ are neighbor input features, $a_{1..4}$ are corresponding neighbor input feature weights, y_0 is the target feature, m is the labeled/unlabeled flag, $\hat{y}_0$ is the label prediction for main input feature, $\hat{y}_{1..4}$ are the label predictions for the neighbor input features.

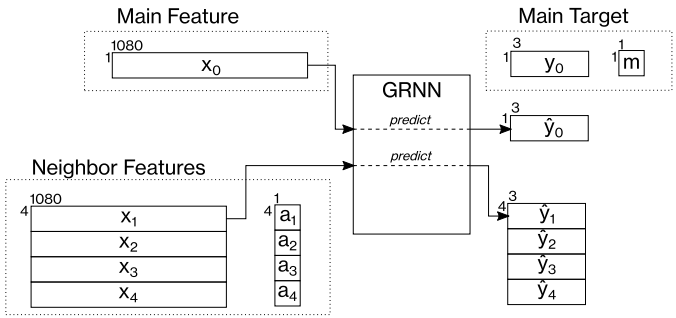


Fig. 2.11. General architecture of a graph regularized neural network (considering 4 neighbor features)

Apart from the introduction of additional inputs (neighbor features, weights, and flags), the actual neural network is just a multilayer perceptron. During the prediction phase, only the main input contributes to the prediction.

In this experiment, several multilayer perceptron architectures were used. The summary of the architectures are presented in Table 2.1. This architecture was found during a previously performed hyperparameter optimization.

Table 2.1. Summary of neural network architectures used for the experimentation

	Architecture											
	1		2		3		4		5		6	
Layer	Act. f.	Size	Act. f.	Size	Act. f.	Size	Act. f.	Size	Act. f.	Size	Act. f.	Size
input	linear	720	linear	720	linear	720	linear	720	linear	720	linear	720
hidden 1	linear	14	linear	4	linear	10	linear	10	linear	10	Leaky ReLU	10
hidden 2	sigmoid	2	sigmoid	32	ReLU	31	ReLU	15	ReLU	15	Leaky ReLU	15
hidden 3	tanh	24	tanh	23			ReLU	15	ReLU	15	Leaky ReLU	15
hidden 4	sigmoid	33	sigmoid	54					ReLU	15	Leaky ReLU	15
hidden 5	linear	50	linear	37					ReLU	15	Leaky ReLU	15
output	linear	2	linear	2	linear	2	linear	2	linear	2	linear	2

Nearby source positions produce similar acoustic features. Therefore, the predicted source positions for the nearby acoustic features should also be similar If they are similar, the graph loss is small. If they are not similar, one would need to penalize the predictor with a large graph loss.

The loss function used for the GRNN training is comprised of two parts: the supervised loss (the difference between the ground truth label and the predicted

label) and the graph loss (the difference between the main input feature label prediction and the weighted sum of neighbor input features label predictions). It can be expressed as

$$L = \mu m \sum_{i \in N_b} (\hat{y}_i - y_i)^2 + (1 - \mu m) \sum_{i \in N_b} \sum_{j \in k_g} a_{ij} (\hat{y}_i - \hat{y}_j)^2, \tag{2.16}$$

where N_b – number of samples in one training batch, k_g – size of the neighborhood, a_{ij} is the neighbor weight, equal to the corresponding element in the affinity matrix, y_i is the target output, $\hat{y}_i$ is the predicted output for main input feature, $\hat{y}_j$ is the predicted output for neighbor input feature, m is the labeled feature indicator, and μ is the supervised-to-unsupervised loss ratio.

2.3.5. Analysis of the Baseline Algorithms

To evaluate the proposed algorithm against existing approaches, two algorithms were selected as a baseline: a geometric algorithm based on finding the intersection point of two DoA radii, and an intensity map algorithm, based on finding a peak value in a two-dimensional signal intensity map.

The geometric sound source localization is based on finding the intersection of the two DoA radii for each frame. The angle of each of the radii to the positive x axis of the coordinate system is the index of the maximum of the 360 element SRP-PHAT spatial spectrum vector calculated for a particular array.

Knowing the angles of the DoA ϕ_0 and ϕ_1 of the sound source at the respective microphone arrays, the coordinates m_0 and m_1 are found using following:

$$m_i = \tan \phi_i; \tag{2.17}$$

$$b_i = M_{i,y} - m_i M_{i,y}; \tag{2.18}$$

$$s_x = \frac{b_{i,x} - b_{i,y}}{m_{i,x} - m_{i,y}}; \tag{2.19}$$

$$s_y = m_{i,0} s_x + b_{i,x}, \tag{2.20}$$

where $i \in [0, 1]$, m_i is the inclination of the i-th DoA radii, $M_{i,(x,y)}$ are the coordinates of the center of the i-th microphone array.

In the intensity map approach, the SRP-PHAT spectra of each of the microphone arrays are mapped from polar to Cartesian coordinate system and superimposed after shifting them accordingly with respect to the origin of the coordinate system. The peak value of the resulting two-dimensional intensity map is then found in two ways:

1. By finding the indices of the maximum value of the intensity map.

2. By finding the peak of the intensity map using local peak finding algorithm.

Let us consider the following example. A frame is taken of the real-world speech signal, which is known to satisfy the thresholding condition described in Section 2.1. SRP-PHAT spectra of both arrays are presented in Fig. 2.12. Consider that these spectra are presented in the polar coordinate system. It can be seen that the DoAs for the first and second arrays are 188° and 204° respectively.

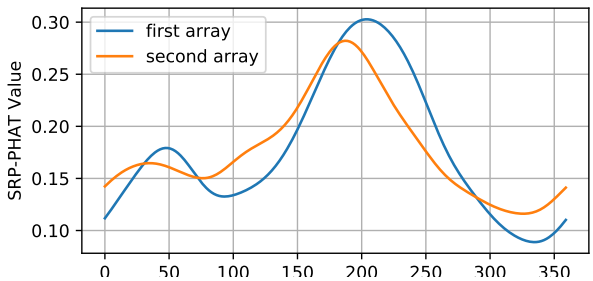


Fig. 2.12. SRP-PHAT spectra of two arrays of a frame, real-world speech signal

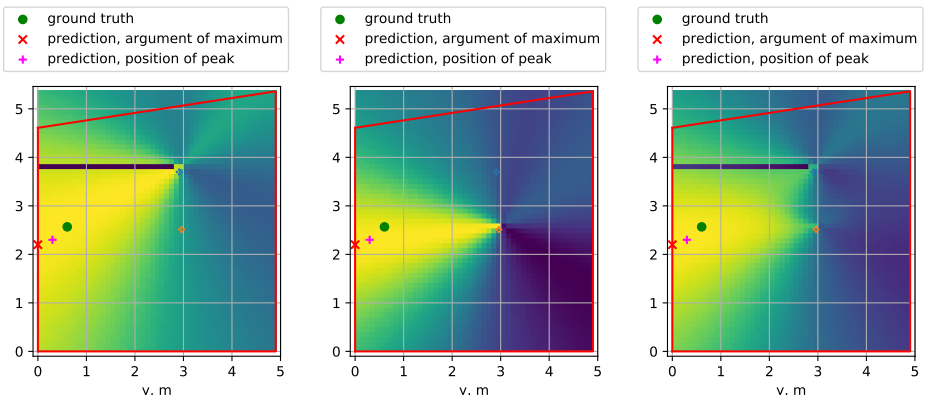


Fig. 2.13. Two-dimensional SRP-PHAT power maps for: a) 1st microphone array, b) 2nd microphone array; c) the combined SRP-PHAT power map

Both of these spectra are mapped from the polar to the Cartesian coordinate system and obtain a two-dimensional SRP-PHAT power maps for both of the arrays (Fig. 2.13a and Fig. 2.13b). two-dimensional SRP-PHAT intensity map as a result of superimposed SRP-PHATspatial spectra of the microphone arrays transformed

from polar to Cartesian coordinate system. After superimposing both maps, the position of the peak value of the resulting map is considered the position of the sound source (Fig. 2.13c). In the figure, green dot represents the real source position, the red cross represent the predicted source position as the argument of the maximum of the map; the magenta cross represent the predicted source positions as the position of the most prominent peak of the map.

Experimental evaluation of the performance of the proposed method and the results of the evaluation are presented in Section 3.2.

2.4. Multiple Sound Source Localization using Correlation Features

The method of single sound source localization using a MLP, presented in the previous section, raises a natural question whether it is possible to localize more than a single sound source within a single array audio frame.

As presented in Section 1.4, a prerequisite for signal demixing or separation is W-disjoint orthogonality. Since one wishes to localize two separate sound sources, it is theorized that the sources must be W-disjoint orthogonal. Since the source might be active simultaneously, which means, both present within a single time sample, the other condition W-disjoint orthogonality must be met – the sources must be active in separate frequency bands. To determine the frequency bands occupied by the source's signal, one must obtain an acoustic feature that carries frequency information. This is the motivation behind the CCFB feature: instead of calculating the cross-correlation for the entire signal, the signal is first split into frequency bands, and then the cross-correlation is calculated in each of the frequency bands. Thus, it is possible to determine the time lag of the cross-correlation peak in each separate frequency band and in turn determine the TDoA of a sound source active in that particular frequency band.

In this section, a method for multiple sound source localization employing a convolutional neural network and Cross-Correlation in Frequency Bands (CCFB) feature is presented.

The investigation described in this section is based on the previous research by the author, presented in (Sakavičius, Serackis 2019), which in turn is based on the investigation by He *et al.* (2018a). In this investigation, multiple sound source DoA estimation (azimuth and elevation) is performed by using a 2D DoA heatmap, and without the need to know the number of active sound sources prior the measurement. This research differs from the one presented by (He *et al.* 2018a) because in such way that a 2D DoA heatmap is used for azimuth and elevation representation instead of a 1D vector – azimuth only representation.

Main goals of this research were:

1. To present a method for multiple SSL using CNN with CCFB per unique microphone pair as input features and DoA map as output feature.
2. To present a method for semi-synthetic dataset generation for CNN training.
3. To present a method and a set of metrics for the evaluation of the performance of a CNN for multiple SSL in reverberant environment.
4. To evaluate the performance of the CNN, trained on a semi-synthetically generated data on a real-world data.

2.4.1. Justification of the Tetrahedral Array Geometry

It can be shown that utilizing a co-planar array, it is impossible to uniquely estimate the azimuth and elevation of the source, since there are two valid candidate positions for every source elevation, that is not co-planar with the array Weng, Guentchev (2001). To overcome this, a non-co-planar microphone array is proposed to be used in this investigation. The simplest non-co-planar geometry is a tetrahedron. Vertex coordinates $\mathbf{M} = [A, B, C, D]$ of a tetrahedron that is centered at $\mathbf{m}_c$ and has a side length of m_{side} are calculated as follows:

$$A = \left[\mathbf{m}_{\text{c}}(x) - \frac{m_{\text{side}}}{2}, \mathbf{m}_{\text{c}}(y) - \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}, \mathbf{m}_{\text{c}}(z) - \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}\right]; \tag{2.21}$$

$$B = \left[\mathbf{m}_{\text{c}}(x), \mathbf{m}_{\text{c}}(y) + \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}, \mathbf{m}_{\text{c}}(z) - \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}\right]; \tag{2.22}$$

$$C = \left[\mathbf{m}_{\text{c}}(x) + \frac{m_{\text{side}}}{2}, \mathbf{m}_{\text{c}}(y) - \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}, \mathbf{m}_{\text{c}}(z) - \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}\right]; \tag{2.23}$$

$$D = \left[\mathbf{m}_{\text{c}}(x), \mathbf{m}_{\text{c}}(y), \mathbf{m}_{\text{c}}(z) + \frac{\sin(\pi/3) \cdot m_{\text{side}}}{2}\right]. \tag{2.24}$$

2.4.2. Preparation of the Neural Network Training Data

For sufficient training of a CNN, a vast amount of data samples is needed. While there are some datasets for such task available (He *et al.* 2018a; Löllmann *et al.* 2018), it was found them to be rather limited for the task investigated in this section, regarding the number of simultaneously active sound sources and sound source positions, as well as the microphone array and room geometry and acoustic properties. There are some methods proposed for the simulation of acoustic data for SSL experiments. In (Vera-Diaz *et al.* 2018), the simulation method only accounts

Table 2.2. Position of the array microphones used for training data generation

Microphone index, i	m_{ix}	m_{iy}	m_{iz}
1	3.06	4.77	2.28
2	3.30	4.63	2.28
3	3.54	4.77	2.28
4	3.30	4.70	2.14

for the amplitude and TDoA of the sound source, and also is described for only one sound source. In (He *et al.* 2019), a method involving a room impulse response (RIR) generator is proposed. A RIR is generated for a particular room geometry and acoustical properties, the position of the sound source and the microphone. Here a very similar method is presented, but instead of using RIR generator, described in (Habets 2006), a `pyroomacoustics` *Python* library (Scheibler *et al.* 2018) is used. Both tools are based on the image source model.

Training data is generated in 3 steps:

1. Semi-synthetic room audio data generation (auralization of the dry signal in a virtual room);
2. Calculation of the CCFB (the input features);
3. Calculation of the DoA map (the target data).

To account for room acoustics, which may degrade the received audio signals, but which may also provide additional acoustic cues for SSL, a dry speech signal is auralized using RIRs obtained using an image source model Scheibler *et al.* (2018). In the experimentation, the dry speech signal was the close-miked meeting signal obtained from AMI Corpus (Carletta *et al.* 2006). An excerpt of the dry speech signal is taken from 30 s to 100 s of the file `AMI_Corpus/ES2016a.Mix-Headset.wav`.

Auralization is performed in a virtual room of size $\mathbf{L} = [L_x, L_y, L_z] = [10.05, 12.1, 4.1]$ m using a virtual tetrahedral microphone array $\mathbf{M} = [\mathbf{m_1}, \mathbf{m_2}, \mathbf{m_3}, \mathbf{m_4}]$; $\mathbf{m}_i = [m_{ix}, m_{iy}, m_{iz}]$, center of which M_C is positioned at $\mathbf{M}_C = [M_{Cx}, M_{Cy}, M_{Cz}] = [3.3; 4.7; 2.21]$ with the side length of 0.487 m. The exact positions of the array microphones used for training data generation are listed in Table 2.2.

Microphone array signals with one and two intermittently active sound sources are simulated in a virtual environment.

The position of each p-th sound source $\mathbf{s}_p = [s_{px}, s_{py}, s_{pz}]$ is uniformly randomly selected so that $\mathbf{s}_p \in \mathbf{L}$ for each of the simulation cases. In the experiments, auralization RIRs are simulated with 5th order reflections, with wall absorption coefficient set to 0.5. We have selected sampling rate $F_s = 44\,100$ Hz for all of the simulations. In this experimentation the value of the speed of sound $v_s = 340\,\mathrm{m\,s^{-1}}$ is used. A set of four (one for each microphone) audio signals

$\mathbf{A} = [A_1, A_2, A_3, A_4]$ is created during the auralization process. The audio signals are divided into N time frames with length $F_N = 2048$ samples, with overlap $F_O = 1024$ samples (50 %) The n-th audio frame is $\mathbf{A}_n, n \in N$.

Cross-correlation in frequency bands (CCFB) is calculated for each audio frame, for each unique pair of microphones. In our case of 4 microphones, 6 CCFB channels are calculated.

Audio frames $\mathbf{A}_n$ are filtered using a filterbank of $K = 16$ bandpass filters of 5$^{\text{th}}$ order to obtain K bandpass-filtered signals $\mathbf{Af}_{n_k} = F_k(\mathbf{A}_n)$, where $F_k(\cdot)$ denotes signal filtering using k-th bandpass filter in the time domain. Center frequencies f_{FBk} of the filterbank filters are linearly spaced on mel scale in the frequency range $\Delta f = [0, f_{k_{\max}}]$ so that $f_{k_{\max}} = (2595 \cdot \log_{10}(1 + (F_s/2)/700))$ and the difference between the center frequencies of the bandpass filters of adjacent frequency bands is $\delta f = f_{k_{\max}}/K$. Center frequency of the k-th filter of the filterbank f_{FBk} is calculated using the following equation:

$$f_{\text{FB}_\text{k}} = (700 \times (10^{(k/2595)} - 1)), \tag{2.25}$$

where $k \in [0, \delta f \ldots f_{k_{\max}}]$ is the mel frequency.

The cross-correlation $C_{i,j,k,n}(\tau) = (Af_{i,n,k} \star Af_{j,n,k})(\tau)$ between each unique pair i, j of filtered audio frame signals $\mathbf{Af}(\tau)_{n_k}$ is calculated for each corresponding frequency band k for each frame n. The output feature is constructed for each frame. $(\star)(\tau)$ denotes cross-correlation function; τ is the time lag. The maximum positive and negative τ of the CCFB sample is limited to $[-64; +64]$ samples, as it is impossible for the time lag to be larger than 64 samples with a given geometry of the microphone array, the sampling rate and the speed of sound. The example of one channel of CCFB is provided in Fig. 2.14a.

Target data sample (a DoA map) is obtained for each audio frame as follows. Firstly, the direction of arrival is calculated geometrically for each p-th source using the following equations:

$$x_p, y_p, z_p = \mathbf{s_p} - \mathbf{M}_C; \tag{2.26}$$

$$r_p = \sqrt{x_p^2 + y_p^2 + z_p^2}; \tag{2.27}$$

$$\theta_p = \arccos \frac{z_p}{r_p}; \tag{2.28}$$

$$\varphi_p = \arctan \frac{y_p}{x_p}. \tag{2.29}$$

The direction $\theta, \varphi = [0, 0]$ coincides with the positive x axis. Obtained θ_p and φ_p values are mapped onto the DoA map with a certain resolution R_{DoA}. The resolution of the DoA, R_{DoA} is equal to the number of divisions of the elevation

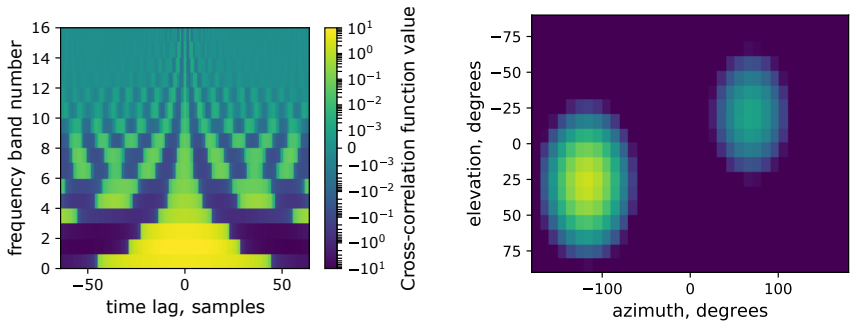


Fig. 2.14. Example of one sample of training data; a) input CCFB; b) target DoA map; two active sound sources, $\sigma = 2$, $R_{\text{DoA}} = 18$

angle range $\Delta_\phi = \pm\pi/2$ rad; and half of the number of divisions of the azimuth range $\Delta_\theta = \pm\pi$ rad. A $R_{\text{DoA}} = 18$ results in a DoA map with 36×18 elements, each point on the DoA map representing a $10^\circ \times 10^\circ$ DoA angle. This is used to limit the size of the output layer of the CNN.

The DoA map is obtained by placing onto a 2D grid a 2D Gaussian kernel (equation (2.30)) centered at the calculated DoA (θ_0 and ϕ_0 respectively) with amplitude A equal to the RMS value of the audio frame (the dry audio signal at the virtual source, before the auralization), and the spread σ.

$$K(\theta, \theta_0, \phi, \phi_0) = A \exp\left(-\left(\frac{(\theta-\theta_0)^2 + (\phi-\phi_0)^2}{2\sigma^2}\right)\right), \tag{2.30}$$

where $\theta \in [-18, 18]$, $\varphi \in [-9, 9]$ represent the DoA map grid. An example of a DoA map with two active sound sources, $R_{\text{DoA}} = 18$ and $\sigma = 2$ is presented in Fig. 2.14b.

2.4.3. Selection of the Neural Network Architecture

Two distinct CNN architectures are considered for evaluation: CONV-WE-CCFB and CONV-CCFB-DOA, which had the same input and output layers for using the same input features and target data, but the inner structure is different.

CONV-WE-CCFB architecture is based on the CNN, proposed by He *et al.* (2018a), with the output layer changed from a 360-dimensional vector to a (36×18)-dimensional array, representing the DoA map. The architecture of the CNN is presented in Fig. 2.15. In each of the convolutional (Conv2D) layers, a 5×5 kernel with 2×2 stride and a ReLU activation was used. Also, there is

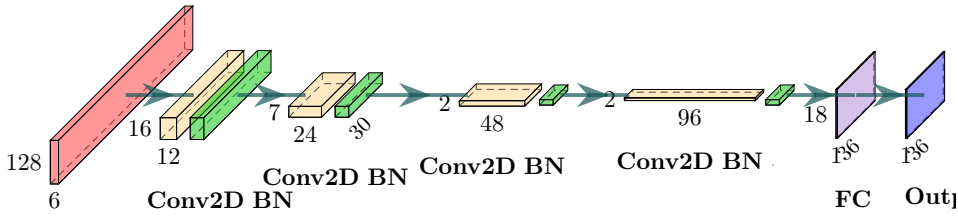


Fig. 2.15. Architecture of the CONV-WE-CCFB network

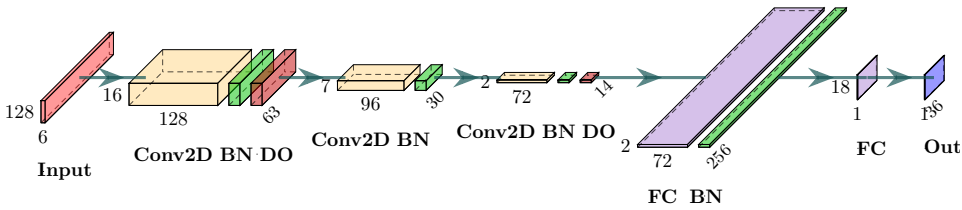


Fig. 2.16. Architecture of the CONV-CCFB-DOA network

a batch normalization (BN) layer after each of the convolutional layers. The last layer is a fully-connected output layer with sigmoid activation.

In this work, a CNN architecture is proposed (CONV-CCFB-DOA, presented in Fig. 2.16) with 3 convolutional layers, each with 128, 96 and 72 filters respectively. For the first layer, both kernel and stride are one-dimensional (kernel size (1,4), strides (1,2)), thus preventing the convolution between frequency bands. For following layers we have used 4×4 kernels with 2×2 strides. A fully-connected layer with 256 neurons follows the last convolutional layer. The output layer is of the same shape as the DoA map (36×18 in our case). For all except the last layer, the activation function was ReLU. A batch normalization layer follows each layer except the output layer. The last layer has a sigmoid activation function. We have introduced a dropout layer (DO) after the first and the third convolutional blocks.

For both networks, a stochastic gradient descent (SGD) optimizer with Nesterov momentum of 0.9 is selected.

A method for multiple sound source localization (azimuth and elevation estimation) using a CNN with CCFB features and DoA heatmap as an output was proposed. It is expected that the CNN would learn the mapping between the CCFB input features and the DoA heatmap – the desired output. Method for input and output feature acquisition and training/testing dataset generation was proposed. Two CNN architectures were presented for further evaluation, which is described in Section 3.4.

2.5. Multiple Acoustic Sources Localization using Spectrum Phase Features

Following up on the previous investigation using GCC-FB features for 2D sound source localization, it was postulated that the GCC-FB feature was not information-rich enough to allow the localization of a sound source. Moreover, assuming the fact that the acoustic features exhibit spatial smoothness, and that the complex acoustic reflection structure within the room is embedded in the acoustic features, which might also help to localize the sound source more robustly, an investigation sound source localization using CNNs trained with such features was carried out. STFT phase components of the array audio frames were used as the input features for the CNN, while the output feature was the same as described in previous section.

A method proposed here is for multiple acoustic source azimuth and elevation estimation using CNN. The neural network trained using the phase component of the STFT, estimated from the microphone array signals, as the input feature and a two-dimensional map of DoA posterior probability, referred to as DoA heatmap from now on, as the output feature. The proposed method is based on the idea of the azimuth estimation for multiple acoustic sources proposed by Chakrabarty, Habets (2019b). However, here the method is extend to estimate the elevation of the acoustic source besides the azimuth angle. This work is an extension of previous research, presented in the Section 2.4, where the same approach was utilized regarding the tetrahedral microphone array geometry and the structure of the target 2D DoA heatmap feature. However, instead of features based on a cross-correlation of frequency bands, now the phase component of the STFT, obtained from the microphone array signals, is used. Thus, the explicit feature extraction step is omitted and the method relies on the CNN to learn the feature extraction during the training.

Acoustic source positions can be estimated from the acoustic signals received by a microphone array. A CNN-based method is proposed to obtain the estimates of the azimuth and elevation of the acoustic sources in respect to the position and orientation of the microphone array. The CNN must be trained by providing training samples consisting of the input features and the corresponding outputs. After training, the CNN provides an estimate of the azimuth and elevation angle for a current set of features presented to the input.

2.5.1. Estimation of the Spectrum Phase Input Features

Extending the work of Chakrabarty, Habets (2019b), the phase component of the STFT calculated for microphone array signals is used as the input feature for the CNN. However, the W-disjoint orthogonality of the microphone array signals was

not explicitly took into account. According to the authors, in case of a N_S-source scenario, each of the source is simulated using the image-source method separately. Then the STFTs of the receiver signals are concatenated and randomly permuted in both time and frequency domains (leaving only the channel order unchanged). In our case, the STFTs are permuted in time and frequency domains, only preserving the original order of the channels, and all the N_S acoustic sources are simulated at once, so their respective spectral components are present in each time frame. The same microphone array geometry was used as was presented in Section 2.4. The preparation of input features is carried out in several steps. First, the STFTs of the simulated microphone signals are calculated. For each of the $N_M = 4$ microphone channels were set the number of Fast Fourier Transform (FFT) points equal to $N_{\text{STFT}} = 512$, with 256 point overlap and a Hanning windowing function. The number of frequency bins in the STFT was $N_f = N_{\text{STFT}}/2 + 1 = 257$. For each simulation $N_T = 4$ temporal STFT frames were obtained. As a result, an array of size $(N_S \times N_M) \times N_f \times N_T$ is created.

Next, the concatenated STFT is randomly permuted along the time and frequency dimensions, keeping the original order of elements only in the channel dimension.

Examples of the prepared input features are presented in Fig. 2.17. There are presented STFT frames for 4 microphones; training STFT sample (noise signal) on the left and testing STFT sample (speech sample) on the right.

As the input features, a single temporal frame of the resulting data structure – a matrix with $N_M \times N_f = 4 \times 257$ elements – is used. Each matrix of input features in the training dataset has an associated desired output – a two-dimensional DoA heatmap.

2.5.2. Preparation of the Two-Dimensional Desired Outputs

In the scope of the proposed method, a 2D DoA heatmap is used as a desired output for each matrix of input features. The heatmap is a matrix of $N \times M$ elements, where each element represents a certain azimuth and elevation angle range. The value of each element represents the probability of an acoustic source being active at a particular azimuth and elevation. Total range of the DoA heatmap represents a 360° azimuth range along θ axis and a 180° elevation range along ϕ axis. The number of elements of the heatmap per azimuth and elevation axes, respectively Q_θ and Q_ϕ, represent the angular resolution of the DoA heat map.

During the generation of the training dataset, to reduce the sparsity of the target feature, Gaussian blurring is additionally applied to the DoA heatmap using a 2D Gaussian kernel with separately controllable spread parameters σ_θ and σ_ϕ on the θ and ϕ axes respectively. Acoustic features exhibit spatial smoothness that is reflected in the feature space (Laufer-Goldshtein *et al.* 2016). Conversely, an ANN

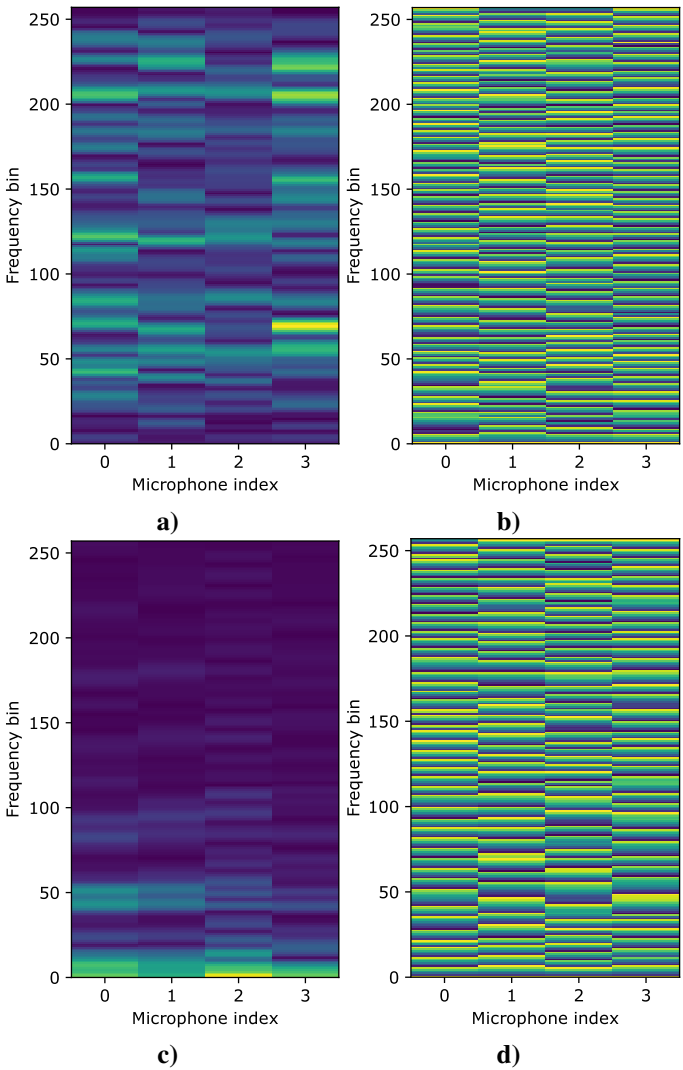
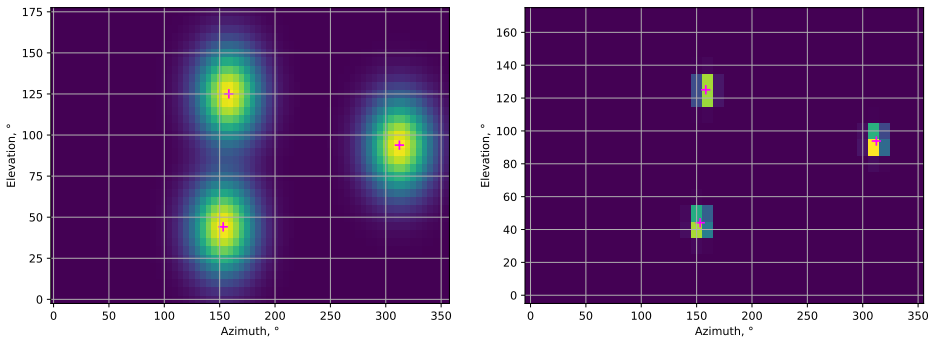


Fig. 2.17. Examples of STFT input features; a) STFT magnitude of a noise signal frame; b) STFT phase of a noise signal frame; c) STFT magnitude of a speech signal frame; d) STFT phase of a speech signal frame



a) **b)**

Fig. 2.18. Examples of 2D DoA heatmap: a) $Q = 5°$, $\sigma = 20°$; b) $Q = 10°$, $\sigma = 5°$; ground truth DoAs are marked with magenta crosses

is expected to classify such neighboring input features to neighboring classes in the output. Therefore, it is speculated that the DoA heatmap blurring operation would allow the CNN to learn to map features that are nearby in the feature space to neighboring DoA classes. The values at the output layer of the ANN represent the posterior probability of a feature being obtained for a sound source at a particular DoA. A feature for a source with a particular DoA can be viewed as having lower but non-zero posterior probability of being obtained for a source with a slightly different (neighboring) DoA. Thus it can be implied that this angular smoothing of the DoA heatmap would be beneficial for the learning of the ANN as well as its robustness.

The values at each grid element are determined by first calculating the azimuth and elevation of the simulated acoustic source with respect to the center of the microphone array. An empty DoA heatmap grid is created, on which a Gaussian kernel centered at exact azimuth and elevation is superimposed for each source DoA. The position of each of the Gaussian peaks corresponds to the 2D DoA of the source.

During the training, CNN learns to extract features from the STFT phase component and to map those extracted features to the DoA heatmap.

Examples of the prepared desired outputs at respectively $Q = 5°$ and $\sigma = 20°$ and $Q = 10°$ and $\sigma = 5°$ are presented in Fig. 2.18.

2.5.3. Post-processing of the Outputs

To obtain the DoAs of the acoustic sources from the DoA heatmap, a peak detection is performed on the heatmap and the indices of the N_S most prominent peak elements of the heatmap are converted to azimuth and elevation angles for each of

the N_S peaks. These angles correspond to the 2D DoA of the acoustic source with respect to the center of the microphone array.

A simple algorithm is used to find a local maxima. This operation dilates the original DoA heat map. After comparison of the dilated and original image, this function returns the coordinates or a mask of the peaks where the dilated heat map equals the original image.

2.5.4. Neural Network Output Layer Shape Modification

In this investigation, a similar architecture of the CNN is used, to as provided by Chakrabarty, Habets in their work (Chakrabarty, Habets 2019b), but the number of elements in each convolutional layer is altered , as well as adjusted the number of output nodes to match the number of elements in the target DoA heatmap.

Chakrabarty, Habets give an explanation that the architecture of the CNN used with N_M-channel STFT phase features can have at most $N_M - 1$ convolution layers, where N_M is the number of microphones (4 in our case), since after $N_M - 1$ layers, performing 2D convolutions is no longer possible as the feature maps become vectors. They have also experimentally demonstrated that indeed $N_M - 1$ convolution layers are required to obtain the best DOA estimation performance for a given microphone array. In the convolution layers, small filters of size 2×1 are applied to learn the phase correlations between neighboring microphones at each frequency sub-band separately. These learned features for each sub-band are then aggregated by the fully connected layers for the classification task.

A CNN with three convolutional layers was proposed as a base architecture for the research, after which a dropout layer is used, and two deep fully-connected layers, followed by a dropout layer. The output layer has the size of $N_{\text{DoA}} = Q_\theta \times Q_\phi$. The dropout rates were fixed to 0.125 and the Binary cross-entropy is used as the loss function.

As in the previous investigation, the output of the CNN is the 2D DoA heatmap, representing the posterior probability of an acoustic feature belonging to a particular spatial class. Spatial classes are arranged as a 2D grid and correspond to the DoA of a sound source. Multiple peaks might be present in a DoA heatmap, meaning that the input feature contains information about two simultaneously active sound sources. Methods for obtaining the input and target features for the training and evaluation of the CNN are presented. CNN architecture consisting of 3 convolutional layer followed by six fully-connected layers and a 2D output layer is described. The experimental evaluation of the proposed method is presented in Section 3.5.

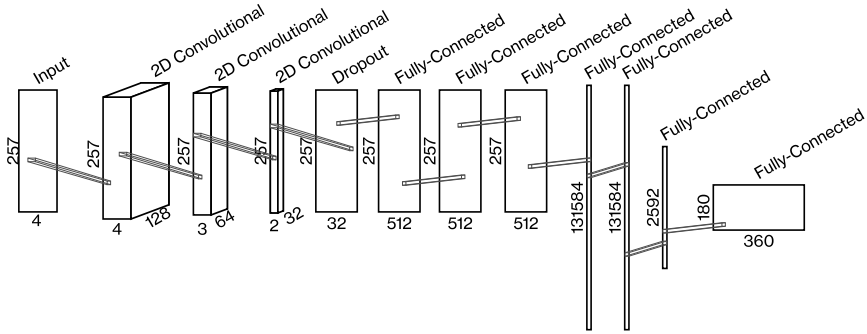


Fig. 2.19. A schematic diagram of the CNN architecture, used for the experimental investigation

2.6. Multiple Sound Source Localization in Three Dimensions

In this section, a method is presented for estimating the position of a single acoustic source or the positions of multiple acoustic sources within an enclosure using a CNN with microphone array signals' STFT phase component as an input feature.

This research is extended on the previous research of 2D DoA heatmap estimation using CNN and STFT phase input features, presented in Section 2.5.

2.6.1. Preparation of the Input Features

The preparation process of input features is analogous to the one presented in Section 2.5. In the figure, STFT magnitude and phase feature examples of noise and speech signals, 1 and 2 simultaneously active acoustic sources; 4 channel (tetrahedral) microphone array; input features are the STFT phase component. No further preprocessing is carried out for the input features. This is viewed as the foremost advantage of the proposed method, as it leaves for the CNN to abstract the input features implicitly. Moreover, STFTs are considered high-dimensional features that contain a lot of information about the propagation of the acoustic waves within an enclosure. This information is inevitably lost if the input feature dimensionality is reduced, thus prohibiting the CNN to learn to use this information for its advantage.

As the input features, we use a single temporal frame of the resulting data structure – a matrix with $N_M \times N_f = 4 \times 257$ elements. STFT features are generated for each source position (or a set of positions in case of $N_S > 1$) for each audio frame.

2.6.2. Preparation of the Three-Dimensional Desired Outputs

In the proposed method, a 3D grid of elements is used as a desired output for each matrix of input features. The grid is a matrix of $K \times L \times M$ elements, where each element represents a point in the metric space, and the value of the element represents the posterior probability of a sound source being active at that point of space. The volume covered by the 3D grid is chosen arbitrarily, and in this investigation coincided with the volume of a cuboid-shaped acoustic enclosure. The number of elements of the 3D grid along x, y and z axes, respectively, can be expressed in terms the density of elements per length unit Q_x, Q_y and Q_z, which represent the spatial resolution of the 3D grid, and the lengths of the sides of the volume X, Y and Z that is represented by the 3D grid:

$$[K, L, M] = [X, Y, Z] \circ [Q_x, Q_y, Q_z]. \tag{2.31}$$

In this investigation, the spatial resolution was equal on all axes: $Q_x = Q_y = Q_z = Q$.

A target feature for CNN training was generated in the following steps:

1. An empty 3D matrix was created. The number of elements in the matrix along each axis defines the spatial resolution of the target feature.
2. A 3D Gaussian kernel function was evaluated on the 3D matrix with the center of the kernel positioned at $\mathbf{s} = [s_x, s_y, s_z]$. The spread of of the Gaussian kernel σ determines the spatial smoothness factor of the target feature. A 3D Gaussian kernel has 3 spread values, one along each axis: σ_x, σ_y, σ_z. In this investigation, spread along all axes were the same: $\sigma_x = \sigma_y = \sigma_z = \sigma$.
3. Step 2 is repeated for N_S times.

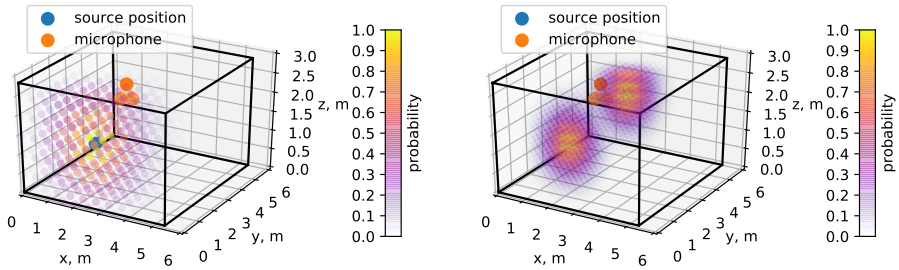
The resulting 3D grid contains a Gaussian kernel with a particular σ placed at the coordinates of the sound source.

CNN then would be trained to estimate such 3D grids for the provided STFT phase input features.

Acoustic features exhibit spatial smoothness that is reflected in the feature space (Laufer-Goldshtein *et al.* 2016). Conversely, an ANN is expected to classify such neighboring input features to neighboring classes in the output. Therefore, here is speculated that the 3D grid blurring operation would allow the CNN to learn to map features that are nearby in the feature space to neighboring spatial classes. The values at the output layer of the ANN represent the posterior probability of a feature being obtained for a sound source at a particular point in space. A feature for a source at a particular spatial position can be viewed as having a lower but non-zero posterior probability of being obtained for a source at a slightly different

(neighboring) position. Thus, we believe that this angular smoothing of the 3D grid would be beneficial for the learning of the ANN as well as its robustness against multipath propagation.

Examples of the prepared desired outputs at respectively $Q = 0.5\,\mathrm{m}$ and $\sigma = 1$ with single active source and $Q = 0.25\,\mathrm{m}$ and $\sigma = 0.5$ with two active sources are presented in Fig. 2.20.



a) **b)**

Fig. 2.20. Examples of 3D DoA heatmap: a) $Q = 0.5\,\mathrm{m}$, $\sigma = 1$, single source with coordinates $\mathbf{s} = [1.2, 2.1, 1.3]\,\mathrm{m}$; b) $Q = 0.25\,\mathrm{m}$, $\sigma = 0.5$, two sources with coordinates $\mathbf{s}_1 = [1.2, 2.1, 1.3]\,\mathrm{m}$ and $\mathbf{s}_2 = [3.1, 3.2, 2.3]\,\mathrm{m}$; ground truth source positions are marked with blue circles

It can be observed, that while the finer grid resolution might provide for better source localization accuracy, the number of the elements in the grid is also greatly increased (by a power of 3). If such grid is used as the desired output for the artificial neural network, the number of training parameters of the network is also increased, which might result in need for longer training times and greater amount of training data.

2.6.3. Modification of the Neural Network Architecture

The architecture of the CNN used in this investigation was based on the CNN architecture used in the previous research, presented in Section 2.5, which is in turn was derived from the one presented by Chakrabarty, Habets (2019b).

The proposed CNN architecture consists of three 2D convolutional layers with 128, 64, and 32 units, respectively, with the convolution kernel size of (2×1) elements. Convolutional layers are followed by a dropout layer with a fixed dropout rate of 0.125. Following the dropout layer are three fully-connected layers each containing (257×4) units, followed again by a dropout layer with a dropout rate of 0.125. Finally, there are a 1028-element fully connected layer and a $K \times L \times M$

fully connected layer which is reshaped into $K \times L \times M$ 3D array of elements at the output of the CNN, with each element of this last layer representing a spatial position in a 3D output grid. Exponential Linear Unit (ELU) as the activation function was used in every layer of the CNN. Binary cross-entropy is used as the loss function and Adaptive Moment Estimation (Adam) optimizer. The diagram of the architecture of the CNN is presented in Fig. 2.21.

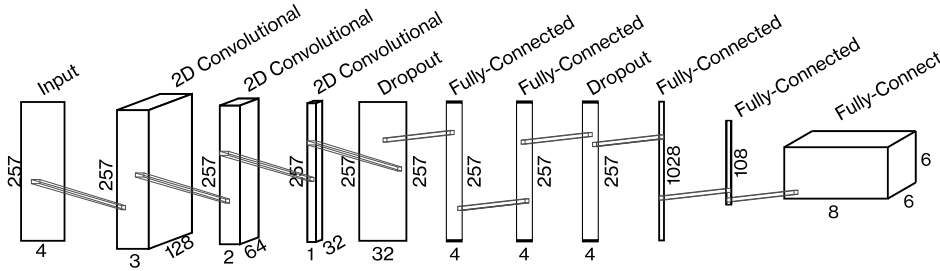


Fig. 2.21. Diagram of the architecture of the CNN

The number of neurons in the output layers of the CNN depends on the number of elements in the 3D output array, which in turn depends on the resolution Q if the spatial 3D grid and the dimensions of the acoustic enclosure. The CNN has to be trained for each different Q, σ of the 3D grid and $[X, Y, Z]$ of the enclosure.

2.6.4. Source Coordinate Estimation from a Three Dimensional Grid

For single source localization, coordinates of the element with maximal value were found and converted to Cartesian coordinates by dividing by the resolution of the 3D grid.

For multiple source localization, the 3D field was thresholded by element values, removing elements that had a value lower than the mean of the entire field. Then the remaining elements were clustered using k-means clustering to N_S clusters. N_S can be an arbitrary number, and the algorithm is supposed to find N_S most probable source positions. Center coordinates of the clusters correspond to the source coordinates. Since the thresholding value is arbitrary, several threshold values are selected, and the centers of the clusters are estimated multiple times; the estimated coordinates of each source are the arithmetic mean of each cluster center estimated at different threshold values. The experimental evaluation of the proposed method is presented in Section 3.6.

2.7. Conclusions of the Second Chapter

1. Microphone array signals' frame thresholding based on signal-to-minimum-error-amplitude ratio (SMEAR) is beneficial for the precision of TDoA estimation, which heavily influences the performance of TDoA-based source DoA and position locators, such as GCC-PHAT or SRP-PHAT.
2. It is possible to use the ILD acoustic feature for a single sound source localization when a large aperture microphone array is used and the acoustic enclosure is not reverberant.
3. The results of the investigation of the high-dimensional acoustic feature dimensionality reduction showed that it is possible to learn the acoustic feature manifold using ISOMAP NLDR algorithm:
   3.1. SRP-PHAT features obtained in a reverberant acoustic enclosure can be embedded into $\mathbb{R}^2$ space and the embeddings exhibit almost no overlap.
   3.2. The embeddings of the SRP-PHAT features exhibit the spatial structure of the sound source positions locally when the reverberation time of the acoustic enclosure is longer.
   3.3. The SRP-PHAT spectra contains information about the structure of the reflections within an enclosure and it can be viewed as an unique acoustic footprint of the sound source location.
4. It is possible to localize multiple sound sources in more than one spatial dimension using a convolutional neural network with trained on synthetic noise signals if the sound sources exhibit W-disjoint orthogonality. To ensure that the W-disjoint orthogonality is pertained within the acoustic feature, this feature must contain multiple values along either time or frequency dimension.
   4.1. Feature which contains multiple values along frequency dimension is the proposed Cross-Correlation in Frequency Bands (CCFB) feature.
   4.2. STFT of an audio frame contains multiple values along both time and frequency axis and thus can be used for sound source localization if the source signals are considered W-disjoint orthogonal.
5. Sound source localization can be viewed as a problem of acoustic feature classification to spatial classes. Thus, it is possible to classify the acoustic features to a 2-dimensional or 3-dimensional matrix of array classes that represent 2-dimensional (azimuth and elevation) or 3-dimensional (azimuth, elevation, distance or Cartesian) coordinates of the sound source.

5.1. The domain of the elements of such matrices are the posterior probability of an acoustic feature belonging to one or more classes.

5.2. Since the acoustic features are shown to exhibit spatial smoothness, such smoothness is reflected in the outputs of the classifiers, where the posterior probability of an acoustic feature to belong to a particular spatial class is also non-zero in adjacent spatial classes.

5.3. A single acoustic feature can be classified to several non-adjacent spatial classes, thus indicating that the feature was obtained for an audio frame in which more than one sound sources were active at different locations.

3

Experimental Investigation of Sound Source Localization

Investigations presented in this chapter aimed the performance estimation of the learning-based sound source localization algorithms proposed in the previous chapter. Experimental investigation results are compared with the baseline state of the art methods.

These objectives were formulated to achieve the aim of the investigation:

1. Evaluation of a performance of the method for a single sound source localization using multi-layer perceptron.
2. Experimental investigation of sound source localization using graph-regularized neural network.
3. Acquisition of a real-world tetrahedral microphone array audio datasets comparison with simulated data.
4. Experimental evaluation of CNN application with CCFB features for sound source azimuth and elevation estimation
5. Experimental evaluation of CNN application with STFT phase features for sound source azimuth and elevation estimation
6. Experimental evaluation of CNN application with STFT phase features for sound source three-dimensional position estimation.

These objectives were achieved via experimentation. The description of the experimental setups and the results are presented in the following sections of this chapter.

The research results presented in this chapter are published in five papers (Sakavičius *et al.* 2017; Sakavičius, Serackis 2019; Sakavičius 2020; Sakavičius 2021, Sakavičius, Serackis 2021) and announced at the international "AIEEE" (Riga, 2017), "eSTREAM" (Vilnius, 2017, 2019) and national "Science – Future of Lithuania" (Vilnius, 2017, 2019) scientific conferences.

3.1. Single Sound Source Localization Using Multilayer Perceptron

In this section, the evaluation of performance of a sound source localization system based on a MLP that is described in Section 2.2, is presented. The proposed system uses four microphones' array for sound signal acquisition, an ILD feature extraction stage, and an MLP for estimation (prediction) of sound source location.

3.1.1. Computer Based Simulations of Multilayer Perceptron

Computer simulation was carried in MATLAB programming environment.

In the computer simulation, MLP was trained using a rectangular array of virtual sound sources, placed around the origin of the coordinate system, and a square virtual microphone array consisting of 4 virtual microphones.

Virtual sound source positions were selected on a rectangular grid, with a step of 1 m and side length of 4 m. The plane of the grid of sound sources corresponded to the plane of the microphone array (see Fig. 3.1).

A square array of 4 virtual microphones were placed around the origin of the Cartesian coordinate system so that the center point of the array was at the origin.

We have tested the variations of the structure of the MLP with the number of neurons in each of the hidden layers being 1, 2, 5, and 10. Thus, we have tested 16 MLP configurations.

For MLP training in simulation, a grid of sound sources was produced, microphone signals were calculated, and the MLP is trained on these signals, with the actual sound source positions as training targets. MLP performance is evaluated using the sound source at a random (unseen) position (in the range of the dimensions of the grid of the positions of the sound sources used for the training of the MLP). Position of the source, predicted by the MLP, is compared to the actual sound source positions. The relative error of the prediction is calculated by

$$e_{\mathrm{rel}_i} = \frac{d(\mathbf{S}_{\mathrm{P}_i}, \mathbf{O})}{d(\mathbf{S}_i, \mathbf{O})}. \tag{3.1}$$

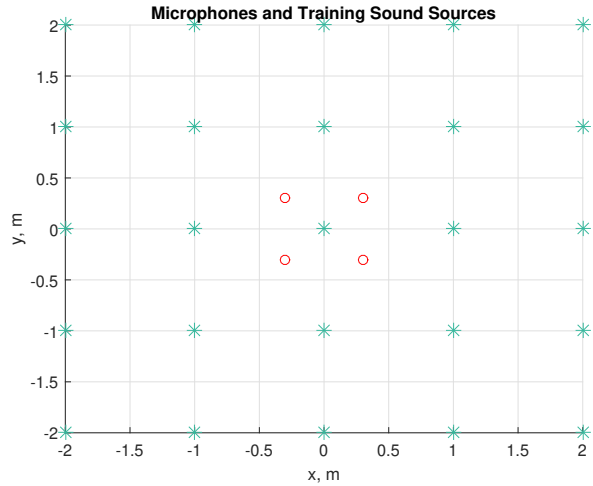


Fig. 3.1. Microphone and sound source setup for computer simulation

Each configuration of the MLP was evaluated 10 times, and the mean values of the relative, angular and distance errors were calculated afterwards, respectively, E_{rel_i}, E_{ang_i} and E_{dist_i}. These values were used to estimate the performance of the MLP for sound source localization.

3.1.2. Localization Experiments using Multilayer Perceptron

Practical experimentation was carried out in a rectangular room with partially damped acoustics with windows in one wall. The dimensions of the room was $(5.7 \times 6.26 \times 3.3)$ m.

A small loudspeaker, consisting of two mid-high-range drivers, each of 7 cm diameter, were used as a point sound source. Speakers in the loudspeaker are arranged on one side of the loudspeaker, next to each other.

A 1000 Hz sine signal was used as the test signal. 25 test points were selected on a rectangular 5 by 5 point grid in the room. Grid step was 1 meter, thus the positions of sound source ranged from -2 meters to 2 meters on both x and y axis. The center of the grid was at the center of the room and corresponded to the center of the microphone array.

Microphone array consisted of 4 Rode MP-5 cardioid condenser microphones. Microphones were mounted on microphone stands, and their capsules were placed in a rectangular array with side length of 60 cm. The microphone array was placed in the middle of the room, so that the center of the array was at $(2.85; 3.13; 1.65)$ m (see Fig. 3.2). Capsules were pointed upwards, so that the microphones were om-

nidirectional to the plane of array. Microphones' signals were recorded to PC using a Tascam US-20x20 digital audio interface.

Loudspeaker was held statically in each of the marked points (triangles in Fig. 3.2), directed to the center of the array, for about 5 seconds. For each test point, a 4 channel recording was produced.

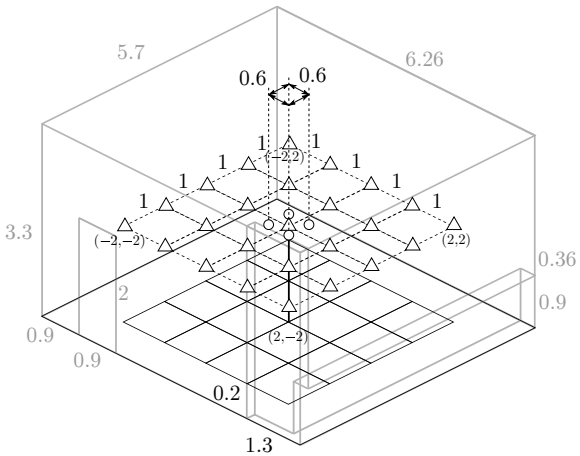


Fig. 3.2. Layout of the room for practical experimentation (microphones' locations marked with circles, sound source locations marked with triangles); dimensions in meters

For MLP training in practical experimentation, 70% of randomly selected real microphone signals were presented to the structure presented in Fig. 2.9 for MLP training, with the actual source positions as training targets. Performance of the MLP was evaluated using previously unused recordings. Same parameters for MLP training were used both in simulation and in practical experimentation, as well as the method for calculating the prediction errors (relative, angular and distance).

The results of the computer based simulation and the practical experimentation of sound source localization using concept of the proposed system are presented in Table 3.1. Best results are highlighted.

In computer simulation, the mean relative prediction error for all configurations was $\bar{E}_{\text{rel}}$ = 5.5%, mean angular prediction error for all configurations was $\bar{E}_{\text{ang}}$ = 5.23 degrees, mean distance prediction error for all configurations was $\bar{E}_{\text{dist}}$ = 1.85 m. In practical experimentation, mean relative prediction error for all configurations was $\bar{E}_{\text{rel}}$ = 8.52%, mean angular prediction error for all configurations was $\bar{E}_{\text{ang}}$ = 3.66 degrees, mean distance prediction error for all configurations was $\bar{E}_{\text{dist}}$ = 0.96 m.

Table 3.1. Results of the computer based simulation and the practical experimentation

h_1 , h_2	Computer Simulation			Practical Experimentation		
	$E_{\text{rel}}^{h_{1,2}}$ %	$E_{\text{ang}}^{h_{1,2}}$ degrees	$E_{\text{dist}}^{h_{1,2}}$ m	$E_{\text{rel}}^{h_{1,2}}$ %	$E_{\text{ang}}^{h_{1,2}}$ degrees	$E_{\text{dist}}^{h_{1,2}}$ m
1, 1	5.244	2.710	2.209	5.219	1.871	1.874
1, 2	5.095	11.522	2.000	7.262	5.441	1.537
1, 5	5.908	8.216	1.756	22.686	3.877	0.770
1, 10	5.312	2.416	2.027	5.556	2.936	1.432
2, 1	5.963	7.105	1.736	9.760	1.050	0.556
2, 2	4.008	8.524	1.958	**1.884**	5.548	1.355
2, 5	6.069	7.292	1.799	7.896	**0.097**	**0.410**
2, 10	**3.993**	4.421	1.753	10.949	1.614	1.411
5, 1	4.592	14.644	2.089	6.887	6.378	1.453
5, 2	4.566	**0.120**	1.759	6.499	4.911	0.729
5, 5	6.190	2.109	**1.577**	4.417	3.329	0.389
5, 10	4.074	5.106	1.600	6.068	5.102	0.733
10, 1	8.000	1.483	1.752	11.319	3.014	0.639
10, 2	6.003	3.238	1.791	4.770	0.984	0.800
10, 5	6.353	4.149	1.913	10.261	7.815	0.446
10, 10	6.625	0.690	1.836	14.904	4.650	0.839

As can be seen from the Table 3.1, in computer simulation, the least angular prediction error (e_{ang_i} = 0.12°) was achieved with the MLP configuration h_1 = 5, h_2 = 2, while least distance prediction error was achieved with the MLP configuration $h_1 = 5$, $h_2 = 5$. Least relative prediction error was 4 % for MLP configuration with h_1 = 2, h_2 = 10. In practical experimentation, least angular prediction error (e_{ang_i} = 0.097°) was achieved with the MLP configuration $h_1 = 5$, $h_2 = 2$, while least distance prediction error was achieved with the MLP configuration $h_1 = 5$, $h_2 = 5$. Least relative prediction error was 1.9 % for MLP configuration with $h_1 = 2$, $h_2 = 2$.

It was found that in the simulated (no reverberation) environment, the sound source localization mean position error was received as good as $1.58\,\text{m}$ for MLP configuration with h_1 = 5, h_2 = 5. In addition, the practical experimentation showed even better results, with for MLP configuration with localization a mean position error as low as $0.41\,\text{m}$ with h_1 = 2, h_2 = 5. Moreover, it was found that MLP based sound source localization system could be trained more efficiently using real-world array signals.

3.2. Experimental Investigation of Graph Regularized Neural Network

In this section an investigation of the performance of a GRNN and suitability of its application for SSL. The performance of the presented GRNN-based sound source localization method was evaluated using both synthesized array audio signals and a real-world array signal dataset.

Simulated array audio dataset was acquired in a simulated enclosure with dimensions $l = 10\,\mathrm{m}$, $w = 12\,\mathrm{m}$, $h = 3\,\mathrm{m}$. The acoustic scene was modeled for simulation using image-source RIR simulation method. The origin of the Cartesian coordinate system corresponds with the front left bottom corner of the simulated enclosure.

A single source position was selected uniformly randomly from the entire $x - y$ plane of the enclosure and the z coordinate was fixed to the same height as the microphones. Band-limited uniform noise was selected as the signal of the simulated sound source. The signal was obtained by filtering a passage with the duration T of uniform white noise using a 5th order Butterworth band-pass filter between the frequencies f_1 and f_2. Initially, values of $f_1 = 500\,\mathrm{Hz}$ and $f_2 = 1000\,\mathrm{Hz}$, $T = 1\,\mathrm{s}$ were used.

Inside the enclosure, N_M circular microphone arrays, each with N_m microphone elements and radius r_M were modeled. Planes of the microphone arrays were parallel to the ground (the normals of the circles coincided with the z axis of the enclosure model). The arrays were modeled at a fixed height $h_m = 2\,\mathrm{m}$. Each of the circular microphone arrays had a radius of $r_M = 0.116\,\mathrm{m}$ and 9 elements ($N_m = 9$).

3.2.1. Position Estimation using Steerd Response Power Features

The investigation of the ability of GRNN to learn the $\mathbb{R}^2$ manifold of the SRP-PHAT dataset embedded in high-dimensional feature space is presented following.

A dataset of SRP-PHAT features and their corresponding ground truth source position coordinate sets were generated using Pyroomacoustics Python package. ISOMAP embeddings were then obtained via the method described in Section 2.3. Hyperparameter optimization of the entire solution was performed and the optimal (by evaluating the source position prediction Cartesian MSE) parameter set was found. Minimum and maximum values of each parameter search space, as well as best performing parameter values are presented in Table 3.2.

To gain a better insight on what is the magnitude of the influence of each of the hyperparameters on the performance of the proposed SSL method, a hyperparameter correlation matrix is presented in Fig. 3.3.

The evaluation of the performance of the proposed method starts with using ISOMAP to obtain embeddings of the high-dimensional SRP-PHAT features into low dimensional (2-dimensional in this particular case) space. As can be seen from Fig. 3.4, the ISOMAP embeddings form a circle with input features ordered by their DoA with respect to the microphone array. In the figure, the red, green and

Table 3.2. Parameters used in parameter optimization/search, one microphone array

Variable	Parameter in code	Description	Min.	Max.	Best
$\mathrm{N}_{\mathrm{k_{emb.}}}$	`isomap_nbrs`	number of ISOMAP nearest neighbors	2	50	44
f_1, Hz	`f1`	band-pass filter bottom cutoff frequency	253	999	659
f_2, Hz	`f2`	band-pass filter top cutoff frequency	1028	7922	2474
f_s, Hz	`fs`	sampling rate	8110	43923	25792
a	`absorption`	absorption coefficient	0.0031	0.9984	0.7871
lr	`learning_rate`	learning rate	0.001	0.0198	0.0172
T, s	`T`	frame duration	0.1033	3.9886	3.3441
O_r	`order`	acoustic simulation order	3	10	5
N_m	`mic_arr_mic_n`	number of microphones in array	2	10	9
N_g	`gnbrs`	number of graph nearest neighbors	1	50	1
R_{rep}	`k_rep`	labeled samples repetition rate	1	200	147
N_E	`n_epochs`	number of training epochs	5	60	53
N_B	`batch_size`	training batch size	1	64	53
μ	`mu`	supervised-to-unsupervised loss ratio	0.01	0.999	0.392
r_M	`mic_arr_radius`	microphone array radius	0.0401	0.4982	0.1162

blue color components of each point correspond to x, y and z coordinates of the sound source position for which the corresponding acoustic feature was obtained

Secondly, ISOMAP embeddings were mapped to the metric coordinate system using a GRNN. By evaluating the proposed method using synthesised audio data, it was found that with only a single microphone array being used, GRNN was only able to learn the DoA mapping of the SRP-PHAT feature (see Fig. 3.5). Naturally, a question arises whether using two microphone arrays instead of one would yield better results.

Multiple experiments with simulated and real-world audio data have been conducted using a setup of two circular microphone arrays, each with radius $r = 45\,\mathrm{mm}$ and $N_m = 4$ microphone elements placed at $0°$, $90°$, $180°$ and $270°$ with respect to the positive x axis of the coordinate system.

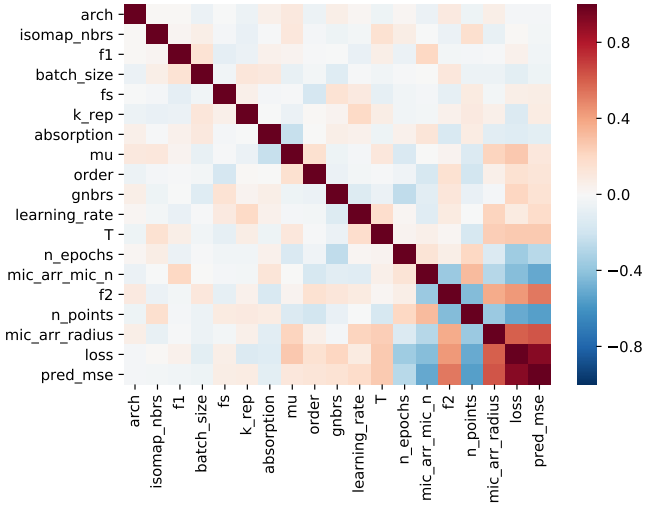


Fig. 3.3. Hyperparameter correlation matrix (1 microphone array)

A dataset for the evaluation of the proposed method with variable N, T, N_m, f_1, f_2, a, r_M and O_r parameters was created. The dataset consist of N samples, each is a pair of vectors: $(N_M \times 360)$-dimensional concatenated SRP-PHAT feature vector and a 3-dimensional coordinate feature vector (x and y coordinate randomly selected, z coordinate fixed). The process to obtain the dataset is described in Section 2.3.3.

SRP-PHAT spatial spectra was calculated using pyroomacoustics inbuilt function, from the simulated microphone audio signals, using FFT length of 512. The spectrum is a 360-element vector, that covers a 360° azimuth, thus the resolution of the spectra is 1°. For each of the N_M microphone arrays, a separate SRP-PHAT spectrum is obtained. To form a feature vector for a single source position, all SRP-PHAT spectra are concatenated.

The dimensionality of SRP-PHAT feature vectors was reduced to 2 dimensions using ISOMAP NLDR algorithm, considering $k = 16$ nearest neighbors. The mapping is presented in Fig. 3.6.

The performance of the GRNN is evaluated by calculating the Mean Squared Error (MSE) between the ground truth coordinates and the GRNN predicted coordinates:

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{ts.}}} \sum_{i \in N_{\mathrm{ts.}}} (y_i - \hat{y}_i)^2, \tag{3.2}$$

where $N_{\mathrm{ts.}}$ is the number of samples in the testing dataset.

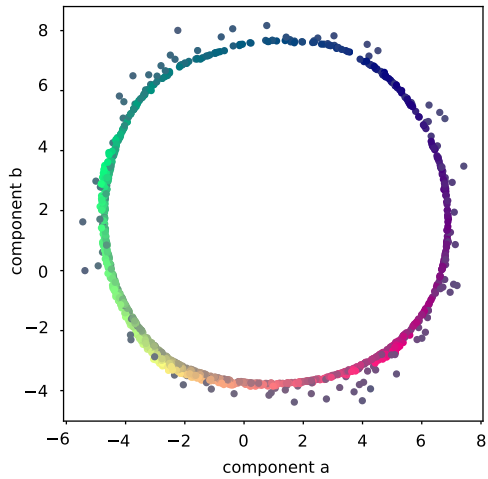


Fig. 3.4. ISOMAP SRP-PHAT feature embedding, features obtained using one circular microphone array with 9 elements; simulated data

For the evaluation of the performance, only samples unseen by the model during the training phase are used (the testing data subset). The results of the source position prediction by the best performing model are presented in Fig. 3.5.

3.2.2. Experimental Tests on Real-World Array Audio

A real-world audio dataset was acquired using 2 circular microphone arrays in an acoustically untreated room with approximate dimensions of 5 m × 5 m, with a height of 3.75 m (see Fig. 3.10 for exact geometry of the room) and a reverberation time $T_{60} = 0.311\,\mathrm{s}$. The microphone arrays had 4 electret elements each and the radius $r_{\mathbf{M}} = 45\,\mathrm{mm}$ (see Fig. 3.7). Microphone arrays were placed on tripods at the approximate height of $1.3\,\mathrm{m}$. The exact coordinates of the first microphone array were $\mathbf{M}_1 = [2.913, 3.699, 1.313]\ \mathrm{m}$ and $\mathbf{M}_1 = [2.960, 2.512, 1.309]\ \mathrm{m}$ (see Fig. 3.9). All distances were measured using a laser rangefinder with a measuring accuracy of $0.5\,\mathrm{mm}$.

Microphone array signals were digitized and transferred to a computer using Tascam US-20x20 digital audio interface. Signals were captured at a sample rate $f_s = 44\,100\,\mathrm{Hz}$ and quantization resolution $Q = 24$ bits. The schematics of a simple electronic interface that allows to connect the electret microphone elements to the digital audio interface's microphone inputs and to power the electret microphone elements using $48\,\mathrm{V}$ Phantom power is presented in Fig. 3.8.

The dataset consists of a set of unlabeled (unknown source position) microphone array audio data, a set of audio data from the labeled source positions (34

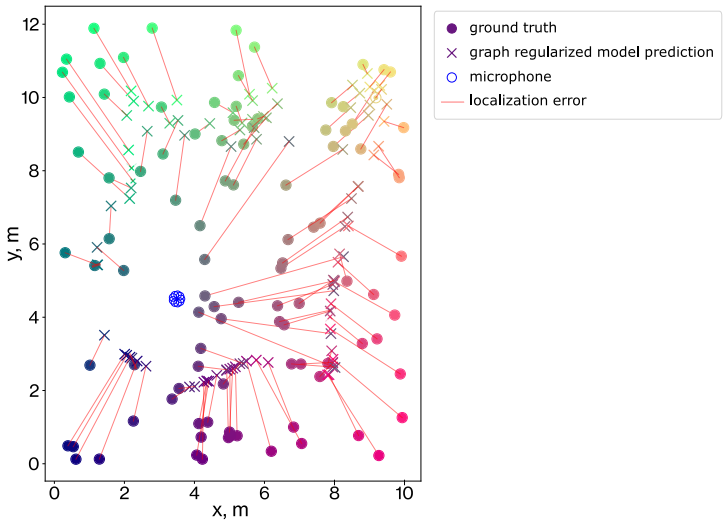


Fig. 3.5. Sound source position estimates by best performing GRNN model and ground truth source positions; single microphone array

points), and a set of audio data of known source trajectories. Each set consists of audio signals obtained using a white noise and a dry speech source signal. The boundaries of the room as well as the known sound source positions are presented in Fig. 3.10. In the figure, the red, green and blue color components of the point color correspond to their x, y and z coordinate respectively.

SRP-PHAT features were obtained using previously described steps. SRP-PHAT features' dimensionality was reduced to 2 components using ISOMAP algorithm and embeddings of the features were obtained. To compare with the simulated acoustic features embeddings presented in Fig. 3.4, the embeddings of a real-world acoustic features are presented in Fig. 3.11.

It can be seen that the distribution of the embeddings is much more complex. This can be attributed to the more complex structures of reflections that are present in a real-world acoustically untreated room versus a simulated room with nothing more than walls. It can be speculated that the reflections that are present in a real-world acoustic enclosures and that are reflected in the acoustic features obtained within such enclosures act as a acoustic fingerprint, being unique in structure at any particular location within the enclosure, as opposed to simulated enclosures, where there are, for example 8 points within a cuboid room where the reflection structure is identical (due to symmetry of the simulated enclosure). This implies that the proposed GRNN-based source localization method would perform better in a real-world acoustic scenarios than in simulated ones.

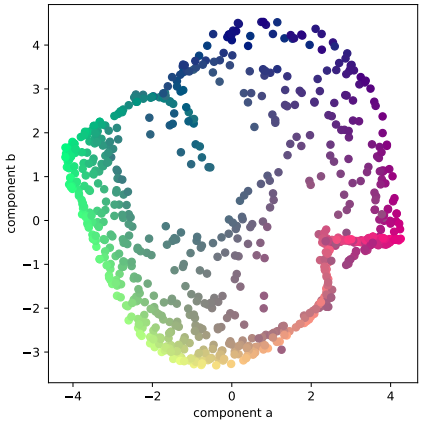


Fig. 3.6. ISOMAP embeddings of unlabeled and labeled SRP-PHAT features, synthesised noise signal

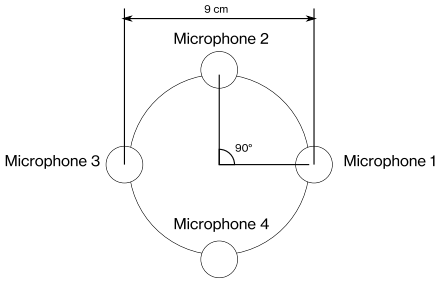


Fig. 3.7. Geometry of the one of the two microphone arrays used to capture real-wold array audio signals for the GRNN experimentation

Since the presented method is intended to localize a speaker within an acoustic enclosure, it was necessary to evaluate the performance of the proposed method using a real-world speech dataset. The audio data that was obtained from the microphone arrays was split into frames with duration of T_τ. For each frame, a 360-element SRP-PHAT spectra was obtained for each array (Fig. 3.12). In the figure, the maximum value of each SRP-PHAT spectrum is marked with a red dot; the color of each point corresponds to a and b embedded coordinates (red component: a; green component: b, blue component constant).

Since the speech signal is not continuous and contains silent passages , some of the frames of the array audio signals do not contain the signal of the source, and only the noise signal. Using these silent frames, it is not possible to localize the (inactive) sound source. The effects of audio signal frame thresholding are presented in Section 2.1. Therefore, such silent frames must be discarded.

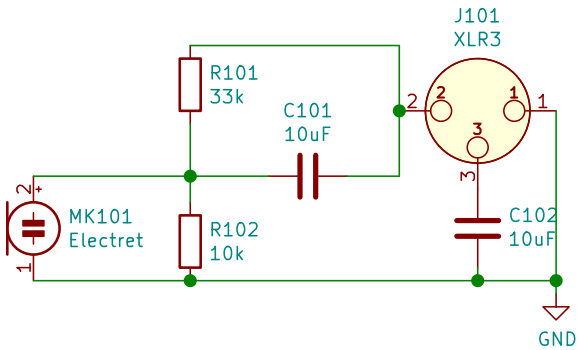


Fig. 3.8. Schematics of a simple electronic interface that allows to connect the electret microphone elements to the digital audio interface's microphone inputs

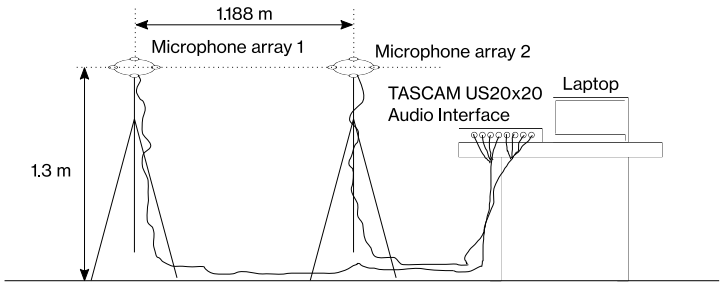


Fig. 3.9. Diagram of the physical setup of the microphone arrays used to capture real-wold array audio signals for the GRNN experimentation

In this investigation, two simple frame thresholding algorithms were used:

1. RMS-based; for each frame, the RMS value of the SRP-PHAT spectra is calculated and then compared to the mean value of all SRP-PHAT RMS values obtained from a particular microphone array.
2. Crest factor-based; same as above, but using the crest-factor instead of RMS values of each frame.

The frame is kept only if both arrays SRP-PHAT spectra meet the conditions. The mean value of either the RMS or the crest factor is multiplied by a constant to give further fine-tuning of the thresholding.

The results of sound source localization using the baseline geometric source localization approach are presented in Fig. 3.13. It can be seen, that in some cases the source position estimation errors are large and the source position is estimated outside the enclosure. This happens when the DoA estimates for both microphone arrays are very close, and thus the intersection of the DoA radii is approaching infinity.

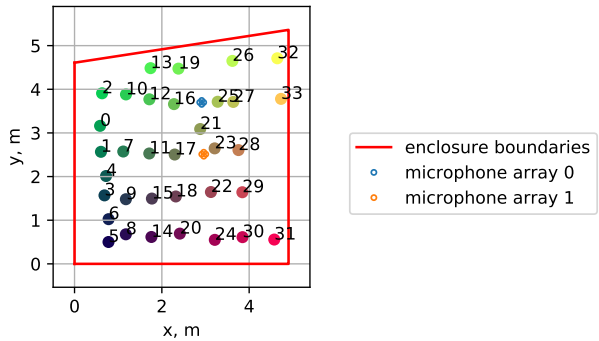


Fig. 3.10. Room boundaries (red), microphone positions (both arrays), and known sound source positions

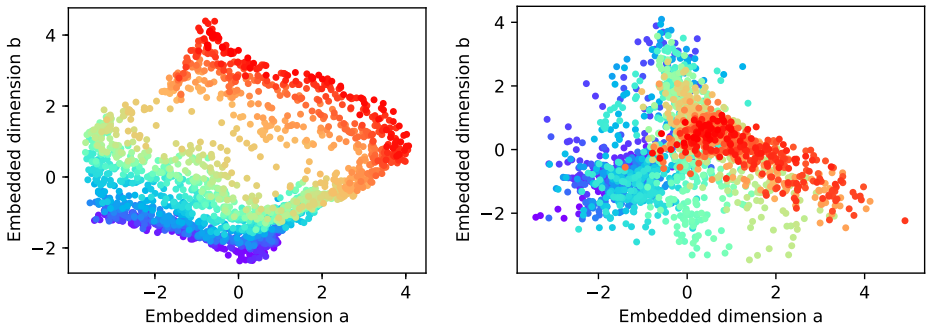


Fig. 3.11. 2-dimensional ISOMAP embeddings of the unlabeled SRP-PHAT features of the real-world signals; a) noise signal; b) speech signal

Using the GRNN sound source localization method presented in this section and in Section 2.3, the positions of the sound source were estimated for real-world speech audio data. The results of the localization are presented in Fig. 3.14. It can be observed in the figure, that the source position estimates are much more accurate than in the previous experiment.

After assembling the training dataset with 2 nearest neighbors, and training the neural network for 50 epochs, the predictions of the source position were more accurate than using the baseline methods.

The summary of source location prediction errors for different source localization methods are presented in Table 3.3. In this table, the results from all experiments were aggregated, and a global source position estimation MAE and STD values are presented. Table 3.4 presents MAE of different source localization methods when evaluated with the best-performing GRNN parameter set.

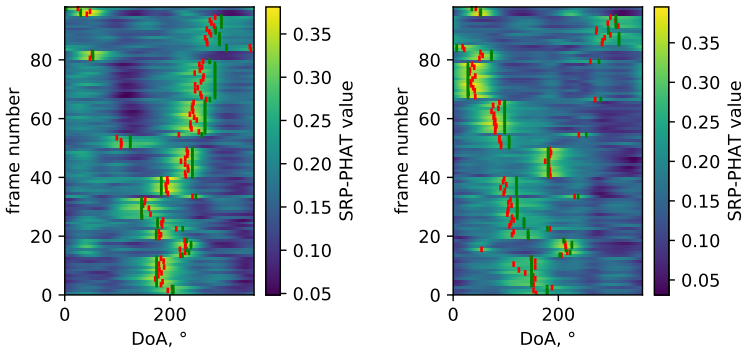


Fig. 3.12. SRP-PHAT spatial spectra of microphone array signals; a) 1st array b) 2nd array; marked actual DoA (green) and SRP-PHAT peak value (red), real-world speech source

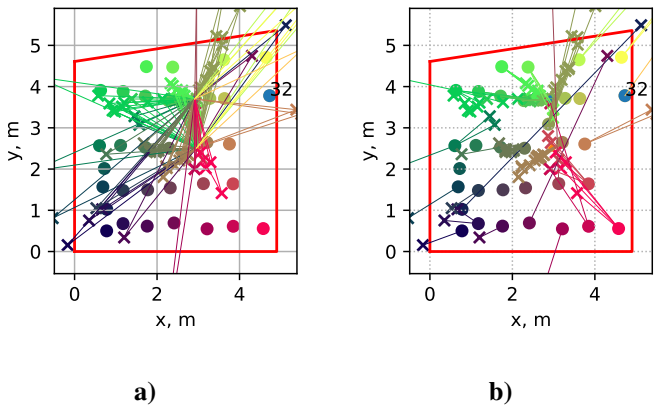


Fig. 3.13. Real-world speech source positions estimated using SRP-PHAT geometric localization algorithm; a) lines represent the DoA radii of each array; b) lines represent localization error

The results of the source location prediction using a GRNN are presented in Fig. 3.14. The distributions of the prediction errors for different source localization methods are presented in Fig. 3.15. It can be seen, that while the source position estimation MAE is similar to all methods, GRNN produces less error variance. After performing a Bayesian hyperparameter optimization for 169 iterations, source position estimation error estimates were obtained.

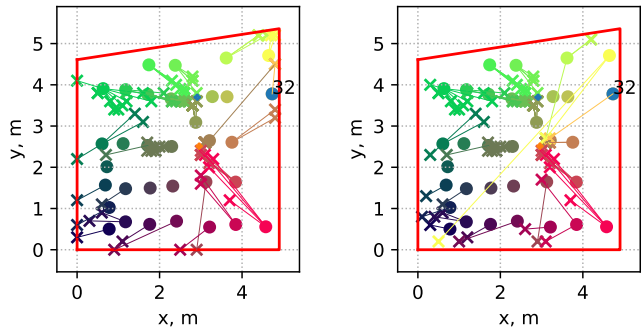


Fig. 3.14. Predicted real-world speech source positions using the SRP-PHAT intensity map approach; a) source positions is the argument of the maximum of the intensity map, b) source positions is the location of the most prominent local peak of the intensity map

Table 3.3. Summary of source location prediction errors for different localization methods (aggregated from all experiments)

Method	MAE, m	STD, m	Improvement, %
Geometric	1.95	1.62	80.1
Intensity Map, argmax	1.13	0.66	4.9
Intensity, peak location	1.12	0.69	3.5
GRNN	**1.08**	0.51	–

The parameters which influence on the source localization accuracy was investigated were: acoustic feature thresholding level, number of nearest neighbors considered for ISOMAP embedding, number of nearest neighbors considered during graph dataset creation, the reintroduction rate of labeled samples during GRNN training, the ratio of supervised to unsupervised loss used during GRNN training and training sample batch size. The relation between the estimated prediction error and the parameters are presented in Figs 3.16–3.19 with further explanation.

As can be seen from the experimental results, the presented method outperforms the baseline methods for almost all parameter configurations. The presented method produces a position estimation error that is $24.2\,\%$ lower than using a geometrical source localization method, and $19.1\,\%$ lower than using the intensity map method at low feature fitness threshold levels. When the acoustic feature selection threshold is high, the performance of all methods becomes comparable. It needs to be addressed that it is impractical to use high threshold values because it is possible that the sound source would not be localized at all (all its features are below the threshold).

Table 3.4. Summary of source location prediction errors for different localization methods (compared to best-performing GRNN parameter configuration)

Method	MAE, m	STD, m	Improvement, %
Geometric	3.06	4.50	68.57
Intensity Map, argmax	1.17	0.74	17.97
Intensity, peak location	1.14	0.75	15.94
GRNN	**0.96**	0.62	–

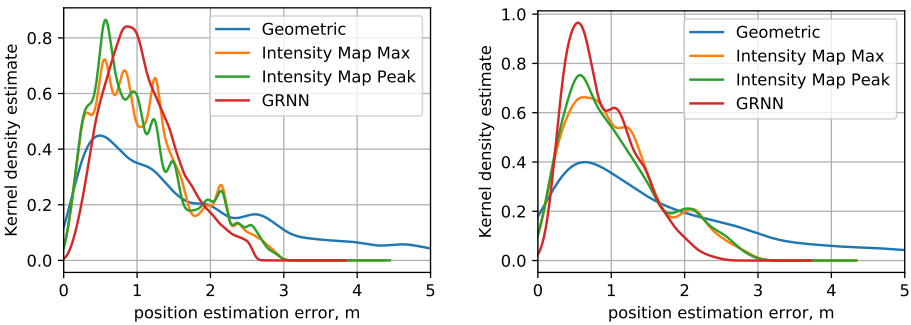


Fig. 3.15. The distributions of the prediction errors for different source localization methods; a) all experiments; b) best performing GRNN parameter set

As seen from the Fig. 3.16b, parameter $N_{k_{emb.}}$ does not affect the performance of the baseline algorithms. This is the expected case, since this parameter is not involved in obtaining the source position estimation using the baseline methods. As for the GRNN approach, it can be observed that $N_{k_{emb.}}$ has little impact on the position estimation. Nevertheless, the presented method performed on average 20.3 % better than the baseline methods when considering the source position estimation RMSE.

Considering the number of the nearest neighbors of the samples in the embedded space when constructing the training graph dataset, Fig. 3.17 shows that the smallest source position estimation error is obtained when only a small number of graph nearest neighbors are selected. This might be due to the non-linearity of the embedded space. The larger the number of considered neighbors, the further the samples are in the embedded space, and the larger the error. As can be seen from the Fig. 3.17b, the number of the nearest graph neighbors considered is influencing the source distance estimation variance the most. The source distance estimation error is the smallest when the proposed algorithm considers 10 neighbors.

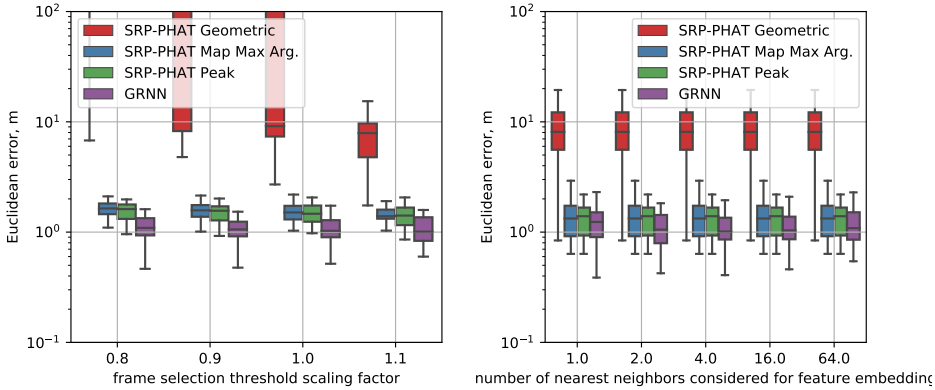


Fig. 3.16. Dependency between the source position estimation error and: a) the acoustic feature thresholding level; b) the number of nearest neighbors considered for ISOMAP embedding (y axis clipped to 1×10^2 for better visibility of GRNN estimates)

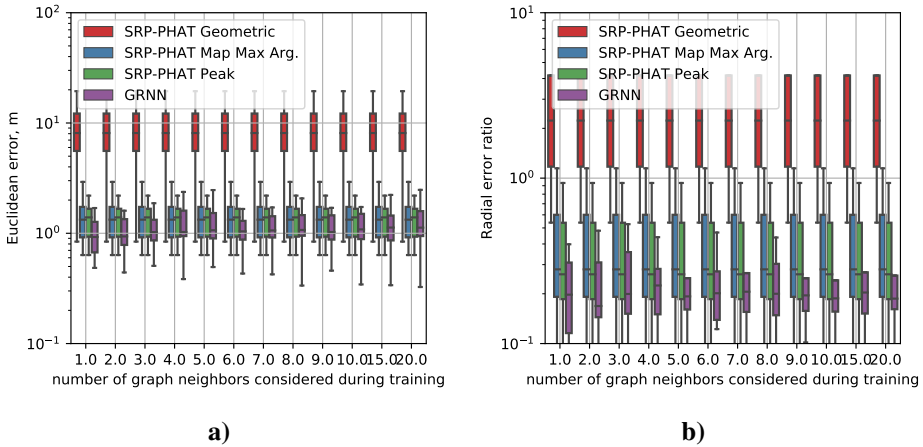


Fig. 3.17. Dependency between the number of nearest graph neighbors considered during the training dataset construction and: a) the source position estimation error; b) source distance estimation error

As can be seen from the Fig. 3.18, the labeled sample repetition rate is not influencing the source position estimation error considerably. The angular error is reduced at high repetition rates, but the source distance estimation error is increased. This might be due to the condition where the labeled sample positions

are condensed around the center of the enclosure, and the supervised loss function forces the network to predict the source positions towards the center. This produces large estimation errors for the sound sources that are further away from the center of the enclosure.

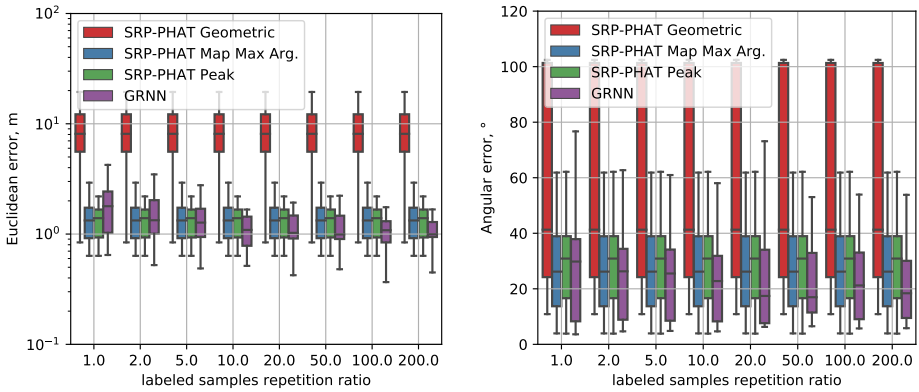


Fig. 3.18. Dependency between the labeled samples repetition rate during GRNN training and: a) the source position estimation error; b) source DoA estimation error

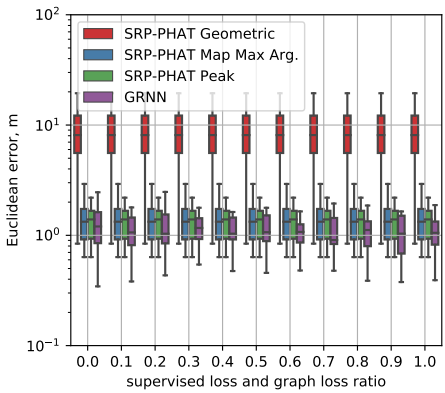


Fig. 3.19. Dependency between the source position estimation error and the ratio between the supervised and unsupervised loses considered during the GRNN training

The most suitable ratio of supervised to unsupervised loss, as can be seen from the Fig. 3.19, is $\mu = 0.6$, which means that the supervised loss has slightly stronger influence during the training of the GRNN. It can be speculated that since there are much more unlabeled samples than labeled ones, the influence of the labeled samples (which contribute during the calculation of the supervised loss) needs to be higher than the influence of the unlabeled samples to achieve a balanced training pattern.

The presented method outperforms the baseline methods for almost all hyperparameter configurations. The presented method produces a position estimation MAE that is averagely 5 times lower than using a geometrical source localization method, and averagely $3.5\,\%$ lower than using the SRP-PHAT intensity map method at low feature fitness threshold levels. The most suitable ratio of supervised to unsupervised loss is found to be $\mu = 0.6$. Overall smallest source position error is achieved with 1 nearest graph neighbor considered during graph training dataset creation.

1. Using an ISOMAP NLDR algorithm, it is possible to embed SRP-PHAT acoustic features to a $\mathbb{R}^2$ space and the embedded dimensions correspond to the spatial dimensions of the acoustic enclosure.
2. Embeddings themselves correspond to the x and y coordinates of the sound source.
3. It is possible to localize a single sound source within an acoustic enclosure with a data-driven algorithm that:
   - Is semi-supervised learning based;
   - Is trained on a an unbalanced ($\mathrm{N_l} << \mathrm{N_u}$) training dataset.

3.3. Real-World Tetrahedral Microphone Array Audio Dataset

For evaluation of the simulation results, a dataset of real-world microphone array signals with one or multiple sound sources present within the acoustic enclosure is needed. There are several audio datasets presented earlier (Le Roux *et al.* 2015), focused on the sound source localization and separation tasks. The LOCATA dataset (Löllmann *et al.* 2018), presented as a part of IEEE-AASP Challenge on Acoustic Source Localization and Tracking, consists of audio recordings of one or two moving and up to 4 static sound sources, captured with a multitude of microphone arrays, with number of microphone per array ranging from 2 to 32. The shortcoming of the LOCATA dataset is that neither the room dimensions nor the distance of the origin of the coordinate system to a corner of the room is not presented, which

imposes a limitation of usage of the LOCATA dataset for evaluation of learning-based SSL methods, such as presented by He *et al.* (2018a, 2019), where the model is trained on semi-synthetic data. Furthermore, the moving sound sources were the human subjects, walking in front of the microphone array and talking, thus there are limited variance of the height of the sound sources relative to the origin of the coordinate system. The Sound Source Localization for Robots (SSLR) Dataset is a collection of real robot audio recordings for the development and evaluation of sound source localization methods, recorded using Softbank robot Pepper, including robot ego-noise and overlapping multiple speech sources (He *et al.* 2018a). The origin of the coordinate system for this dataset is the center of the microphone array. Moreover, the sound sources remain stationary, while the robot head is panning to sides, thus the microphone-room spatial relationship is constantly changing, which is not the case in many ambient intelligence and surveillance systems, where the array is stationary for the duration of operation. Therefore, this dataset may not be well suited for evaluation of the performance of static arrays. Drone Egonoise and localization (DREGON) dataset (Strauss *et al.* 2018) is aimed at evaluating SSL using microphone arrays embedded in an unmanned aerial vehicle (UAV). The dataset contains both clean and noisy in-flight audio recordings continuously annotated with the 3D position of the target sound source using an accurate motion capture system. The dataset includes the description of the room geometry and its reverberation time. In addition, the speech signals were emitted by a static sound source. The downside of this dataset is that the microphone array is mounted on the UAV and is not stationary or vice versa, the sound source is stationary. Considered the shortcomings of the aforementioned datasets, we present a dataset for the evaluation of the performance of sound source localization algorithms that is captured by a static tetrahedral microphone array with one and two static, simultaneously active sound sources. The presented dataset includes thorough and explicit measurements of the room and the positions of the microphones and the sound sources with the origin of the coordinate system coinciding with one corner of the room.

3.3.1. Description of the Experimental Setup

For all audio recordings, a Tascam US20x20 USB audio interface was used. All recordings were performed with a sampling rate of 44 100 Hz and 16 bit quantization resolution. All spatial measurements were made manually using a measuring tape with a precision of ± 0.005 m. The dataset consists of audio files (4 channel audio files for microphone array signals and mono audio files for the corresponding source signal), impulse response measurement data in *MATLAB*® compatible format (.mat) and in .WAV file and a spreadsheet file with the corresponding infor-

mation about the positions of the sound sources, the microphone and the signals of the sound sources.

3.3.2. Properties of the Room

The dimensions of the room were $5.40 \times 5.86 \times 2.64$ m. The origin of the coordinate system of the dataset coincided with a corner of the room. Three of four of the walls of the room were made of painted masonry, while the fourth wall was a plaster wall. The volume of the room was $89.869\,\mathrm{m}^3$ and the total surface area of the room was $145.048\,\mathrm{m}^2$.

The furniture of the room consisted of three plywood tables, three chairs, several computers and computer monitors, which were not taken into account to not over-complicate the process of dataset acquisition. The absorption coefficients of each of the wall were not directly measured but rather calculated from the measurement of the T_{60} reverberation time value using Sabine's equation:

$$T_{60} = \frac{24 \ln 10}{c_{20}} \frac{V}{Sa} \approx 0.1611 \frac{V}{Sa}, \tag{3.3}$$

where c_{20} is the speed of sound at $20\,^{\circ}\mathrm{C}$, V is the volume of the cuboid room, S is the total surface area of the room, and a is the average absorption coefficient of the surfaces of the room.

The reverberation time can be calculated using Schroeder's method of backward integration of the room impulse response (RIR) (Schroeder 1965):

$$F_c \approx 2000 \left(\frac{T_{60}}{V} \right)^{0.5}. \tag{3.4}$$

For the RIR measurements, a Mackie Thump12 powered loudspeaker was used as a sound source (axis of the loudspeaker directed to the capsule of the microphone). The measurement microphone was Sonarworks XREF20. RIRs were captured using a *MATLAB*® tool Room Impulse Measurer. Provided by the tool are the two most widely used IR measurement techniques: Maximum-Length-Sequence (MLS) and Swept Sine. MLS technique is based on the excitation of the acoustical space by a periodic pseudo-random signal. The impulse response is obtained by calculating a circular cross-correlation between the measured output of the system and the excitation signal. The Swept Sine measurement technique uses an exponential time-growing frequency sweep as and the excitation signal. The output of the system is recorded and deconvolution is used to recover the impulse response from the swept sine tone. The impulse response of the room was measured at three different combinations of the signal source and the measurement microphone positions (positions are presented in Table 3.5).

Table 3.5. Position of the impulse response measurement loudspeaker and microphone and the results of the T_{60} measurements

	Source			Microphone				
No.	x, m	y, m	z, m	x, m	y, m	z, m	IR measurement signal	T_{60}, ms
1	4.16	3.74	1.03	1.395	0.73	1.52	MLS	576
							Sine Sweep	526
2	4.49	2.43	2.95	3.825	1.145	1.49	MLS	607
							Sine Sweep	533
3	4.98	1.4	1.03	4.765	3.275	0.34	MLS	520
							Sine Sweep	551

The T_{60} reverberation time was calculated using Schroeder's backwards integrated room impulse decay method and calculating the intersection of the slope with the $-60\,\mathrm{dB}$ (Hak *et al.* 2012; Schroeder 1965). The T_{60} time for each of the 6 trials and used the average of the results as a single value, $T_{60} = 0.552\,\mathrm{s}$.

3.3.3. Properties of the Microphone arrays

For the acquisition of the real-world audio data, two tetrahedral microphone arrays with different baseline lengths B were used: $B = 0.3\,\mathrm{m}$ and $B = 0.6\,\mathrm{m}$, called *ARRAY30* and *ARRAY60* respectively. This approach was chosen to allow the evaluation of the influence of the baseline length of the microphone array on the performance of the sound source localization algorithms. Since tetrahedral arrays were used, B is the distance between each of the microphones. The center of the microphone array $\mathbf{m}_\mathrm{C}$ is the mean of the coordinates of the microphones $\mathbf{m}_i$ of the array:

$$\mathbf{m}_\mathrm{C} = \frac{1}{N_\mathrm{m}} \sum_{i=0}^{N_\mathrm{m}} \mathbf{m}_i. \tag{3.5}$$

Maximum TDoA $\Delta T_{A_\mathrm{max}}$, observable using the array of baseline length B is

$$\Delta T_{A_\mathrm{max}} = \frac{B}{c_{20}}. \tag{3.6}$$

For ARRAY30, $T_{A_\mathrm{max}30} = 8.82 \times 10^{-4}\,\mathrm{s}$. At $f_s = 44\,100\,\mathrm{Hz}$, this corresponds to 38 samples. For ARRAY60, the $T_{A_\mathrm{max}60} = 1.76 \times 10^{-3}\,\mathrm{s}$ or 77 samples.

The coordinates of the microphones of the array with $B = 0.3\,\mathrm{m}$ and the coordinates of the microphones of the array with $B = 0.6\,\mathrm{m}$ are provided in Table 3.6 and Fig. 3.20. Note that the geometry of the array does not exactly match a

tetrahedron. This is due to the error of manually placing the microphones onto the microphone stands.

Table 3.6. Position of microphones of the *ARRAY30* array microphones (30 cm aperture) and the *ARRAY60* array microphones (60 cm aperture)

		Microphone coordinate		
Array	Microphone index i	x, m	y, m	z, m
ARRAY30	1	1.45	1.14	1.42
	2	1.425	0.84	1.42
	3	1.58	0.975	1.63
	4	1.295	1.025	1.63
	Array center $\mathbf{m}_\mathrm{C}$	1.4375	0.995	1.525
ARRAY60	1	1.49	1.325	1.36
	2	1.385	0.715	1.34
	3	1.72	0.975	1.78
	4	1.12	1.055	1.78
	Array center $\mathbf{m}_\mathrm{C}$	1.429	1.018	1.565

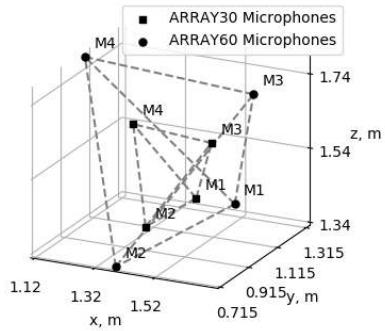


Fig. 3.20. The positions of the ARRAY30 and ARRAY60 microphones; dashed lines denote the edges of the tetrahedrons

Each tetrahedral array consists of four identical conndenser microphones (RØDE M2). Since the directivity pattern of the RØDE M2 microphone is cardioid shaped, we have positioned the microphones in such a way that the acoustic axes of the microphones were oriented up-wards, so that the directivity of the microphones would be close to omnidirectional in a horizontal plane. The position reference point of each microphone coincided with the center of its membrane.

3.3.4. Setting of the Sound Sources

The real-world audio data was recorded using each of the previously described arrays with one or two simultaneously active sound sources. The sound sources were represented by two small loudspeakers: battery-powered JBL GO loudspeaker (Source 1, SJ), mounted on a tripod to allow for a convenient positioning; and Yamaha MSP3 amplified two-way compact monitor loudspeaker (Source 2, SY), placed on a portable pedestal or a table. The position of the sound source is determined by a reference point.

For both sound sources the reference points were located in the center of the front grid of the speakers.The speech signals that were reproduced through the speakers were obtained from the AMI Corpus (Carletta *et al.* 2006), headset microphone mix (file ES2019a.Mix-Headset.wav). To allow for the two simultaneously active sound sources to reproduce different signals, we have selected two excerpts from the file, each with a duration of 60 s. The first excerpt (E1) began at the 70-th second of the source audio file, and the second excerpt (E2) began at the 310-th second of the file. Ten positions for Source 1 were randomly selected from a uniform distribution in the entire volume of the room.

While all three coordinates were randomly chosen for the tripod-mounted Source 1, Source 2 could only be placed on a fixed height pedestal or the table. Thus its z coordinate z_2 is limited to two values: 0.85 m and 0.865 m above ground; x and y coordinates are the same for both source positions. The coordinates of the Source 1 (x, y, z_1) and Source 2 (x, y, z_2) of the selected positions are presented in Table 3.7. As can be seen from the Table 3.7, the average of coordinates of all source positions are very close to the geometric center of the room and differs from it no more than 8.25% (for x coordinate). The positions of the sources and the centers of both arrays are also presented in Fig. 3.21.

By converting the Cartesian coordinates of the positions of the sound sources to polar coordinates, with the centers of the microphone arrays at the origin of the polar coordinate system, DoAs of sound sources were obtained (presented in Fig. 3.22). DoA with azimuth $\theta = 0$ and elevation $\phi = 0$ corresponds to the positive x axis of the Cartesian coordinate system.

For the single active sound source case, only Source 1 was used, and it was placed at all ten positions (coordinates of which are expressed as (x, y, z_1)). For the two active sound source case, ten positions of the Source 2 were selected from the Table 3.7 sequentially, while the positions of the Source 1 were selected from the Table 3.7 and randomly permuted, resulting in 10 combinations presented in Table 3.8. The speech signal excerpts were assigned to the sound sources in an alternating manner.

Table 3.7. Sound source positions of used for the acquisition of the dataset

Position No.	x, m	y, m	z_J, m	z_Y, m
1	4.0	4.85	1.3	0.85
2	4.2	2.7	1.665	0.85
3	1.81	5.55	1.57	0.85
4	3.02	3.38	0.57	0.85
5	0.43	3.7	2.42	0.865
6	1.06	2.14	0.94	0.85
7	0.43	1.04	1.72	0.865
8	2.71	2.15	1.665	0.85
9	3.47	0.38	0.84	0.85
10	1.33	5.08	2.38	0.865
Standard deviation	1.423	1.734	0.613	0.007
Average coordinate	2.246	3.097	1.507	0.8545
Room center coordinate	2.69	2.925	1.42	1.42

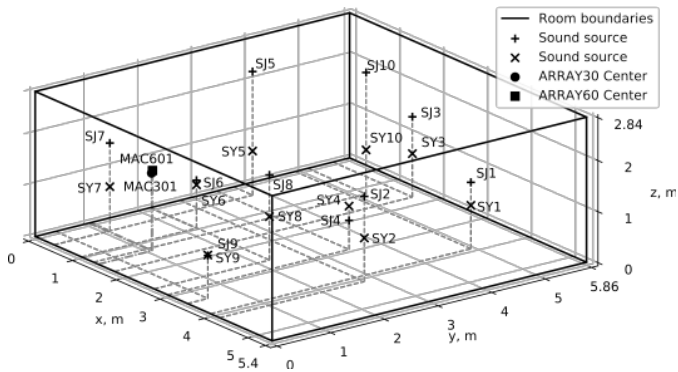


Fig. 3.21. Positions of the sources (SJ_i and SY_i where $i = 1, 2, \ldots, 10$ denotes the positions of Source 1 (JBL GO) and Source 2 (Yamaha MSP3) respectively, as presented in Table 3.7) and the centers of ARRAY30 (MAC301) and ARRAY60 (MAC601) within the room

3.3.5. Estimation of the Room Additional Acoustic Properties

To obtain the average absorption coefficient of the room a, a value of the T_{60} reverberation time is needed. This value was calculated from the impulse response of the room. The reverberation time T_{60} was calculated for each of the obtained RIRs using Schroeder's backward integration method. The results are presented in Fig. 3.23. The average T_{60} value was $T_{60} = 552\,\text{ms}$, with standard deviation of $33.6\,\text{ms}$.

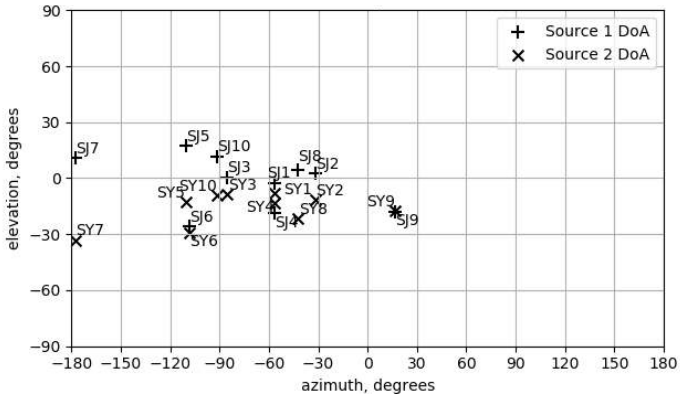


Fig. 3.22. DoAs for source positions presented in Table 3.7, relative to the center of the ARRAY30

Table 3.8. Positions of sound sources in case of two simultaneously active sound sources and sources' corresponding signals

Source 1		Source 2	
Position number	Signal excerpt	Position number	Signal excerpt
1	E2	2	E1
2	E1	6	E2
3	E2	7	E1
4	E1	3	E2
5	E2	10	E1
6	E1	1	E2
7	E2	5	E1
8	E1	9	E2
9	E2	4	E1
10	E1	8	E2

The absorption coefficient was calculated using (3.3) with $T_{60} = 0.552\,\mathrm{s}$:

$$a = 0.1611 \frac{V}{S \cdot 0.552} = 0.206. \tag{3.7}$$

Schroeder's frequency was calculated using a measured room volume and T_{60}:

$$F_c \approx 2000 \left(\frac{0.552}{89.869} \right)^{0.5} = 156.79\,\mathrm{Hz}. \tag{3.8}$$

The measurements of RIRs were compared with the computer simulation of a virtual room with the same dimensions and the placement of the IR measure-

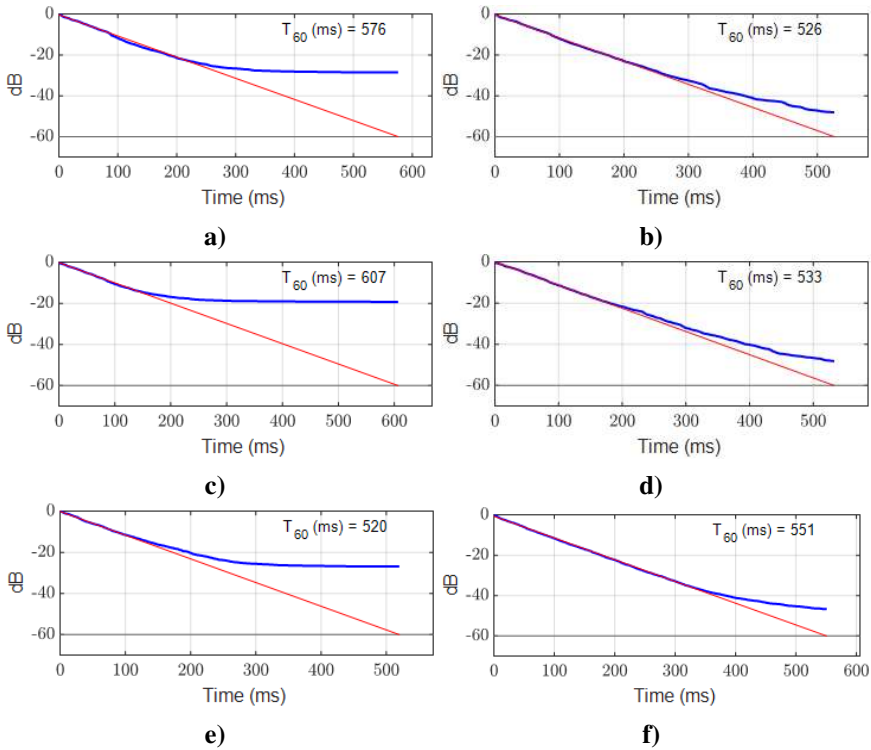
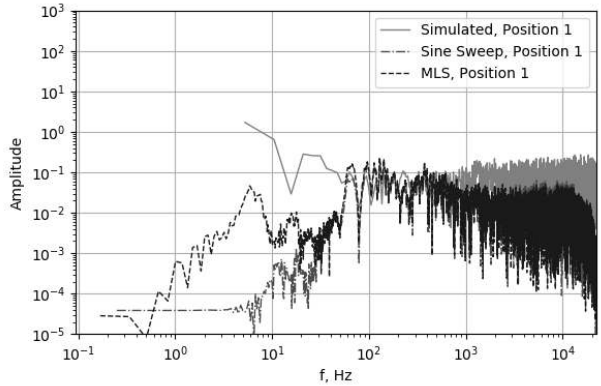


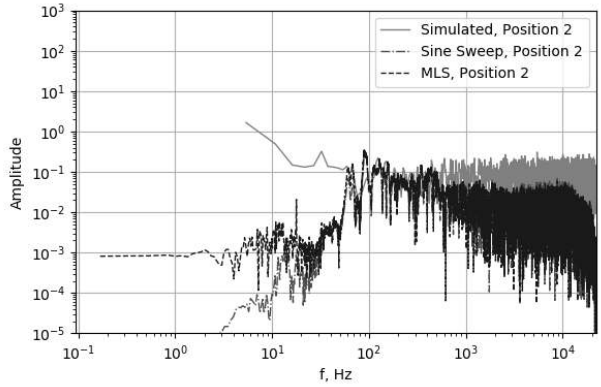
Fig. 3.23. Results of the T_{60} estimation using Schroeder's backward impulse response integration; a) MLS, position 1; b) Swept Sine, position 1; c) MLS, position 2; d) Swept Sine, position 2; e) MLS, position 3; f) Swept Sine, position 3

ment sound source and microphone, using Python programming language and pyroomacoustics package, which uses image-source method for impulse response calculation (Scheibler *et al.* 2018). For the simulation, the absorption coefficient a, calculated in (3.7) was used, while the maximum order of reflection was set to 10. By performing the Fast Fourier Trans-form (FFT) of the RIRs, transfer functions of the room were obtained (magnitude spectra of the transfer functions presented in Fig. 3.24).

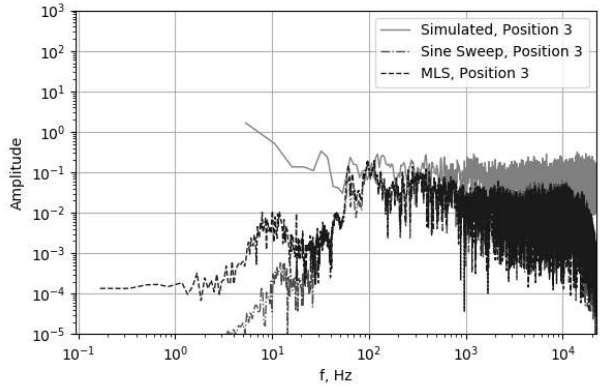
As can be observed from the magnitude spectra of the transfer functions in all RIR measurement positions, the simulation is relatively accurate only in the approximate frequency range from 60 Hz to 500 Hz. This range starts at a frequency that is more than twice lower than Schroeder's frequency of the room and does not encompass the widely used telephone band (ITU-T, Rec. P.342, 2009).



a)



b)



c)

Fig. 3.24. Magnitude spectra of the transfer functions obtained from the RIR measurements (using Sine Sweep and MLS methods and computer simulation) at positions of sources and microphones presented in Table 3.5; a) position 1; b) position 2; c) position 3

Thus, the auralization results using simulated RIRs might be inaccurate and unsuitable for reliable evaluation of the performance of sound source localization algorithms using speech signals. For all three measurement positions, the amplitude of the simulated transfer function is significantly higher in the low-frequency range than in the measured RIRs. This can be addressed to a) the unsuitability of the image source for RIR simulation in the low frequency range (wave-based phenomena, such as diffraction and interference, are not properly recreated (Siltanen *et al.* 2010)) and b) inaccuracy of the real-world RIR measurements, as it relies on the linearity of the transfer functions of the transducers (measurement sound source and microphone, which are not linear. The diffraction effect is stronger at low frequencies where the wavelength is longer than or comparable to the dimensions of the reflecting objects (Siltanen *et al.* 2010), that is, lower than Schroeder's frequency.

The frequency response of Thump12 loudspeaker presents a steep roll-off in the sound pressure level below 70 Hz and above 6 kHz, so it is impossible to obtain a fully accurate RIR using neither Swept Sine nor MLS method using such loudspeaker. Considering these findings, it is advisable to evaluate SSL algorithms not only synthetic or semi-synthetic audio data but also on real-world audio data as the simulated audio signals might not accurately reflect the real-world situation.

To sum up this section, a dataset of four different scenarios (two tetrahedral microphone arrays with different baseline lengths, one and two active sound sources for each type of array) was created, with ten different source positions (in case of two active sound sources – 10 two source position combinations) for each scenario. Positions of sound sources were distributed evenly in the room, with average of coordinates of all sources differing from the geometric center of the room no more than 8.25% (for x coordinate).

A set of 6 room impulse responses was measured using three different combinations of source-microphone positions, using two IR acquisition techniques: MLS and Swept Sine. The reverberation time T_{60} was estimated from the RIR using Schroeder's method, and the average reverberation time T_{60} was determined to be $0.552\,\mathrm{s}$. The average surface absorption coefficient was derived from the reverberation time and the geometry of the room and was determined to be $a = 0.206$. The Schroeder's frequency of the room was calculated to be $F_c = 156.76\,\mathrm{Hz}$.

A computer simulation of a virtual room with the same geometry and acoustical parameters as the real-world room was performed. From the comparison of results, it was determined that the magnitude spectra of real-world and simulated RIRs differ considerably both in low and high-frequency ranges, and the simulation is relatively accurate only in the approximate frequency range from 60 Hz to 500 Hz. Thus, if a sound source localization method or algorithm is being developed, its evaluation on real-world audio data is crucial as the simulated audio signals might not accurately reflect the real-world situation.

3.4. Source Localization using Correlation Based Features

In this section, presented is the process of the generation of the training datasets and the training of the CNN. First, the training/testing dataset format and dataset creation process are presented. Then, the CNN training procedure is outlined. Finally, the results are provided and discussed.

3.4.1. Synthesis of the Dataset Records

In total, 3 training/testing datasets have been generated, each containing 20000 samples, differing in the number of sound sources and the Gaussian kernel spread. Summary of the datasets is presented in Table 3.9. All other parameters regarding the auralization, training and target data generation were as described in Section 2.4.2. For all datasets, the resolution of the DoA map was $R_{\mathrm{DoA}} = 18$, resulting in a 36×18 DoA map elements.

Table 3.9. Summary of different datasets used for CNN evaluation

Dataset	No. of sources	σ
dataset1	1	2
dataset2	2	1
dataset4	2	2

Training data was generated in 2 second duration cases. For each case, the virtual sound sources were simulated at random positions and remained stationary throughout the case. With the frame length of 2048 samples with 1024 sample overlap, one case produced 84 samples. Thus, there were 238 different cases of sound source position.

3.4.2. Training of the Convolutional Neural Network

Each CNN architecture was trained for 100 epochs on 20000 samples, with 4 samples per batch. We have evaluated the performance of two CNN architectures, CONV-WE-CCFB and CONV-CCFB-DOA, described in section 2.4, using 3 datasets, created by the method described in the same section. We have evaluated the performance of both CNN architectures with learning rates of $\mathrm{lr} = 0.001$ and $\mathrm{lr} = 0.01$.

The mean absolute angular error (MAE) and the standard deviation of the absolute angular error (ESD) were evaluated for a grid of points inside the volume of the virtual room with grid spacing in all directions being $1\,\mathrm{m}$. For evaluation of the absolute angular error, we have found the centroids of the blobs in the DoA map

using connected components labeling with thresholding and for both the ground truth and the estimation and calculated the Euclidean distance between them. In case of $N > 1$ sources, N blobs were located in the ground truth and the estimation DoA map. Euclidean distances between all centroids of the blobs were calculated, and N smallest values were observed. After iterating through all grid points, MAE and ESD were calculated. For MAE and ESD evaluation, we have used newly generated samples and not the samples from the training dataset.

3.4.3. Evaluation of Source Localization using Correlation Features

The ground truth and the estimation of a single active sound source DoA map with $\sigma = 2$ using CONV-WE-CCFB network is presented in Fig. 3.25, and using CONV-CCFB-DOA network – in Fig. 3.26. Also in these figures present are the markers ($\times$) representing the coordinates of the centroids of the blobs, detected in the DoA maps (indicated in red). In the estimation, the azimuth and elevation error is also indicated. A pattern caused by overfitting may be observed in CONV-CCFB-DOA estimation, while CONV-WE-CCFB does not present such property.

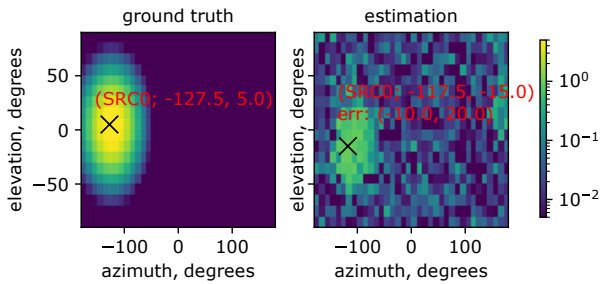


Fig. 3.25. Ground truth and estimation of a single sound source DoA map with $\sigma = 2$, CONV-WE-CCFB network

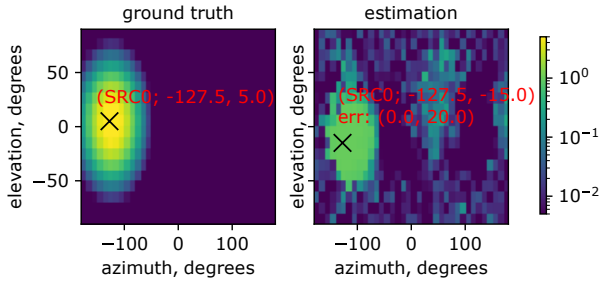


Fig. 3.26. Ground truth and estimation of a single sound source DoA map with $\sigma = 2$, CONV-CCFB-DOA network

In Fig. 3.27 and Fig. 3.28 the ground truth and estimation of a two sound sources DoA map with $\sigma = 1$ by CONV-WE-CCFB and CONV-CCFB-DOA networks respectively are presented. There are no distinct patterns in any of the DoA map estimates, indicating that both neural networks may have learned to generalize. Results of estimation if the DoA map of two simultaneously active sound sources with a CNN (CONV-CCFB-DOA) trained on a dataset in which was only one active sound source, are presented in Fig. 3.29. As can be seen from the figure, CNN was unable to produce more than one distinct blob in the estimation.

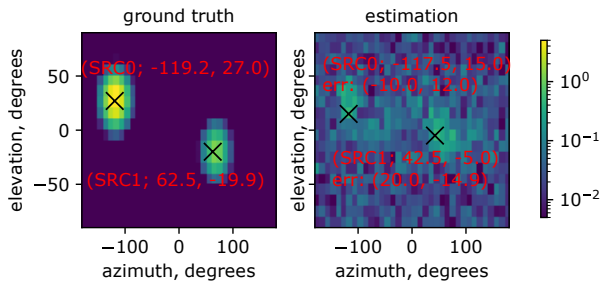


Fig. 3.27. Ground truth and estimation of a two sound sources DoA map with $\sigma = 1$, CONV-WE-CCFB network

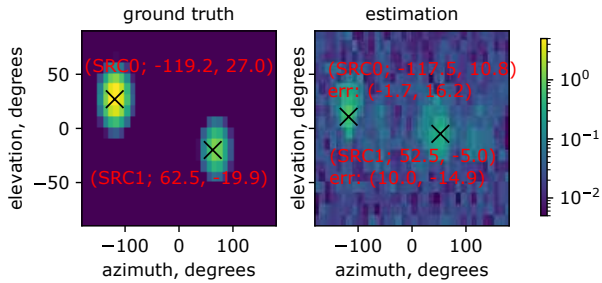


Fig. 3.28. Ground truth and estimation of a two sound sources DoA map with $\sigma = 1$, CONV-CCFB-DOA network

In Table 3.10, the results of CNN training and the estimation error evaluation are presented. We have separately selected the best results for single sound source cases and for two sound source cases. For a single active sound source, CONV-CCFB-DOA architecture performed better regarding all metrics.

For two sound source cases CONV-CCFB-DOA had smallest MAE when trained with learning rate $\mathrm{lr} = 0.01$. Second best result was achieved using CONV-WE-CCFB network, trained with $\mathrm{lr} = 0.001$, and in this case, the ESD was smallest, thus suggesting that such architecture is the most suitable for multiple sound source localization using of all evaluated cases.

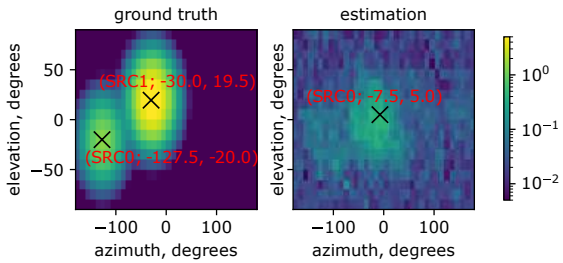


Fig. 3.29. Ground truth and estimation of a two sound sources DoA map with $\sigma = 2$, CONV-CCFB-DOA network trained on dataset with single active sound source

Table 3.10. Results of the CNN training and estimation error evaluation (best results highlighted)

No. of sources	σ	Learning rate	Model	MAE, °	ESD, °
1	2	0,001	CONV-WE-CCFB	29.97	57.64
1	2	0,001	CONV-CCFB-DOA	**22.67**	**48.02**
2	1	0,001	CONV-WE-CCFB	25.61	**19.58**
2	1	0,001	CONV-CCFB-DOA	27.17	36.75
2	2	0,001	CONV-CCFB-DOA	31.44	34.08
2	1	0,01	CONV-WE-CCFB	28.30	34.98
2	1	0,01	CONV-CCFB-DOA	**25.22**	29.29

Least CNN test loss was achieved with CONV-CCFB-DOA network, but the MAE and ESD were moderate for this case. Highest training accuracy was achieved with CONV-CCFB-DOA when the target DoA map was generated with $\sigma = 2$.

For CNNs trained on dataset1 with the same learning rate, the differences in MAE and ESD for the two different CNN architectures were not significant, with $32.1\,\%$ MAE difference and $32.1\,\%$ ESD difference. For CNNs trained on dataset2 with the same learning rate, MAE difference was $5.7\,\%$ and ESD difference was $46.7\,\%$. Tenfold increase of the learning rate generally increased the training accuracy for both networks, but decreased the test accuracy for CONV-WE-CCFB by $1.4\,\%$ while increasing the test accuracy for CONV-CCFB-DOA by $26.6\,\%$. Also, the increase of the learning rate increased the MAE by $10.5\,\%$ and ESD by

78.7 % for CONV-WE-CCFB network, while decreasing MAE by 7.2 % and ESD by 20.3 % for CONV-CCFB-DOA network.

Further investigation is needed on the preparation of the input features and the training data (audio frame sizes, number of filterbank bands, filter order, DoA map resolution, mapping of the probability density values, and other parameters).

Further research on optimal ANN architectures for SSL is needed. Usage of depthwise convolution layers and depth-separable convolutional layers might be of benefit, since the information between channels of the CCFB feature would be preserved. Also, an investigation of using separate convolutional layers for each channel, merging them at a later point might be of interest. Usage of multiple types of features (complex hybrid network architecture with merging of several sub-networks) could be investigated. To sum up, an ANN hyperparameter search and optimization might provide a deeper insight on the solution of multiple sound source localization using ANNs. A research on training the model in different acoustic spaces to obtain a generalized estimator for sound source DoA, that works in any acoustic situation, might be of interest.

CCFB as input features can be utilized for multiple sound sources DoA map estimation, which can in turn be used for sound source localization and separation. A method for obtaining the training data for the CNN (CCFBs and DoA maps) was proved to be effective.

From the results presented in Section 3.4.3, these main points can be concluded:

1. DoA map estimation using CCFB as input features is a viable method for SSL. Both CNN architectures may be used for sound source localization. Sound source DoA estimation absolute angular error best case for single active sound source was 22.67° and the worst case was 29.97°. For two active sound sources, best case MAE was 25.22° and worst case was 31.44°.
2. CNN trained on a dataset in which one source was intermittently active may be used to estimate the DoA of a single source, while CNN trained with a dataset in which two sources were intermittently active may be used to estimate the DoAs of both one and two simultaneously active sound sources.
3. The proposed CNN architecture (CONV-CCFB-DOA) outperforms the architecture, adapted from He *et al.* (2018a) (CONV-WE-CCFB) for a single sound source localization. This may be addressed to the greater numbers of trainable parameters of the CONV-CCFB-DOA network, (total of 24 204 380 parameters, versus CONV-WE-CCFB total of 651 972 parameters).

4. Generally, a larger and more complex CNN performs better for single sound source localization and benefits from higher learning rates, but the effects of overfitting may become apparent.

3.5. Two-Dimensional Source Localization using Phase Based Features

In this section, experiments to estimate the azimuth and elevation of single and multiple sound sources using a CNN with STFT input features and the results of such experimentation are presented.

Firstly, the creation of the training/testing dataset that is used for the evaluation of the performance of the CNN is presented. Then, the procedure of the evaluation is described. Lastly, the results of the experimentation are provided and discussed.

3.5.1. Preparation of the Training and Testing Dataset

To evaluate the performance of the proposed method, a set of datasets for training and testing were synthesized. Training datasets were synthesized with white noise as the sources' signals and the target DoA maps were synthesized with $Q \in [5, 10, 20]^\circ$ and $\sigma \in [5, 10, 15, 20]$. Training datasets contain 100000 samples each. Training datasets were created with the STFT frequency random permutation, also without permutation, with one, two, or three active sound sources. Each sample in the datasets contains a matrix of input features and a desired output.

The testing datasets were created with speech signals from AMI Corpus Carletta *et al.* (2006) without STFT scrambling, assuming W-disjoint orthogonality of speech signals.

The proposed structure of CNN was trained on each of the training datasets and evaluated its performance using a testing dataset with the corresponding DoA heatmap grid resolution and Gaussian spread. A Keras implementation of CNN training was used during experimental investigation.

The microphone array's signals were synthesized using an image source model implemented in Pyroomacoustics package (Scheibler *et al.* 2018). The acoustic signals were simulated in a cuboid shaped acoustic enclosure with dimensions matching a real room described in Section 3.3. The tetrahedral microphone array was set to have an arbitrarily selected side length of $0.4\,\mathrm{m}$ and its center was placed at an arbitrary location within an acoustic enclosure.

For all experiments the geometry of the microphone array, its position, and orientation remained constant. Simulated acoustic source coordinates were selected from an uniform random distribution within the volume of the simulated acoustic

enclosure. CNN was trained on a training dataset with 100000 samples during 5 epochs with learning rate of 0.001.

3.5.2. Evaluation of the Performance of the Proposed Method

To compare the performance of the proposed method with alternatives, the Steered Response Power Phase Transform (SRP-PHAT) algorithm was used as a baseline. `pyroomacoustics` Python package implementation was used for SRP-PHAT calculation, which allows to estimate the response power of the beamformer and present it as a 2D (azimuth and elevation) heatmap, which is compatible with the output of the proposed method. SRP-PHAT DoA heatmaps were estimated at the same resolution as with the proposed CNN-based method.

The Mean Average Error (MAE) of source 2D DoA predictions were obtained using the proposed method and the baseline method. DoA estimation error is the Euclidean distance in the polar coordinate system between the estimated source DoA and the ground truth DoA.

The ground truth DoA is calculated geometrically from known source and microphone array positions. The estimated DoA is obtained from the DoA heatmap using a simple 2D peak detection algorithm. The DoA estimation errors are obtained in 2 steps:

1. Euclidean distances between all pairs or ground truth and estimated DoAs are calculated.
2. N_S smallest errors are selected as the DoA prediction errors for N_S sources.

This two-step approach allows to determine the angular distance between the ground truth and the estimated closest candidate positions.

During the experimental investigation, for each STFT input frame the DoA heatmaps and DoA prediction errors were estimated to test the proposed method and the baseline method. If the peak detection algorithm locates the number of peaks under inequality $N_{\text{Est.}} < N_S$, only $N_{\text{Est.}}$ errors are calculated.

The MAE is calculated using the following equation:

$$\text{MAE} = \frac{1}{N_T} \sum_{i \in N_T} \sum_{j \in N_{\text{Est.}}} e_{ij}. \tag{3.9}$$

The performance of the proposed method was evaluated using several DoA heatmap resolutions and Gaussian kernel spreads (σ). Azimuth and elevation resolution were equal: $Q_\theta = Q_\phi = Q$, as well as azimuth and elevation Gaussian kernel spreads: $\sigma_\theta = \sigma_\phi = \sigma$. Experiments were performed at resolution values $Q \in [5, 10, 20]$ and Gaussian kernel spread values $\sigma \in [5, 10, 15, 20]$ with three

active souns sources. The results are presented in Fig. 3.30. In this figure, MAE of DoA prediction for each testing sample is presented.

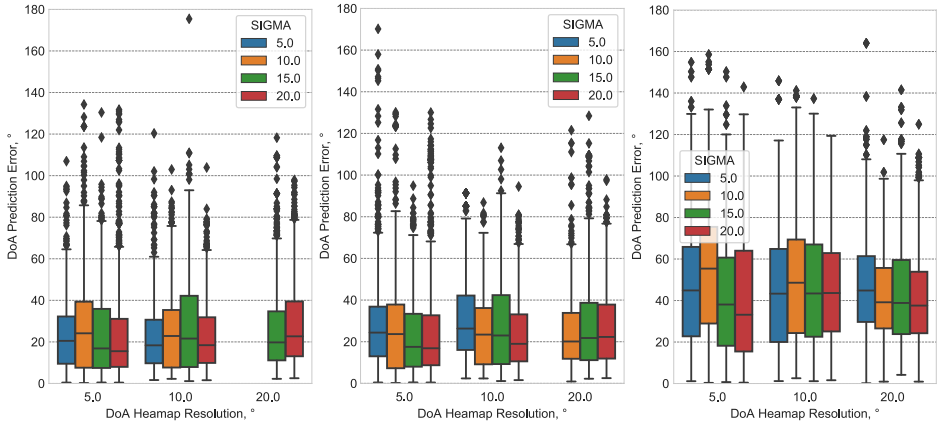


Fig. 3.30. Errors of DoA estimation of three sound sources using the proposed and the baseline method: a) STFT not permuted; b) STFT with permuted time and frequency dimensions; c) SRP=PHAT; data was unavailable for CNN trained on STFT features with permuted time and frequency dimensions with $\sigma \in [5, 10]$ and CNN trained on regular STFT features with $\sigma = 5$

To evaluate the performance of the proposed method when subjected to background and acquisition system noise, experimentation with the best-performing Q and σ configuration was carried out with varying Signal-to-Noise Ratio (SNR) of the simulated microphone array signals. For the evaluation, the training dataset was augmented by adding an uncorrelated noise signal sampled from the uniform distribution to the original signal to obtain a signal with a specific SNR. The MAE of DoA estimation of three simultaneously active speech sources was obtained with testing signals with SNR = $[30, 20, 10]$ dB, and the results are presented in Fig. 57. It can be seen that the angular MAE of three sound source DoA estimation increases with increased noise level (decreased SNR) for both the proposed method and the baseline method. Nevertheless, the method has reached DoA estimation MAE as low as $23.13°$ with 30 dB SNR and $27.21°$ with 10 dB SNR. To compare, SRP-PHAT method gives MAE $51.6°$ and $52.36°$ at respective SNR values. To sum up, the proposed method allows to achieve at least 48 % lower DoA estimation angular MAE than SRP-PHAT at all evaluated SNR values.

To determine the influence of the CNN architecture on the performance of the proposed method, 3 architecture variations were additionally evaluated, having only a single convolutional layer, two convolutional layers and the originally

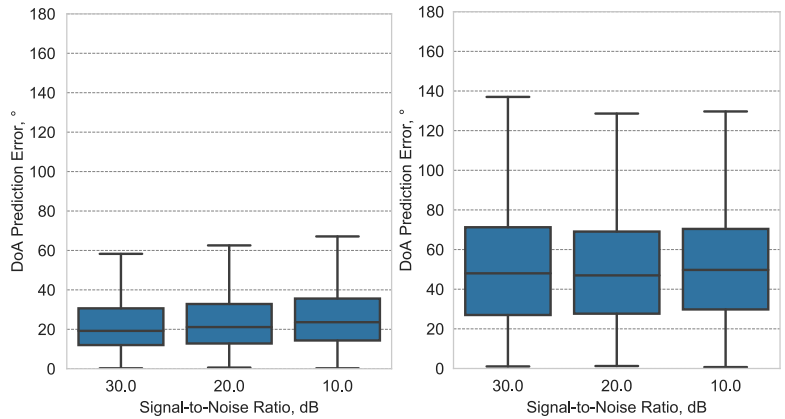


Fig. 3.31. Angular errors of source DoA estimation at different input signal SNR values: a) the proposed method; b) the baseline method; $Q = 5°$, $\sigma = 20$

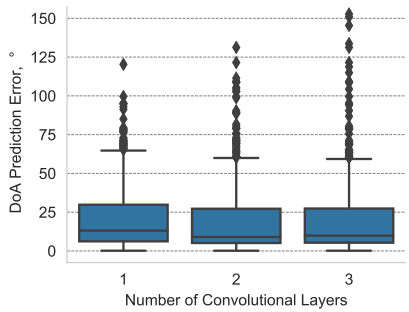
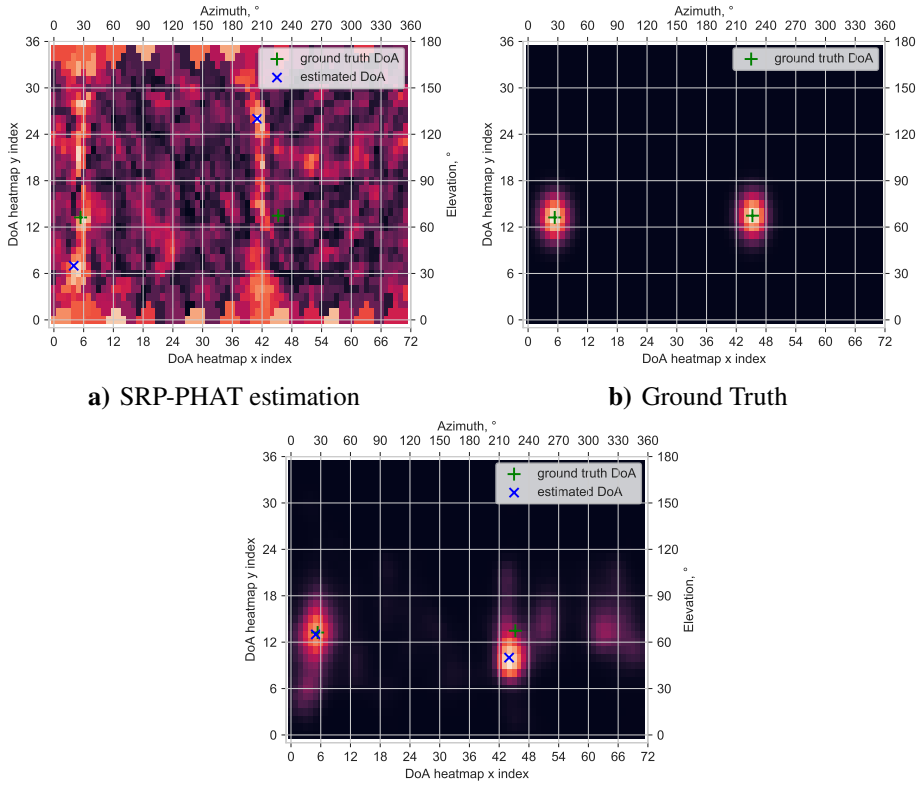


Fig. 3.32. Angular errors of source DoA estimation at different number of CNN convolutional layers; $Q = 10°$, $\sigma = 10$

proposed architecture with three convolutional layers, with $10°$ angular resolution output layer (36x18 elements), trained on a dataset with target feature $\sigma = 10$. After evaluation of these CNN architecture variations on a dataset with 3 active speech sources, it was discovered, that higher number of convolutional layer contributes positively in reducing the MAE of source DoA estimation, as shown in Fig. 58. With only a single convolutional layer in the CNN, source DoA estimation MAE was $19.8°$, while increasing the number of convolutional layers to 3 allowed to achieve source DoA estimation MAE of $18.14°$, which is a 8.4 % improvement.

Examples of DoA heatmaps are presented in Fig. 3.33. These examples were obtained for an array audio frames with two speech sources active at DoAs situated respectively at $(-153.1°, -23.8°)$ and at $(46.3°, -22.6°)$. An example of a spatial



a) SRP-PHAT estimation

b) Ground Truth

c) CNN Estimation

Fig. 3.33. Examples of DoA heatmap output: a) SRP-PHAT spatial power spectrum; b) ground truth (used as a target for training of the CNN; $Q_\theta = Q_\phi = 5°$); c) CNN estimated DoA heatmap ($Q_\theta = Q_\phi = 5°$); same STFT input feature was used for both SRP-PHAT and the proposed method

power spectrum extracted using SRP-PHAT algorithm is presented in 3.33a. Here the SRP objective function is evaluated on a grid with an angular resolution $Q_\theta = Q_\phi = 5°$). An example ground truth DoA heatmap that is used to train the CNN is presented in Fig. 3.33b. The angular resolution of the DoA heatmap is the same as SRP-PHAT spatial spectrum. The Gaussian spread selected to prepare the the desired outputs for this CNN training was $\sigma_\theta = \sigma_\phi = 10$. An example of CNN DoA heatmap estimation using the proposed method is presented in Fig. 3.33c.

As can be seen from the Fig. 3.30, the proposed CNN-based source localization method outperforms the baseline SRP-PHAT algorithm in estimating the azimuth and elevation of multiple acoustic sources. While the lowest source DoA estimation MAE was 25° for the baseline method, at $Q = 5°$ and $\sigma = 20$, pre-

sented CNN-based method achieved MAE of $16°$ with the same Q and σ values. This can be interpreted as performance increase by $36\,\%$.

Generally, the method presented here outperformed the baseline in all experiments by at least $29\,\%$, with the largest performance increase by $70\,\%$ at $Q = 5°$ and $\sigma = 10$. Thus, it can be concluded that the proposed CNN-based multiple acoustic source 2D DoA estimation algorithm allows for a more precise source DoA estimation than the SRP-PHAT-based method.

3.6. Three-Dimensional Source Localization using Phase Based Features

In this section, the experimental evaluation of CNN application with STFT phase features for sound source 3D position estimation is presented. As in the previous sections, first, the dataset format and its creation procedure is outlined. Then the training of the CNN is briefly presented. Lastly, the results of the CNN performance evaluation are presented and discussed.

3.6.1. Preparation of the Training and Evaluation Datasets

To train the CNN and to evaluate the performance of the CNN at various Q and σ as well as the number of sound sources, multiple datasets were generated.

The performance of the CNN trained on datasets with one or two simultaneously active sound sources is evaluated with 3D grid resolution $Q \in [0.25, 0.5, 1]$ and Gaussian blurring of the 3D grid $\sigma \in [0.25, 0.5, 1]$. Source positions were selected randomly within the limits of a simulated acoustic enclosure with dimensions of $5.4\,\mathrm{m}$, $5.86\,\mathrm{m}$ and $2.84\,\mathrm{m}$ in x, y and z dimensions, respectively. Microphone positions $\mathbf{m}_i = [m_{ix}, m_{iy}, m_{iz}]$, $i \in [1, 2, 3, 4]$ of a tetrahedral microphone array are presented in Table 3.11.

Table 3.11. Positions of the microphones of the tetrahedral microphone array

Microphone index, i	m_{ix}, m	m_{iy}, m	m_{iz}, m
1	3.0	2.0	2.0
2	3.4	2.0	2.0
3	3.2	2.35	2.0
4	3.2	2.12	2.35

For a particular N_S, source positions were generated once and used to generate dataset variants with different signals (noise or speech), Q and σ; 2 datasets in total. These are the source positions ground truth datasets (see Fig. 3.34).

For the two-source dataset, source positions were generated in sets of two source positions per set, and the array audio was simulated for each of the source position sets. For one position set, multiple frames (samples) are generated.

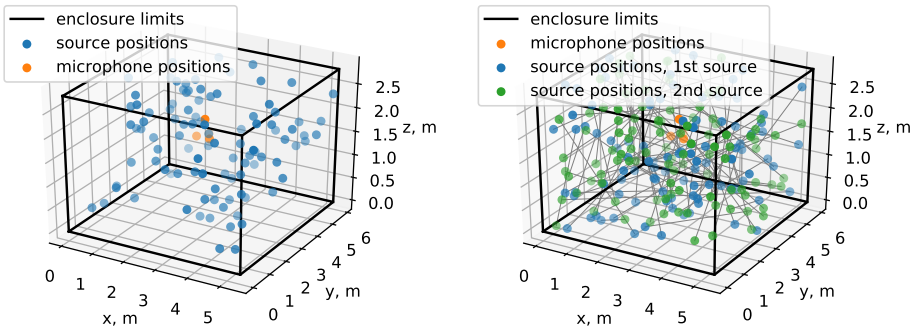


Fig. 3.34. Source positions used for dataset generation; a) single source; b) two sources (lines connect source positions for the same position set)

STFT datasets were created once for each N_S, with noise and speech signals, 4 datasets in total. These are the input feature datasets. Synthetic STFT datasets with noise and speech signals were generated using image-source method Scheibler *et al.* (2018). Example STFT features are shown in Fig. 3.35.

Training/testing target 3D grids were generated for each N_S and for each Q and σ; 18 datasets in total (9 for a single source, 9 for two sources). These are the training target datasets.

Training input and target feature datasets were generated only using noise signals at 1×10^5 source positions, with one STFT frame per position, resulting in 1×10^5 training samples. Noise signals were generated dynamically during the dataset creation; samples of these signals were sampled randomly from an uniform distribution and a gain of 0.9 was applied, creating a white noise signal.

Evaluation datasets were generated using both noise and speech signals at 100 source positions with 314 STFT frames at each position, resulting in 31400 evaluation samples. Multiple frames per single source position were generated because the speech signal is non-stationary and the prediction result for an input frame depends on the audio content of a particular audio frame from which the input feature was generated; thus it is desired to evaluate each source position using more than one speech signal frame. For the creation of the speech evaluation dataset, speech signals were randomly selected for each source position from the AMI Corpus (Carletta *et al.* 2006), from a subset of dry microphone recordings that are of 5 s or greater duration (longer records were truncated to 5 s).

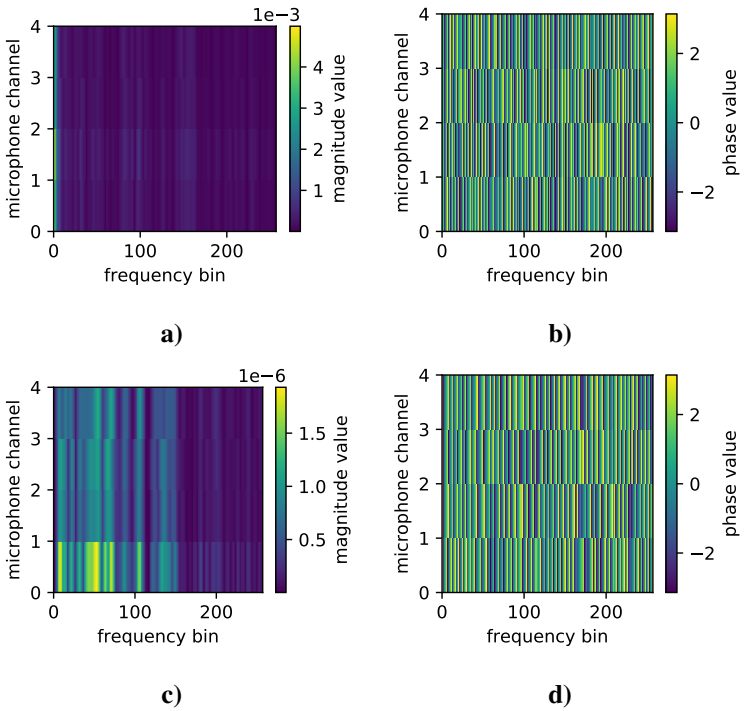


Fig. 3.35. STFT feature examples of noise and speech signals; a) Single noise source STFT magnitude; b) Single noise source STFT phase; c) Two speech sources STFT magnitude; d) Two speech sources STFT phase

For the noise signal evaluation STFT dataset, white noise signals were generated in the same manner as described earlier and saved to files for reuse, in contrast with the training dataset, where white noise signals were generated on fly.

3.6.2. Evaluation of the Convolutional Neural Network

Same CNN architecture, except with different output layer shape (to accommodate the number of 3D grid elements dictated by the Q of the grid), was trained on noise signal datasets containing 1×10^5 samples that were created for each of the $Q \in [0.25, 0.5, 1]$ and $\sigma \in [0.25, 0.5, 1]$ as described in Section 3.6.1. CNN were trained for 100 epochs. For each Q and σ, CNN was trained anew and the model saved separately for later performance evaluation.

To evaluate the performance of each trained CNN model, testing datasets were used. For each of the testing datasets, a 3D grid output feature was predicted by the CNN for each STFT phase input feature of the dataset.

Table 3.12. Source position estimation MAE values at different Q and σ; minimum MAE highlighted

		3D grid maximum		k-means clustering			
		1 source				2 sources	
		Noise	Speech	Noise	Speech	Noise	Speech
Q, m	σ, m	MAE, m	MAE, m	MAE, m	MAE, m	MAE, m	MAE, m
0.25	0.25	2.51	2.60	0.79	0.94	2.74	2.74
0.25	0.50	1.26	1.39	**0.62**	**0.76**	1.10	1.09
0.25	1.00	**0.99**	**1.10**	0.81	0.91	1.18	1.17
0.50	0.25	2.18	2.32	0.67	0.86	1.19	1.17
0.50	0.50	2.29	2.35	0.69	0.82	**1.08**	**1.08**
0.50	1.00	1.05	1.14	0.84	0.94	1.18	1.18
1.00	0.25	2.73	2.73	0.97	1.10	1.41	1.40
1.00	0.50	1.92	2.00	0.81	0.91	1.17	1.16
1.00	1.00	1.11	1.22	0.89	0.99	1.20	1.20

From the predicted 3D grid, the source coordinates were estimated using the methods described in Section 3.6. For a single sound source, 3D grid maximum or k-means cluster centers were estimated; for two sound sources, only the k-means clustering approach was used.

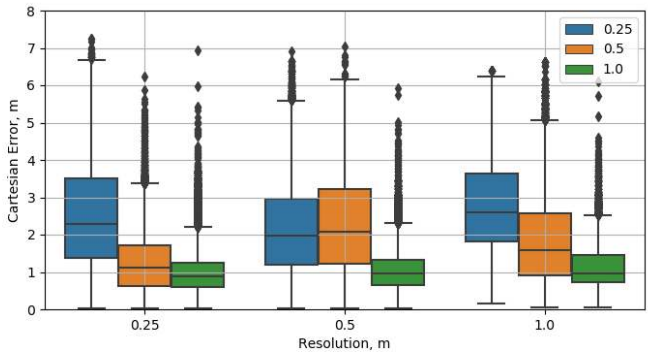
After evaluating all trained CNN architectures, MAEs were calculated between the estimated source(s) position(s) and the ground truth source(s) positions(s). The results are provided in Table 3.12.

Source position estimation errors are presented in box-plot representation in Fig.s 3.36, 3.37 and 3.38, respectively, for 1 source position estimation from 3D grid maximum, 1 source position estimation from 3D grid thresholding and k-means clustering, and 1 sources position estimation from 3D grid thresholding and k-means clustering. It can be seen from the Fig. 3.36, that the dispersion of the source position estimation error is higher for small σ values and decreases with increased σ. This tendency can be interpreted as the evidence of the ability of the CNN to learn the spatial smoothness of the acoustic features. The σ of the Gaussian function that is used during the creation of the desired output features for CNN training represents the probability density function of acoustic feature classification. Larger σ translates to higher spatial smoothness of the acoustic feature classification. It can be concluded that σ values that are higher than the spatial resolution Q of the CNN predicted 3D grid allows to achieve lower source position estimation errors and lower dispersion of these errors.

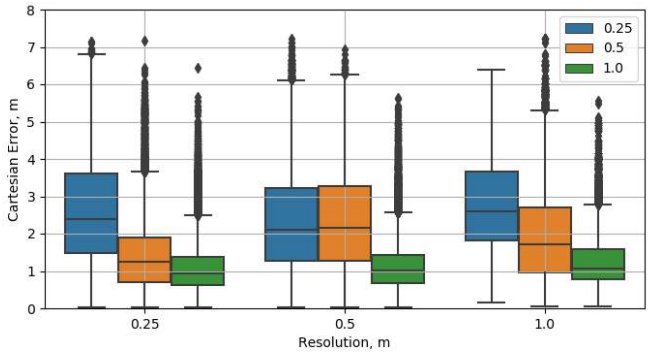
The same tendency can be observed in two sound source localization scenario when the source positions are estimated using thresholding and k-means clustering, although only when the spatial resolution of the predicted 3D grid is fine ($Q = 0.25\,\mathrm{m}$).

Nevertheless, when the source positions are estimated from thresholded and clustered predicted 3D grid, the influence of the σ and even the resolution is much smaller compared to when the source positions are estimated from the 3D grid global maximum. This can be attributed to the fact that the thresholding level depends on the mean of all predicted 3D grid values, and while higher σ values result in more high-valued elements in the grid, the thresholding level rises accordingly, and the relative number of thresholded grid elements to the total number of grid elements remains relatively constant for the same number of active sound sources.

It is also worth noting that using the clustering-based source position estimation method, the uncertainty of source position can be lower than the resolution of the output grid, because the coordinates of the estimated cluster centers are not quantized to this resolution – cluster center coordinates are continuous.



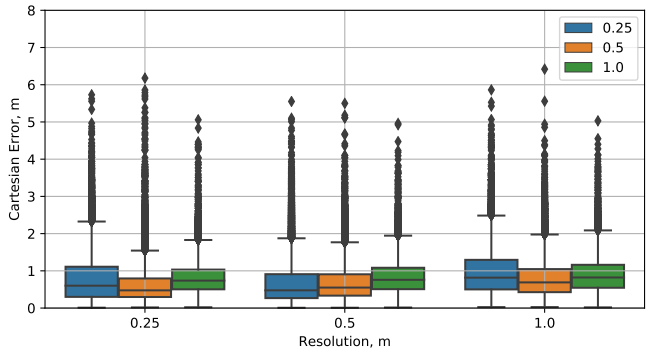
a)



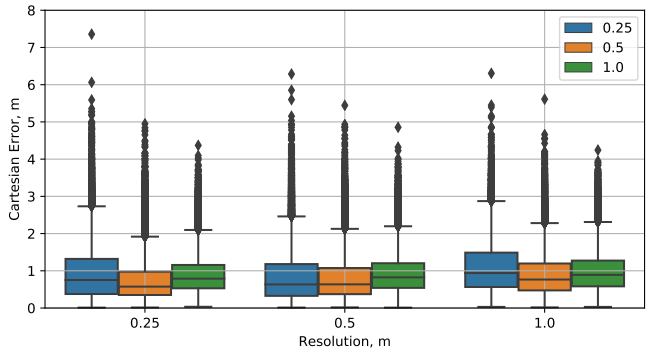
b)

Fig. 3.36. Single source position estimation errors at different resolution and sigma values; a) noise source; b) speech source; coordinates obtained from 3D grid maximum

Additionally, as can be seen from Table 3.12 and Figs 3.36, 3.37 and 3.38, the single source position estimation error is smaller for noise sources compared to the speech sources. This can be attributed to the fact that the CNN was trained using samples with noise sources, so the features learned by the convolutional layers of the CNN might be better suited for the classification of input features obtained with an active noise source.



a)



b)

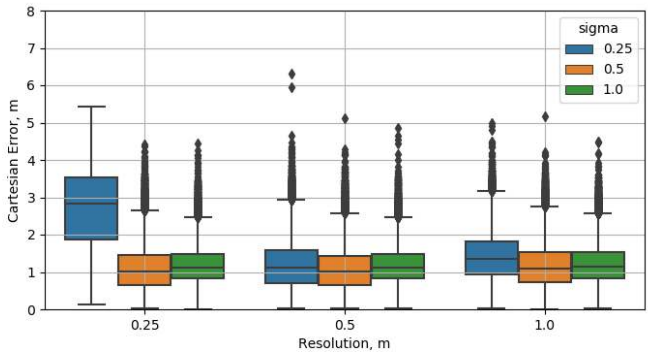
Fig. 3.37. Single source position estimation errors at different resolution and sigma values; a) noise source; b) speech source; coordinates obtained via k-means clustering

On the other hand, for two source scenario, the source estimation errors for both noise and speech signals differ only up to $1\,\%$. It can be speculated that with more speech sources present in the acoustic scene, the spectra of the microphone array signals become less sparse and thus becomes more similar to the spectra of the microphone array signals, obtained with the noise sources present in the

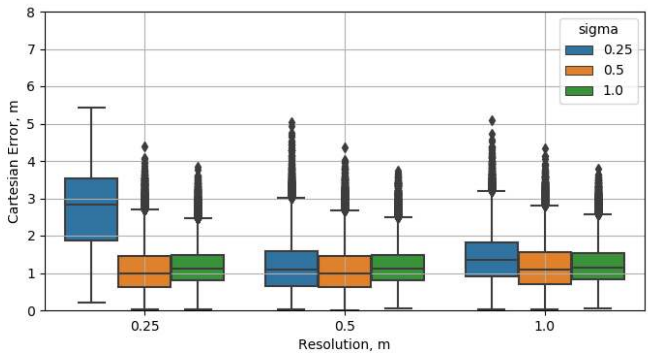
acoustic scene. Thus, the acoustic features presented to the CNN are more similar for noise and speech source localization scenarios.

An illustration of a ground truth 3D grid with a single active source ($Q = 0.25\,\mathrm{m}$, $\sigma = 1\,\mathrm{m}$) and the corresponding 3D grid, predicted by the CNN, is shown in Fig. 65 (noise source). The center of the Gaussian blob in the ground truth 3D grid corresponds to the position of the sound source. In the predicted 3D grid, the coordinates of the element with the maximum value, converted to metric coordinates using Q factor, are considered the estimated source coordinates (in case of the 3D grid maximum coordinate acquisition method).

Source position estimation via 3D grid thresholding and k-means clustering is illustrated in Fig. 3.40. Clustering method is available for both single source and multiple source scenarios.

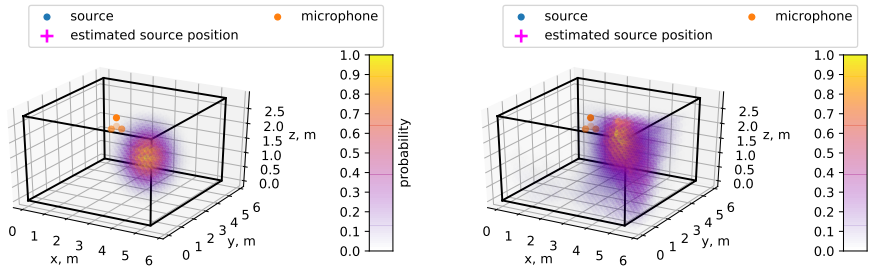


a)



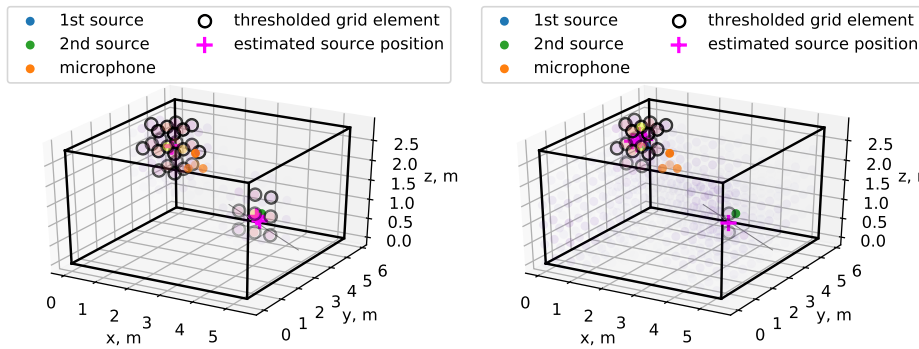
b)

Fig. 3.38. Source position estimation errors at different resolution and sigma values; a) noise sources; b) speech sources



a) ground truth **b)** estimation

Fig. 3.39. Examples of CNN output 3D grid for single noise source; a) ground truth; b) CNN estimation



a) ground truth **b)** estimation

Fig. 3.40. Examples of CNN output 3D grid for two sound sources; a) ground truth; b) CNN estimation

3.6.3. Discussion of the Experimental Investigation

As can be seen from the Table 3.12, the lowest estimation MAE for single noise source localization using 3D grid maximum finding as the coordinate estimation method was $0.99\,\mathrm{m}$ with grid resolution $Q = 0.25\,\mathrm{m}$ and $\sigma = 1\,\mathrm{m}$.

For single speech source localization, the lowest MAE of $1.1\,\mathrm{m}$ was achieved at the same Q and σ values. When the k-means clustering source coordinate estimation method is used, the lowest MAE for both noise ($0.62\,\mathrm{m}$) and speech ($0.76\,\mathrm{m}$) signals are achieved at $Q = 0.25\,\mathrm{m}$ and $\sigma = 0.5\,\mathrm{m}$, which is a 37 % improvement for noise source localization and 31 % improvement for speech source localization.

For 2 source localization, the smallest MAE is $1.08\,\mathrm{m}$ for both noise and speech signals were achieved at $Q = 0.5\,\mathrm{m}$ and $\sigma = 0.5\,\mathrm{m}$.

1. It was shown, that it is possible to localize one or two sound sources within a 3-dimensional space using a CNN with STFT phase component of the tetrahedral microphone array signals as the input feature.
2. Using thresholded CNN output 3D grid k-means cluster centers as the source position estimate
   2.1. it is possible to estimate the position of single noise source with estimation MAE of $0.62\,\mathrm{m}$ and of speech source with estimation MAE of $0.76\,\mathrm{m}$ with grid resolution $Q = 0.25\,\mathrm{m}$ and $\sigma = 0.5\,\mathrm{m}$.
   2.2. it is possible to estimate the position of two noise or speech sources with estimation MAE of $1.08\,\mathrm{m}$, with grid resolution $Q = 0.5\,\mathrm{m}$ and $\sigma = 0.5\,\mathrm{m}$.
3. Using thresholded CNN output 3D grid k-means cluster centers as the source position estimate instead of 3D grid maximum coordinates, there is at least 31 % decrease in mean absolute error of a single sound source position estimation.

3.7. Conclusions of The Third Chapter

1. Moderate precision of single sound source localization may be received using a MLP with only ILD features sent to the input of the network, with estimated sound source position prediction mean absolute error was as low as $1.58\,\mathrm{m}$ in computer simulation and $0.41\,\mathrm{m}$ in practical experimentation.
2. Semi-supervised GRNN trained on SRP-PHAT features can localize a single sound source with source position estimation MAE that is averagely 5 times lower than using a geometrical source localization method, and averagely $3.5\,\%$ lower than using the SRP-PHAT intensity map method at low feature fitness threshold levels. The most suitable ratio of supervised to unsupervised loss is found to be $\mu = 0.6$. Overall smallest source position error is achieved with 1 nearest graph neighbor considered during graph training dataset creation.
   2.1. Using an ISOMAP NLDR algorithm, it is possible to embed SRP-PHAT acoustic features to a $\mathbb{R}^2$ space and the embedded dimensions correspond to the spatial dimensions of the acoustic enclosure.
   2.2. Embeddings themselves correspond to the x and y coordinates of the sound source.
   2.3. It is possible to localize a single sound source within an acoustic enclosure with a data-driven algorithm that:

- Is semi-supervised learning based;
- Is trained on a an unbalanced ($\mathrm{N_l} << \mathrm{N_u}$) training dataset.

3. A CNN with CCFB as input features can be used to estimate single sound source DoA with mean absolute angular error as low as 22.67° and two sound sources with mean absolute angular error as low as 25.22°.
4. A CNN with an STFT phase input feature can estimate azimuth and elevation of two sound sources with mean absolute angular error as low as 16° with DoA heatmap resolution $Q = 5^\circ$ and $\sigma = 20^\circ$, and outperform a baseline SRP-PHAT algorithm by $36\,\%$.
5. It is possible to localize one or two sound sources within a 3-dimensional space using a CNN with STFT phase component of the tetrahedral microphone array signals as the input feature.

   5.1. Using centers of k-means clusters of thresholded CNN output 3D grid as the source position estimate:

      i) it is possible to estimate the position of single noise source with estimation MAE of 0.62 m and of speech source with estimation MAE of 0.76 m with grid resolution Q = 0.25 m and σ = 0.5 m.

      ii) it is possible to estimate the position of two noise or speech sources with estimation MAE of 1.08 m, with grid resolution Q = 0.5 m and σ = 0.5 m.

   5.2. Using thresholded CNN output 3D grid k-means cluster centers as the source position estimate instead of 3D grid maximum coordinates, there is at least 31 % improvement in the accuracy of single sound source position estimation.

General Conclusions

The hypotheses were confirmed by the investigation of the results presented in this dissertation.

1. Using a graph regularized artificial neural network trained with a semi-supervised learning strategy with SRP-PHAT input features it is possible to achieve up to 5 times lower mean absolute error of a single sound source localization than using SRP-PHAT-based geometrical source localization method.
2. Using a convolutional neural network with cross-correlation in frequency bands as an input feature, the mean absolute localization error of two sound sources remains above 25°.
3. Phase component of the spectra of the microphone array signals can be successfully used as an acoustic feature to localize multiple sound sources in two-dimensional and three-dimensional space using artificial neural networks.
4. Using a convolutional neural network with microphone array signal spectrum phase component as an input feature, it is possible to estimate the direction of arrival of three sounds sources with a mean absolute error of 16°, with the angular resolution of the network output $Q = 5^\circ$ and Gaussian kernel spread $\sigma = 20^\circ$, which is a $36\,\%$ improvement compared to SRP-PHAT evaluated on a grid of same resolution.

5. It is possible to localize one and two sound sources in a three-dimensional space using a convolutional neural network with spectrum phase component of the tetrahedral microphone array signals as the input feature. Using the k-means clustering-based method for the source position estimation instead of the convolutional neural network three-dimensional output grid maximum coordinates, there is at least 31 % decrease in the mean absolute error of the estimation of a single sound source position.

References

Adavanne, S.; Politis, A.; Nikunen, J.; Virtanen, T. 2019a. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, *IEEE Journal of Selected Topics in Signal Processing* 13(1): 34–48. [see 27, 28 p.]

Adavanne, S.; Politis, A.; Virtanen, T. 2018. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Rome, Italy. [see 27 p.]

Adavanne, S.; Politis, A.; Virtanen, T. 2019b. Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network. [see 28 p.]

Allen, J. B.; Berkley, D. A. 1976. Image method for efficiently simulating small-room acoustics, *The Journal of the Acoustical Society of America* 65(4): 943–950. [see 28 p.]

Argentieri, S.; Danes, P.; Soueres, P. 2015. A survey on sound source localization in robotics: From binaural to array processing methods, *Computer Speech and Language* 34(1): 87–112. [see 1, 12, 26 p.]

Astapov, S.; Berdnikova, J.; Preden, J.-S. 2015. Optimized acoustic localization with srp-phat for monitoring in distributed sensor networks, *International Journal of Electronics and Telecommunications* 59(4): 383–390. [see 13 p.]

Athanasopoulos, G.; Verhelst, W.; Sahli, H. 2015. Robust speaker localization for real-world robots, *Computer Speech and Language* 34(1): 129–153. ISSN 0885-2308. [see 2 p.]

Bianco, M.; Gannot, S.; Gerstoft, P. 2020. Semi-supervised source localization with deep generative modeling, in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Eespo, Finland. [see 27 p.]

Bohlender, A.; Spriet, A.; Tirry, W.; Madhu, N. 2021. Exploiting temporal context in CNN based multisource DoA estimation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 1594–1608. [see 26 p.]

Brandstein, M.; Silverman, H. 1997. A robust method for speech signal time-delay estimation in reverberant rooms, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Munich, Germany: IEEE Computer Society Press, 375–378. ISBN 978-0-8186-7919-3. [see 10, 32 p.]

Brutti, A.; Omologo, M.; Svaizer, P. 2008. Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection, in *Hands-Free Speech Communication and Microphone Arrays*, Trento, Italy: IEEE, 69–72. ISBN 978-1-4244-2337-8. [see 18 p.]

Cao, Y.; Iqbal, T.; Kong, Q.; An, F.; Wang, W.; Plumbley, M. 2021. An improved event-independent network for polyphonic sound event localization and detection, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada. [see 28 p.]

Cao, Y.; Iqbal, T.; Kong, Q.; Galindo, M.; Wang, W.; Plumbley, M. D. 2019. Two-stage sound event localization and detection using intensity vector and generalized cross-correlation, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 26 p.]

Carletta, J.; ; *et al.* 2006. The AMI Meeting Corpus: A Pre-announcement, in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*: MLMI'05, Berlin, Heidelberg: Springer-Verlag, 28–39. ISBN 978-3-540-32549-9. [see 53, 92, 103, 109 p.]

Chakrabarty, S.; Habets, E. 2019a. Multi-scale aggregation of phase information for reducing computational cost of CNN based DoA estimation, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain. [see 26 p.]

Chakrabarty, S.; Habets, E. 2019b. Multi-speaker DoA estimation using deep convolutional networks trained with noise signals, *IEEE Journal of Selected Topics in Signal Processing* 13(1): 8–21. [see 26, 57, 61, 64 p.]

Champagne, B.; Bedard, S.; Stephenne, A. 1996. Performance of time-delay estimation in the presence of room reverberation, *IEEE Transactions on Speech and Audio Processing* 4(2): 148–152. ISSN 10636676. [see 18 p.]

Chazan, S.; Hammer, H.; Hazan, G.; Goldberger, J.; Gannot, S. 2019. Multi-microphone speaker separation based on deep DoA estimation, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain. [see 28 p.]

Chytas, S.; Potamianos, G. 2019. Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 27 p.]

Datum, M. S.; Palmieri, F.; Moiseff, A. 1996. An artificial neural network for sound localization using binaural cues, *The Journal of the Acoustical Society of America* 100(1): 372–383. ISSN 0001-4966. [see 32 p.]

Diaz-Guerra, D.; Miguel, A.; Beltran, J. 2021. Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 300–311. [see 27 p.]

DiBiase, J. H.; Silverman, H. F.; Brandstein, M. S. 2001. *Robust localization in reverberant rooms*. Springer. 157–180 p. [see 16, 32 p.]

Ding, H.; Bao, Y.; Huang, Q.; Li, C.; Chai, G.Three-dimensional localization of point acoustic sources using a planar microphone array combined with beamforming, *Royal Society Open Science* 5(12): 181 407. [see 27 p.]

Do, H. T. H. 2009. *Real-time SRP-PHAT Source Location Implementations on a Large-aperture Microphone Array*. , Brown University. [see 19 p.]

El Badawy, D.; Dokmanic, I. 2018. Direction of Arrival With One Microphone, a Few LEGOs, and Non-Negative Matrix Factorization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(12): 2436–2446. ISSN 2329-9290. [see 12 p.]

Elko, G. W.; Anh-Tho Nguyen Pong 1997. A steerable and variable first-order differential microphone array, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Munich, Germany, 223–226 vol.1. [see 21 p.]

Grondin, F.; Glass, J.; Sobieraj, I.; Plumbley, M. 2019. Sound event localization and detection using CRNN on pairs of microphones, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 26 p.]

Grumiaux, P.-A.; Kitic, S.; Girin, L.; Guerin, A. 2021a. Improved feature extraction for CRNN-based multiple sound source localization, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland. [see 26 p.]

Grumiaux, P.-A.; Kitic, S.; Girin, L.; Guérin, A. 2021b. A Survey of Sound Source Localization with Deep Learning Methods, *arXiv:2109.03465 [cs, eess]* Available online at: http://arxiv.org/abs/2109.03465. [see 26 p.]

Guizzo, E.; Gramaccioni, R. F.; Jamili, S.; Marinoni, C.; Massaro, E.; Medaglia, C.; Nachira, G.; Nucciarelli, L.; Paglialunga, L.; Pennese, M.; *et al.* 2021. L3das21 challenge: Machine learning for 3d audio signal processing, *arXiv:2104.05499* . [see 28 p.]

Habets, E. A. 2006. Room impulse response generator, *Technische Universiteit Eindhoven, Technical Report* 2(2.4): 1. [see 53 p.]

Hack, P. 2015. *Multiple Source Localization with Distributed Tetrahedral Microphone Arrays*. , University of Music and Performing Arts Graz. [see 13 p.]

Hak, C. C. J. M.; Wenmaekers, R. H. C.; L. C.J. Luxemburg, V. 2012. Measuring room impulse responses : impact of the decay range on derived room acoustic parameters, *Acta Acustica united with Acustica* 98(6): 907–915. ISSN 1610-1928. [see 90 p.]

Hao, Y.; Kucuk, A.; Ganguly, A.; Panahi, I. 2020. Spectral fluxbased convolutional neural network architecture for speech source localization and its real-time implementation, *IEEE Access* 8: pp. [see 26 p.]

He, W.; Motlicek, P.; Odobez, J.-M. 2018a. Deep neural networks for multiple speaker detection and localization, in *International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 74–79. [see 12, 22, 27, 51, 52, 55, 88, 102 p.]

He, W.; Motlicek, P.; Odobez, J.-M. 2018b. Joint localization and classification of multiple sound sources using a multi-task neural network, in *Proceedings of the Interspeech Conference*, Hyderabad, India, 312–316. [see 28 p.]

He, W.; Motlicek, P.; Odobez, J.-M. 2019. Adaptation of multiple sound source localization neural networks with weak supervision and domainadversarial training, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 770–774. [see 53, 88 p.]

Hirvonen, T. 2015. Classification of spatial audio location and content using convolutional neural networks, *Journal of the Audio Engineering Society* . [see 26 p.]

Huang, Y.; Tong, J.; Hu, X.; Bao, M. 2021. A Robust Steered Response Power Localization Method for Wireless Acoustic Sensor Networks in an Outdoor Environment, *Sensors* 21(5): 1591. [see 27, 28 p.]

Huang, Y.; Wu, X.; Qu, T. 2018. DNN-based sound source localization method with microphone array, Beijing, China. [see 27 p.]

Huang, Y.; Wu, X.; Qu, T. 2020. "A time-domain unsupervised learning based sound source localization method," in Int, *IEEE International Conference on Information Communication and Signal Processing (ICICSP 2020)* 26–32. [see 26, 27 p.]

Hubner, F.; Mack, W.; Habets, E. 2021. Efficient training data generation for phase-based DoA estimation, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada. [see 26 p.]

Jenrungrot, T.; Jayaram, V.; Seitz, S.; Kemelmacher-Shlizerman, I. 2020. The cone of silence: speech separation by localization, in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada. [see 27 p.]

Jourjine, A.; Rickard, S.; Yilmaz, O.; Yilmaz, O. 2000. Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2985–2988. [see 26 p.]

Kapka, S.; Lewandowski, M. 2019. Sound source detection, localization and classification using consecutive ensemble of crnn models, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 26, 27, 28 p.]

Kim, Y. 2014. Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 1746–1751. [see 26 p.]

Kim, Y.; Ling, H. 2011. Direction of arrival estimation of humans with a small sensor array using an artificial neural network, *Progress In Electromagnetics Research* 27: 127–149. [see 26 p.]

Knapp, C.; Carter, G. 1976. The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24(4): 320–327. [see 18 p.]

Kotus, J. 2013. Multiple sound sources localization in free field using acoustic vector sensor, *Multimedia Tools and Applications* 74(12): 4235–4251. ISSN 1380-7501, 1573-7721. [see 1, 2 p.]

Kucuk, A.; Ganguly, A.; Hao, Y.; Panahi, I. 2019. Real-time convolutional neural network-based speech source localization on smartphone, *IEEE Access* 7: 169–969. [see 27 p.]

Laufer-Goldshtein, B.; Talmon, R.; Gannot, S. 2016. Semi-Supervised Sound Source Localization Based on Manifold Regularization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(8): 1393–1407. ISSN 2329-9290, 2329-9304. [see 29, 58, 63 p.]

Le Roux, J.; Vincent, E.; Hershey, J. R.; Ellis, D. P. 2015. Micbots: Collecting large realistic datasets for speech and audio research using mobile robots, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia: IEEE, 5635–5639. ISBN 978-1-4673-6997-8. [see 87 p.]

Lin, Y.; Wang, Z. 2019. A report on sound event localization and detection, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 27 p.]

Lopatka, K.; Kotus, J.; Czyzewski, A. 2011. Application of Vector Sensors to Acoustic Surveillance of a Public Interior Space, *Archives of Acoustics* 36(4). ISSN 0137-5075. [see 2 p.]

Lu, Z. 2019. Sound event detection and localization based on CNN and LSTM, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 27 p.]

Löllmann, H. W.; Evers, C.; Schmidt, A.; Mellmann, H.; Barfuss, H.; Naylor, P. A.; Kellermann, W. 2018. The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking, in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, 410–414. [see 12, 13, 28, 32, 52, 87 p.]

Ma, N.; Brown, G.; May, T. 2015. Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions, in *Proceedings of the Interspeech Conference*, Dresden, Germany, 160–164. [see 27 p.]

Ma, W.; Liu, X. 2018. Compression computational grid based on functional beamforming for acoustic source localization, *Applied Acoustics* 134: 75–87. ISSN 0003-682X. [see 26 p.]

Moing, G.; Vinayavekhin, P.; Agravante, D.; Inoue, T.; Vongkulbhisal, J.; Munawar, A.; Tachibana, R. 2021. Data-efficient framework for real-world multiple sound source 2D localization, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada. [see 27 p.]

Noh, K.; Choi, J.-H.; Jeon, D.; Chang, J.-H. 2019. Three-stage approach for sound event localization and detection, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 27 p.]

Opochinsky, R.; Laufer-Goldshtein, B.; Gannot, S.; Chechik, G. 2019. Deep ranking-based sound source localization, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New-Paltz, NY, USA, 283–287. [see 27 p.]

Pak, J.; Shin, J. 2019. Sound localization based on phase difference enhancement using deep neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(8): 1335–1345. [see 27 p.]

Park, S.; Suh, S.; Jeong, Y. 2020. Sound event localization and detection with various loss functions, in *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*, 1–5. [see 26 p.]

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* 12(85): 2825–2830. [see 45 p.]

Perotin, L.; Serizel, R.; Vincent, E.; Guerin, A. 2018. CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector, in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 241–245. [see 27 p.]

Pertila, P.; Cakir, E. 2017. Robust direction estimation with convolutional neural networks based steered response power, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, USA, 6125–6129. [see 26 p.]

Phan, H.; Pham, L.; Koch, P.; Duong, N.; McLoughlin, I.; Mertins, A. 2020. Audio event detection and localization with multitask regression network, in *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*. [see 27 p.]

Pujol, H.; Bavu, E.; Garcia, A. 2019. Source localization in reverberant rooms using deep learning and microphone arrays, in *Proceedings of the 23rd International Congress on Acoustics (ICA)*, vol. 149, Aachen, Germany, 4248–4263. [see 27 p.]

Rabenstein, R.; Annibale, P. 2017. Acoustic Source Localization under Variable Speed of Sound Conditions, *Wireless Communications and Mobile Computing* 2017. ISSN 1530-8669. [see 8 p.]

Rickard, S. 2002. On the approximate W-disjoint orthogonality of speech, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, 529–532. [see 26 p.]

Roden, R.; Moritz, N.; Gerlach, S.; Weinzierl, S.; Goetze, S. 2015. On sound source localization of speech signals using deep neural networks, in *Proceedings of the Deutsche Jahrestagung Akustik (DAGA)*, Nuremberg, Germany. [see 26, 27 p.]

Ronchini, F.; Arteaga, D.; Pérez-Lopez, A. 2020. Sound event localization and detection based on CRNN using rectangular filters and channel rotation data augmentation, in *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*. [see 27 p.]

Salvati, D.; Drioli, C.; Foresti, G. 2018. Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions, *IEEE Transactions on Emerging Topics in Computational Intelligence* 2(2): 103–116. [see 27 p.]

Scheibler, R.; Bezzam, E.; Dokmanic, I. 2018. Pyroomacoustics: a Python package for audio room simulation and array processing algorithms, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, Canada, 351–355. [see 53, 95, 103, 109 p.]

Schroeder, M. R. 1965. New Method of Measuring Reverberation Time, *The Journal of the Acoustical Society of America* 37(3): 409–412. ISSN 0001-4966. [see 89, 90 p.]

Schymura, C.; Bönninghoff, B.; Ochiai, T.; Delcroix, M.; Kinoshita, K.; Nakatani, T.; Araki, S.; Kolossa, D. 2021. Pilot: Introducing transformers for probabilistic sound event localization, in *Proceedings of the Interspeech Conference*, Brno, Czechia. [see 27 p.]

Siltanen, S.; Lokki, T.; Savioja, L. 2010. Rays or waves? understanding the strengths and weaknesses of computational room acoustics modeling techniques, in *Proceedings of the International Symposium on Room Acoustics*, Melbourne, Australia. [see 97 p.]

Silverman, H. F.; Ying Yu; Sachar, J. M.; Patterson, W. R. 2005. Performance of real-time source-location estimators for a large-aperture microphone array, *IEEE Transactions on Speech and Audio Processing* 13(4): 593–606. [see 21 p.]

Singla, R.; Tiwari, S.; Sharma, R. 2020. A sequential system for sound event detection and localization using CRNN, in *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*. [see 27 p.]

Sivasankaran, S.; Vincent, E.; Fohr, D. 2018. Keyword-based speaker localization: localizing a target speaker in a multi-speaker environment, in *Proceedings of the Interspeech Conference*, Hyderabad, India. [see 27 p.]

Strauss, M.; Mordel, P.; Miguet, V.; Deleforge, A. 2018. DREGON: Dataset and Methods for UAV-Embedded Sound Source Localization, in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid: IEEE, 1–8. ISBN 978-1-5386-8094-0. [see 88 p.]

Subramanian, A.; Weng, C.; Watanabe, S.; Yu, M.; Yu, D. 2021. Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition, *Computer Speech & Language* . [see 27 p.]

Sundar, H.; Wang, W.; Sun, M.; Wang, C. 2020. Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4642–4646. [see 27 p.]

Suvorov, D.; Dong, G.; Zhukov, R. 2018. Deep residual network for sound source localization in the time domain, *arXiv preprint arXiv:1808.06429* . [see 27 p.]

Takeda, R.; Komatani, K. 2016a. Discriminative multiple sound source localization based on deep neural networks using independent location model, in *IEEE Spoken Language Technology Workshop (SLT)*, virtual Shenzhen, China, 603–609. [see 27 p.]

Takeda, R.; Komatani, K. 2016b. Sound source localization based on deep neural networks with directional activate function exploiting phase information, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 405–409. [see 27 p.]

Takeda, R.; Komatani, K. 2017. Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2217–2221. [see 27 p.]

Vargas, E.; Hopgood, J.; Brown, K.; Subr, K. 2021. On improved training of CNN for acoustic source localisation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 720–732. [see 26, 27 p.]

Vecchiotti, P.; Ma, N.; Squartini, S.; Brown, G. 2019. End-to-end binaural sound localisation from the raw waveform, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 451–455. [see 27 p.]

Vera-Diaz, J.; Pizarro, D.; Macias-Guarasa, J. 2018. Towards endto-end acoustic localization using deep learning: from audio signal to source position coordinates, *Sensors* 18(10): 3418. [see 27, 52 p.]

Vesperini, F.; Vecchiotti, P.; Principi, E.; Squartini, S.; Piazza, F. 2016. A neural network based algorithm for speaker localization in a multi-room environment, in *IEEE International Workshop for Machine Learning for Signal Processing*, Salerno, Italy, 1–6. [see 27 p.]

Wang, Z.; Zhang, X.; Wang, D. 2019. Robust speaker localization guided by deep learning-based time-frequency masking, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(1): 178–188. [see 27 p.]

Weng, J.; Guentchev, K. Y. 2001. Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning, *The Journal of the Acoustical Society of America* 110(1): 310–323. ISSN 0001-4966. [see 52 p.]

Wu, Y.; Ayyalasomayajula, R.; Bianco, M.; Bharadia, D.; Gerstoft, P. 2021. SSLIDE: sound source localization for indoors based on deep learning, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada. [see 27 p.]

Xiao, X.; Xu, C.; Zhang, Z.; Zhao, S.; Sun, S.; Watanabe, S.; Wang, L.; Xie, L.; Jones, D. L.; Chng, E. S.; *et al.* 2016. A study of learning based beamforming methods for speech recognition, in *CHiME 2016 workshop*, San Francisco, USA, 26–31. [see 32 p.]

Xiao, X.; Zhao, S.; Zhong, X.; Jones, D.; Chng, E.; Li, H. 2015. A learning-based approach to direction of arrival estimation in noisy and reverberant environments, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2814–2818. [see 27 p.]

Yalta, N.; Nakadai, K.; Ogata, T. 2017. Sound source localization using deep learning models, *Journal of Robotics and Mechatronics* 29(1): 37–48. [see 27 p.]

Yasuda, M.; Koizumi, Y.; Saito, S.; Uematsu, H.; Imoto, K. 2020. Sound event localization based on sound intensity vector refined by DNNbased denoising and source separation, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, "virtual Barcelona, Spain", 651–655. [see 26 p.]

Youssef, K.; Argentieri, S.; Zarader, J. 2013. A learning-based approach to robust binaural sound localization, in *Proceedings of the IEEE International Workshop on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2927–2932. [see 26, 27 p.]

Zermini, A.; Yu, Y.; Xu, Y.; Wang, W.; Plumbley, M. 2016. Deep neural network based audio source separation, in *Proceedings of the IMA International Conference on Mathematics in Signal Processing*, Birmingham, UK. [see 27 p.]

Zhang, J.; Ding, W.; He, L. 2019. Data augmentation and priori knowledge-based regularization for sound event localization and detection, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. [see 27 p.]

List of Scientific Publications by the Author on the Topic of the Dissertation

Papers in the Reviewed Scientific Journals

Sakavičius, S.; Serackis, A. 2021. Estimation of Azimuth and Elevation for Multiple Acoustic Sources Using Tetrahedral Microphone Arrays and Convolutional Neural Networks, *Electronics* 10(21): 1–12. ISSN 2079-9292. DOI: 10.3390/electronics10212585.

Sakavičius, S. 2021. Investigation of Signal Thresholding Effects on the Accuracy of Sound Source Localization, *International journal of advanced research (IJAR)* 9(9): 80–85. ISSN 2320-5407. DOI: 10.21474/IJAR01/13377.

Sakavičius, S. 2020. Dataset for Evaluation of the Performance of the Methods of Sound Source Localization Algorithms using Tetrahedral Microphone Arrays. *Science – Future of Lithuania. Electronics and electrical engineering* 12 (2020): 1–8. ISSN 2029-2341.

Papers in Other Editions

Sakavičius, S.; Serackis, A. 2019. Estimation of Sound Source Direction of Arrival Map using Convolutional Neural Network and Cross-Correlation in Frequency Bands, in *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1–6.

Sakavičius, S.; Serackis, A.; Plonis, D. 2017. Investigation of the Influence of Attack and Decay Parameters on the Performance of Loudness Control Algorithm, in *5th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE'2017)*, 1–6.

Sakavičius, S.; Plonis, D.; Serackis, A. 2017. Single Sound Source Localization using Multi-Layer Perceptron, in *2017 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1–4.

Summary in Lithuanian

Įvadas

Problemos formulavimas

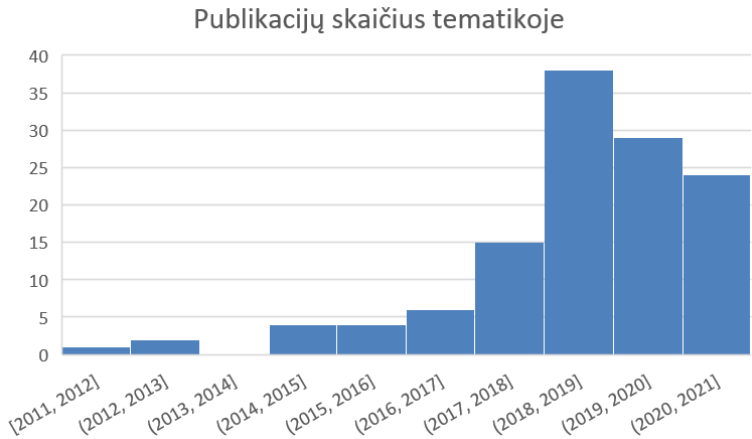
Garso šaltinio lokalizavimas yra svarbus elementas tokiose srityse kaip žmogaus ir kompiuterių sąveika, robotika, autonominės transporto priemonės, saugumas, telekonferencijų sistemos, garso inžinerija ir kitose. Garso šaltinio lokalizavimas apima kalbėtojo vietos nustatymą auditorijoje, įvykių aptikimą aplinkoje ir jų stebėjimą, robotų navigaciją nežinomose aplinkose (Argentieri *et al.* 2015; Kotus 2013).

Dažnai reikia lokalizuoti garso šaltinį tokiu tikslumu, kuris yra artimas ar net lenkia žmogaus gebėjimus lokalizuoti garso šaltinį – nustatyti garso šaltinio kryptį (angl. *Direction of Arrival*) 15° tikslumu. Be to, dažnai reikia nustatyti ne tik garso sklidimo kryptį, bet ir atstumą nuo šaltinio iki imtuvo. Užduotis tampa sudėtingesnė, kai yra poreikis lokalizuoti kelis garso šaltinius aktyvius vienu metu. Dažniausiai žinomi metodai veikia arba pasirinkdami vieną stipriausią garso šaltinį ir slopindami kitus (toks scenarijus vadinamas „vienas prieš daugelį“), arba lokalizuodami kelis garso šaltinius, kurie aktyvūs tuo pačiu metu.

Darbo aktualumas

Didėjantį susidomėjimą mokymu grįstais garso šaltinių lokalizavimo metodais galima iliustruoti šioje srityje publikuotų mokslinių straipsnių skaičiumi, kuris nuo 2011 iki 2019 metų didėjo eksponentiškai (S0.1 pav.).

Šiuolaikiniai garso šaltinio lokalizavimo metodai nėra pakankamai atsparūs aplinkos triukšmams, nepalankioms akustinėms aplinkos sąlygoms, tokioms kaip aidus kambarys,



S0.1 pav. Publikacijų disertacijos tematikoje skaičiaus kitimas 2011–2021 metais

reverberacija. Šiuolaikiniai garso šaltinio lokalizavimo metodai yra pagrįsti sklidimo laiko skirtumo matavimu, tačiau nepakankamai dėmesio skiriama aplinkos triukšmo įvertinimui, reverberacijos ir specifinių akustinių reiškinių mažinimui, kurių didžiausias poveikis metodo rezultatams pastebimas uždaroje erdvėje (**?**). Esamų garso šaltinio lokalizavimo metodų, grįstų sklidimo laiko skirtumo nustatymu, tikslumas mažėja kai aplinkoje pasireiškia reverberacija ir kai aplinkoje yra triukšmo šaltiniai. Todėl pastaruoju metu aktyviai tyrinėjami mokymu grįsti garso šaltinio lokalizavimo metodai. Šiuo metu žinomais mokymu grįstais garso šaltinio lokalizavimo metodais galima nustatyti kelių garso šaltinių sklidimo kryptį, bet ne atstumą iki jų. Taip pat negalime nustatyti garso šaltinių koordinačių trimatėje erdvėje.

Galima bandyti nustatyti kelių garso šaltinio koordinates trimatėje erdvėje formuluojant regresijos uždavinį, tačiau tam iš anksto turi būti žinomas šaltinių skaičius akustinėje scenoje. Taikant prižiūrimo mokymo strategiją, reikalingas didelis kiekis žymėtų pavyzdžių. Tačiau pavyzdžių žymėjimas yra sudėtingas ir imlus laikui uždavinys. Neprižiūrimo mokymosi arba pusiau prižiūrimo (hibridinio) mokymo strategijos leistų sumažinti arba atsisakyti mokymo pavyzdžių žymėjimo, taip sumažinant sprendimų konkrečiai akustinei scenai kūrimo trukmę ir kainą.

Tyrimų objektas

Šio darbo tyrimo objektai yra mokymu grįsti metodai vieno ir daugelio garso šaltinių lokalizavimui, gebantys veikti atsižvelgiant į erdvės akustines savybes, foninį triukšmo lygį, garso šaltinių signalų parametrus ir garso šaltinių judėjimą.

Darbo tikslas

Šio darbo tikslas – pasiūlyti originalius mokymu grįstus metodus garso šaltinio lokalizavimui aidžiose aplinkose.

Darbo uždaviniai

Disertacijos tikslui pasiekti suformuluoti du uždaviniai:

1. Pasiūlyti prižiūrimu mokymu grįstus metodus kelių garso šaltinių dvimačiam lokalizavimui aidžioje aplinkoje ir ištirti jų veikimą;

2. Pasiūlyti prižiūrimu mokymu grįstus metodus kelių garso šaltinių trimačiam lokalizavimui aidžioje aplinkoje ir ištirti jų veikimą;

3. Pasiūlyti hibridiniu mokymu grįstus metodus vieno garso šaltinio lokalizavimui aidžioje aplinkoje ir ištirti jų veikimą.

Tyrimų metodika

Šios disertacijos rengimo metu atlikti tyrimai padalinti į dvi dalis. Pirmoje dalyje buvo bandoma lokalizuoti kelis garso šaltinius aidžioje aplinkoje dvimatėje ir trimatėje erdvėje. Pirmiausia buvo aptartos naujausios garso šaltinių lokalizavimo technologijos ir metodai, o vėliau pasiūlyti mokymu grįsti metodai kaip alternatyva dabartiniams geriausiai veikiantiems algoritmams. Pateikiamas akustinių požymių, kurie gali būti naudojami garso šaltinio padėties nustatymui, tyrimas. Kaip minėto tyrimo dalis buvo pristatyti mokymo metodai, skirti daugelio garso šaltinių lokalizavimui dvimatėje ir trimatėje erdvėje, naudojant dviejų ir trijų matmenų dirbtinio neuronų tinklo (DNT) išėjimo sluoksnio struktūras ir eksperimentiškai įvertintas jų veikimas.

Antroje dalyje buvo ištirtas pusiau prižiūrimu mokymu grįstas garso šaltinio lokalizavimo metodas, apimantis grafu reguliarizuotą DNT (GRDNT). Tai perspektyvi alternatyva šiuo metu žinomiems neprižiūrimu ir pusiau prižiūrimu mokymu grįstiems garso šaltinio lokalizavimo metodams.

Eksperimentiškai siūlomi garso šaltinio lokalizavimo metodai įvertinti atlikus mikrofonų gardelių signalų kompiuterinį modeliavimą naudojant atspindžių (angl. *Image Source*) metodą, o taip pat realių mikrofonų gardelių signalų įrašymą ir apdorojimą kompiuteryje. Buvo sukurti imitaciniai mikrofonų gardelių signalai ir DNT modeliai, kurie buvo apmokyti „Python“ aplinkoje su „pyroomacoustics“ akustinio modeliavimo paketu ir „TensorFlow“ mašininio mokymo paketu. Kitos kompiuterinio modeliavimo užduotys ir skaičiavimai buvo atlikti „Matlab“ arba „Python“ aplinkose. Realūs mikrofonų gardelių signalai buvo gauti naudojant keturių elementų plokščias mikrofonų gardeles ir keturių elementų tetraedrines gardeles , keičiant jų apertūrą.

Darbo mokslinis naujumas

Rengiant šią disertaciją, buvo sukurta nemažai naudingos programinės įrangos, surinkti garso signalų duomenų rinkiniai ir pasiūlyti nauji, originalūs garso šaltinio lokalizavimo metodai:

1. Paruoštas ir viešai paskelbtas tetraedrinių mikrofonų gardelių signalų rinkinys vieno ir dviejų garso šaltinių lokalizavimui aidžioje aplinkoje tirti.

2. Pasiūlyti sąsūkos DNT grįsti metodai kelių garso šaltinių sklidimo krypties nustatymui (dvimatėje erdvėje) naudojant mikrofonų signalų koreliaciją dažnių juostose ar spektro fazės komponentės požymius ir ištirtas jų veikimas.

3. Pasiūlytas sąsūkos DNT grįstas metodas kelių garso šaltinių pozicijos nustatymui trimatėje erdvėje naudojant mikrofonų gardelės signalų spektro fazės komponentę ir išėjimo sluoksnio verčių grupavimą.

4. Ištirtas hibridiniu mokymu ir grafu reguliarizuotu DNT grįstas metodas vieno šaltinio lokalizavimui dvimatėje erdvėje.

Darbo rezultatų praktinė reikšmė

Surinktas ir viešai paskelbtas tetraedrinų mikrofonų gardelių signalų duomenų rinkinys su vienu ir dviem aidžioje aplinkoje esančiais garso šaltiniais: pažymėta ne tik šaltinių ir mikrofonų vieta, bet ir pateikta informacija apie patalpos matmenis bei akustinius parametrus. Šis duomenų rinkinys leidžia palyginti realius ir imituotus mikrofonų gardelių signalus ir nustatyti signalų imitavimo tikslumą. Pateikta imituotų akustinių signalų duomenų rinkinių kūrimo metodika grafu reguliarizuotų neuronų tinklų tyrimams. Pasiūlytas naujas būdas keliems garso šaltiniams lokalizuoti trimatėje erdvėje atsisakant skaičiavimams imlaus požymių išskyrimo. Pasiūlyta metodika garso šaltinių koordinačių nustatymo trimatėje erdvėje tikslumui padidinti naudojant grupavimą.

Ginamieji teiginiai

1. Naudojant GRDNT, mokytą taikant hibridinę mokymo strategiją ir SRP-PHAT požymius įėjime, galima sumažinti vieno garso šaltinio lokalizavimo dvimatėje erdvėje vidutinę paklaidą iki 4% lyginant su SRP-PHAT intensyvumo žemėlapio maksimumo nustatymo metodu.

2. CCFB požymius galima taikyti garso šaltinio krypties nustatymui dvimatėje erdvėje ir pasiekti 23 laipsnių vidutinę paklaidą vieno garso šaltinio atveju ir 26 laipsnių vidutinę paklaidą dviejų garso šaltinių atveju.

3. Taikant spektro fazės komponentę kaip požymį ir naudojant sąsūkos DNT galima pasiekti iki 36 % mažesnę trijų garso šaltinių lokalizavimo dvimatėje erdvėje klaidą nei taikant plačiai taikomą SRP-PHAT algoritmą.

4. Naudojant pasiūlytą sąsūkos DNT išėjimo sluoksnio pakeitimą ir spektro fazės komponentę kaip požymį, aidžioje aplinkoje galima pasiekti 1,08 m vidutinę paklaidą trimatėje erdvėje lokalizuojant du kalbėtojus.

Darbo rezultatų aprobavimas

Darbo rezultatai paskelbti šešiuose moksliniuose straipsniuose. Trys publikacijos atspausdintos recenzuojamuose mokslo žurnaluose, trys publikacijos atspausdintos Lietuvos ir tarptautinių konferencijų straipsnių rinkiniuose. Pagrindiniai disertacijos rezultatai paskelbti penkiose konferencijose:

- Dviejose Lietuvos jaunųjų mokslininkų konferencijose „Mokslas - Lietuvos ateitis“, 2017 ir 2019 metais Vilniuje, Lietuvoje;
- Dviejose tarptautinėse konferencijose „Electrical, Electronic and Information Sciences (eStream)“, 2017 ir 2019 metais, Vilniuje, Lietuvoje;
- Tarptautinėje konferencijoje „Advances in Information, Electronic and Electrical Engineering (AIEEE)“, 2017 metais, Rygoje, Latvijoje.

o

Disertacijos struktūra

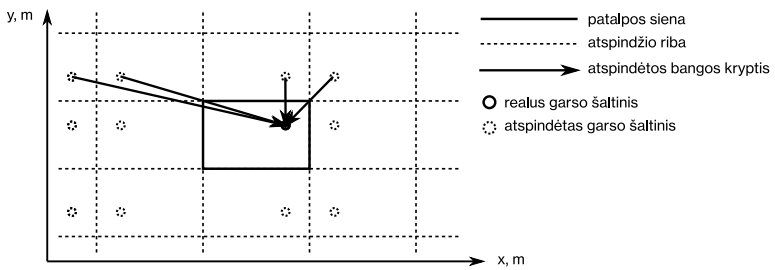
Disertaciją sudaro: įvadas, trys skyriai, bendrosios išvados. Darbo apimtis yra 150 puslapių kuriuose yra pateikta: 97 formulės, 69 paveikslai ir 19 lentelių. Disertacijoje remtasi 101 kitų autorių literatūros šaltiniais.

1. Garso šaltinio lokalizavimo metodų apžvalga

Šiame skyriuje pateikiamos pagrindinės akustikos sąvokos, naudojamos visoje disertacijoje. Čia minimas akustinis scenarijus yra akustinių parametrų rinkinys parametrų terpės, akustinės erdvės, garso lauko, garso šaltinių, imtuvų (mikrofonų) ir susijusių apdorojimo sistemų, o taip pat lokalizavimo užduotis, kurią ketinama atlikti. Šie parametrai bus išsamiai aptarti šiame skyriuje.

Akustinė erdvė – tai bet kokia erdvė, užpildyta akustine terpe, kurioje gali sklisti akustinės bangos. Ideali akustinė erdvė yra izotropinė ir begalinė. Tokiose erdvėse akustinės bangos keliauja nekeisdamos greičio ar krypties (t. y. be lūžio ar atspindžio). Be to, kadangi nėra akustinių ribų, kiti bangų sklidimo reiškiniai (sklaida, difrakcija ir difuzija) taip pat nepasireiškia.

Kita vertus, akustinės erdvės, kurias riboja sienos, laikomos akustiškai uždaromis. Garso šaltinio lokalizavimas akustiškai uždarose erdvėse (patalpose) yra daug sudėtingesnis, palyginti su lokalizavimu atvirose erdvėse. Patalpose garso bangos, kurias skleidžia garso šaltinis, yra linkę atsispindėti nuo sienų, taip sukuriant atspindžius (S1.1 pav.).



S1.1 pav. Šaltinio atvaizdai, sukuriami garso bangų atspindžių nuo patalpos ribinių paviršių

Ryškiausias akustinių atspindžių produktas yra aidas arba reverberacija. Reverberaciją galima vertinti kaip daugelio atsispindėjusių akustinių bangų sumą imtuve, kurių kiekvienas turi skirtingą delsos trukmę ir signalo slopinimą dėl skirtingų sienų garso energijos absorbcijos savybių.

Įprastinių sklidimo laiko vertinimu grįstų garso šaltinio lokalizavimo metodų veikimas žymiai prastėja, kai akustinėje scenoje yra stipri reverberacija – garso bangų atspindžiai neleidžia užtikrintai nustatyti sklidimo vėlinimo trukmės ir to pasekoje lokalizuoti šaltinių. Tai riboja jų taikymą aidžiose aplinkose, ypač kai dominantis signalas yra šneka, ir kai negalima įvertinti ar kompensuoti kanalo efektus prieš įvertinant laiko uždelsimą (Brandstein, Silverman 1997; Silverman *et al.* 2005).

Valdomos kryptingumo charakteristikos (angl. *Steered Beamformer*) pagrindu veikiantys garso šaltinio lokalizavimo metodai yra pagrįsti principu, kad mikrofonų gardelės kryptingumas gali būti valdomas matematiškai naudojant mikrofonų signalų delsą ir sumavimą (arba filtravimą ir sumavimą). Patį paprasčiausią tokio tipo garso šaltinio lokalizavimo metodą galima įgyvendinti skaičiuojant dviejų mikrofonų signalų koreliaciją ir ieškant vėlinimo laiko pagal tai, kur gaunama didžiausia koreliacijos reikšmė. Taip veikia paprasčiausias bendrosios kryžminės koreliacijos (angl. *Generalized Cross-Correlation*, GCC) garso šaltinio lokalizavimo metodas. Taikant fazės transformaciją (angl. *PHAse Transform*, PHAT), sudaromas plačiai žinomas GCC-PHAT garso šaltinio lokalizavimo metodas. Tobulinant metodus toliau, galima naudoti ne vieną, o daugiau mikrofonų porų, apskaičiuojant laiko skirtumus visoms poroms. Tam, kad būtų maksimaliai padidintas mikrofono gardelės jautrumas norima kryptimi, sukuriamas valdomas atsako galios garso šaltinio lokalizavimo metodas (angl. *Steered Response Power*, SRP). Analogiškai, pritaikius fazės transformaciją, gaunamas SRP-PHAT garso šaltinio lokalizavimo metodas.

Kita garso šaltinių lokalizavimo algoritmų klasė yra aukštos raiškos spektro analize pagrįsti algoritmai, tokie kaip MUSIC (angl. *MUltiple SIgnal Classification*) ir ESPRIT (angl. *Estimation of Signal Parameters via Totational Invariance*), kurie leidžia tiksliai lokalizuoti kelis garso šaltinius vienu metu, tačiau yra neatsparūs aplinkos triukšmams ir reverberacijai, o taip pat yra imlūs skaičiavimams.

Siekiant išvengti minėtų garso šaltinio lokalizavimo metodų trūkumų, pastaruoju metu tyrinėjami mokymu grįsti garso šaltinių lokalizavimo metodai. Tokie metodai yra pagrįsti sąryšio tarp akustinių požymių ir garso šaltinio vietos mokymu. Garso šaltinio lokalizavimo problema gali būti formuluojama kaip regresijos arba kaip klasifikavimo uždavinys. Regresijos atveju, algoritmo išėjime sukuriamos vertės, atitinkančios vieno ar kelių garso šaltinių koordinates. Tokio tipo garso šaltinio lokalizavimo metodų trūkumas yra tas, kad jie gali būti taikomi tik riboto ir fiksuoto šaltinių skaičiaus lokalizavimui. Kita mokymu grįstų garso šaltinio lokalizavimo metodų grupė yra klasifikatoriai, kurie klasifikuoja akustinius požymius į akustines klases, kurios atitinka erdvinius šaltinio parametrus: sklidimo kryptį arba koordinates.

2. Mokymu grįstų garso šaltinių lokalizavimo metodų teoriniai tyrimai

Ruošiant disertaciją, buvo atlikti šeši eksperimentiniai tyrimai: vieno garso šaltinio lokalizavimo taikant daugiasluoksnį perceptroną kaip požymį naudojant tarpimtuvinio signalų

lygio skirtumą, signalo analizės kadrų trukmės įtakos garso šaltinio lokalizavimo tikslumui, vieno garso šaltinio lokalizavimo taikant grafu reguliarizuotą neuronų tinklą ir pusiau prižiūrimą mokymo strategiją su ribotos apimties žymėtų duomenų rinkiniu, sąsūkos neuroniniu tinklu grįsto metodo daugelio garso šaltinių lokalizavimui dvimatėje erdvėje su koreliacijos dažnių juostose požymiu, sąsūkos neuroniniu tinklu grįsto metodo daugelio garso šaltinių lokalizavimui dvimatėje erdvėje su spektro fazės komponentės požymiu ir sąsūkos neuroniniu tinklu grįsto metodo daugelio garso šaltinių lokalizavimui trimatėje erdvėje su spektro fazės komponentės požymiu. Toliau glaustai pristatomi minėti tyrimai ir jų rezultatai.

Erdvėje sklindančios garso bangos frontas gali būti aproksimuojamas sferos paviršiumi. Sferos paviršiaus plotas yra kvadratu proporcingas jos spinduliui, taigi ir atstumui nuo garso šaltinio iki bangos fronto. Garso intensyvumas yra lygus garso galiai ploto vienetui. Imtuvui tolstant nuo garso šaltinio, garso intensyvumas (taip pat ir garso slėgis) mažėja kvadratine priklausomybe. Išnaudojant šią garso bangų sklidimo savybę, buvo pasiūlyta naudoti mikrofonų gardelės signalų kadrų galią kaip akustinį požymį. Akustinių požymių ryšys su šaltinio koordinatėmis gali būti išmoktas daugiasluoksnio perceptrono struktūros DNT. Lokalizavimo uždavinys šiuo atveju formuluojamas kaip regresijos problema. Naudojant plokščią apskritiminę mikrofonų gardelę su keturiais elementais, galima gauti keturis garso galios požymius vienam gardelės signalų analizės kadrui. Atstumai nuo garso šaltinio iki kiekvieno iš gardelės mikrofonų yra skirtingi, todėl skiriasi ir juos pasiekiančio garso signalo intensyvumas.

Tyrime naudotas DNT turi keturis įėjimus akustiniams požymiams, du paslėptuosius sluoksnius ir išėjimo sluoksnį su dviem neuronais. Išėjimo sluoksnyje norima gauti garso šaltinio dvimates koordinates. Du paslėptieji sluoksniai buvo pasirinkti todėl, kad su vienu paslėptuoju sluoksniu DNT negalėjo išmokti sąryšių tarp įėjimo požymių ir norimo atsako. Pasiūlyto metodo veikimas buvo išbandytas taikant imituotus mikrofonų gardelės signalus ir realius, mikrofono gardele įrašytus garso signalus. Tiek imituoti, tiek realūs gardelės signalų įrašai buvo surinkti garso šaltiniui esant viename iš dvimatės erdvinės gardelės taškų, kurios centre buvo mikrofonų gardelė, o dvimatės gardelės žingsnis buvo 1 m.

Eksperimentiniai tyrimai su realiais mikrofonų gardelių signalais (Löllmann *et al.* 2018) parodė, kad yra priklausomybė tarp garso šaltinio lokalizavimo tikslumo ir tokių signalo parametrų kaip: diskretizavimo dažnis, perteklinio diskretizavimo santykis, šnekos aptiktuvo tipas ir jo parametrai. Disertacijoje pateiktas Signalo kadrų atrinkimo įtaka garso šaltinio lokalizavimo tikslumui tyrimas parodo, kad norint padidinti garso šaltinio lokalizavimo metodo tikslumą, būtina iš garso signalo išskirti atkarpas, kuriose yra lokalizuojamo garso šaltinio signalas, atmetant tokias atkarpas, kuriose vyrauja aplinkos triukšmai.

Šnekos signalai pasižymi laike kintančiomis savybėmis. Kai kurios šnekos fonemos yra panašios į triukšmo signalą, kai tuo tarpu kitose signalas yra periodinis. Taip pat nepastovi yra ir šnekos signalo amplitudė. Tyrime buvo aiškinamasi, kaip kinta garso šaltinio lokalizavimo tikslumas parenkant iš mikrofono signalo tik kai kuriuos kadrus, kuriuose signalo ir triukšmo amplitudžių santykis viršija pasirinktą slenkstinę vertę. Spėjama, kad signalo kadrų atrinkimas gali padidinti šaltinio sklidimo krypties nustatymo tikslumą. Laikoma, kad kadrams, kuriuose vyrauja atsitiktinis signalas, neįmanoma nustatyti sklidimo laiko trukmės, o kadruose su periodiniu signalu – galima. Tyrime pasiūlytas kadrų atrinkimo kriterijus – signalo amplitudės ir vidutinės triukšmo signalo amplitudės santykis (angl.

Signal Amplitude to Mean Error Amplitude Ratio, SMEAR). Metodas buvo patikrintas naudojant LOCATA (Löllmann *et al.* 2018) šnekos duomenų rinkinį su trimis SMEAR slenkstinėmis vertėmis: 2, 3 ir 5. Šaltinio sklidimo krypties įverčiai palyginti su žinoma šaltinio sklidimo kryptimi. Tyrimas parodė, kad atrenkant signalo kadrus pagal SMEAR kriterijų, klaidingų šaltinio sklidimo krypties nustatymo atvejų sumažėja, tačiau didinant SMEAR slenksčio vertę, prarandamas didelis kiekis signalo kadrų.

Žymėtų garso signalų duomenų rinkinių sudarymas yra sudėtingas ir imlus laikui. Kita vertus, gana nesudėtinga gauti didelį nežymėtų pavyzdžių garso duomenų rinkinį. Todėl tokių duomenų panaudojimas apmokyti DNT taikant neprižiūrimo ar pusiau prižiūrimo (hibridinio) mokymo strategijas yra patrauklus ir vertas detalesnio tyrimo. Daroma prielaida, kad daugiamačiai akustiniai požymiai yra išsidėstę sumažinto matiškumo daugdaros (angl. *Manifold*), esančios daugiamatėje požymių erdvėje. Ši prielaida buvo pasiūlyta Laufer-Goldshtein *et al.* (2016).

Akustiniai požymiai, gauti erdvėje gretimoms garso šaltinio padėtims, yra artimi ir požymių erdvėje. Be to, daugiamačių požymių matiškumas gali būti sumažintas iki keleto matmenų taikant matiškumo mažinimo algoritmus, pvz., ISOMAP. Tuo atveju, jei akustinės scenos visi parametrai yra fiksuoti, ir keičiasi tik šaltinio padėtis, akustinių požymių koordinatės perskaičiuotoje (angl. *Embedded*) erdvėje atitinka šaltinių, kuriems veikiant buvo gauti akustiniai požymiai, fizines koordinates (tačiau šis sąryšis nėra tiesinis). Manoma, kad šis netiesinis sąryšis tarp akustinių požymių koordinačių sumažinto matiškumo erdvėje ir šaltinio koordinačių fizinėje erdvėje gali būti išmoktas DNT taikant hibridinę mokymo strategiją (grafu reguliarizuotas DNT – GRDNT).

Tyrime siūloma DNT įėjime kaip akustinius požymius naudoti SRP-PHAT erdvinius spektrus. SRP-PHAT spektrai gaunami mikrofonų gardelės signalų analizės kadrams. Siekiant mokyti GRDNT taikant hibridinę mokymo strategiją, siūloma taikyti reguliarizavimą grafu. Reguliarizavimo grafu esmė yra ta, kad GRDNT mokant su dideliu kiekiu nežymėtų pavyzdžių, tinklo išėjimas reguliarizuojamas (šio tyrimo atveju – į metrinę erdvę siekiant prognozuoti garso šaltinio koordinates) naudojant nedidelį kiekį žymėtų pavyzdžių.

Žymėti ir nežymėti pavyzdžiai yra susieti grafu. Grafo kraštinės atitinka atstumus tarp mokymo pavyzdžių akustinių požymiu sumažinto matiškumo erdvėje (t. y. ant daugdaros paviršiaus). Mokymo duomenų rinkinyje kiekvienam (žymėtam ar nežymėtam) pavyzdžiui yra nurodomi jo artimiausi kaimyniniai pavyzdžiai sumažinto matiškumo erdvėje (ant daugdaros paviršiaus).

Mokant GRDNT, kartu su pagrindiniu pavyzdžiu pateikiami ir kaimyniniai pavyzdžiai. Laikantis prielaidos, kad akustiniaims požymiams pasireiškia erdvinis glodumas, norima, kad GRDNT išėjimo vertės pagrindiniam požymiui būtų artimos išėjimo vertėms kaimyniniams požymiams. Taip siekiama, kad GRDNT išmoktų požymių erdvinį glodumą, tuo pačiu išmokdamas teisingą sąryšį tarp akustinių požymių ir garso šaltinio koordinačių. DNT mokomas naudojant specialią tikslo funkciją, sudarytą iš dviejų dalių: prižiūrimo mokymo klaidos, gaunamos žymėtiems pavyzdžiams, ir neprižiūrimo mokymo klaidos, gaunamos nežymėtiems pavyzdžiams:

$$L = \mu m \sum_{i \in N_b} (\hat{y}_i - y_i)^2 + (1 - \mu m) \sum_{i \in N_b} \sum_{j \in k_g} a_{ij} (\hat{y}_i - \hat{y}_j)^2, \tag{S2.1}$$

čia N_b – pavyzdžių skaičius vienoje mokymo imtyje, k_g – kaimyninių požymių skaičius, a_{ij} – kaimyninio požymio svorio (atstumo) koeficientas, y_i – norimas atsakas žymėtam požymiui, $\hat{y}_i$ – tinklo atsakas žymėtam požymiui, $\hat{y}_j$ – tinklo atsakas nežymėtam kaimyniniam požymiui, m – žymėto požymio indikatorius, μ – tikslo funkcijos komponenčių žymėtiems ir nežymėtiems požymiams santykis.

Sprendžiant kelių garso šaltinių lokalizavimo uždavinį taikant prižiūrimo mokymosi strategiją, pasiūlytas kelių garso šaltinių lokalizavimo dvimatėje erdvėje (skirtas ieškoti azimuto kampo ir aukščio) metodas, pagrįstas sąsūkos DNT taikymu su koreliacijos dažnių juostose (angl. *Corss-Correlation in Frequency Bands*, CCFB) požymiu ir dvimate DNT išėjimo struktūra. Pasiūlytas metodas kelių šaltinių lokalizavimo problemą sprendžia kaip klasifikavimo uždavinį, kai įėjimo požymiai yra priskiriami vienai ar kelioms erdvinėms klasėms. Erdvinės klasės siūlomo metodo atveju sudaro dvimatę DNT išėjimo sluoksnio struktūrą, kur kiekviena erdvinė klasė atitinka tam tikrą azimuto kampo ir aukščio derinį. DNT išėjimo dvimatės struktūros elementų vertės atitinka tikimybes, kad tam tikras įėjimo požymis priklauso tam tikrai išėjimo erdvinei klasei. Vienas įėjimo požymis gali būti priskiriamas kelioms klasėms – tai reiškia, kad mikrofonų gardelės signalo kadre, kuriam gautas įėjimo akustinis požymis, buvo daugiau nei vienas aktyvus garso šaltinis.

Pasiūlyta naudoti CCFB akustinį požymį kaip tinklo įėjimą. Šis požymis buvo pasiūlytas He *et al.* (2018a), tačiau autoriai šaltinį lokalizavo tik vienmatėje erdvėje (ieškodami tik azimuto kampo). Disertacijoje siūlomas metodo patobulinimas leidžia šaltinį lokalizuoti dvimatėje erdvėje. CCFB požymis gaunamas mikrofonų gardelės signalo kadrą filtruojant 16 juostų juostinių filtrų masyvu, ir kiekvienam gautam filtruotam signalui skaičiuojant koreliaciją tarp mikrofonų porų signalų. Naudojant keturių mikrofonų gardelę, gaunamos šešios mikrofonų poros, todėl ir CCFB požymis yra šešių kanalų. CCFB požymis yra apribojamas vėlinimo laiko iki ± 64. Pageidaujamas DNT atsakas formuojamas sukuriant nulių matricą, turinčią $360^\circ/Q_x$ elementų x ašyje ir $180^\circ/Q_y$ elementų y ašyje. Q yra DNT išėjimo sluoksnio kampinė raiška, nurodanti, kokį sklidimo krypties kampą atitinka vienas sluoksnio elementas. Toliau kiekvienam garso šaltiniui nulių matricoje pridedama Gauso funkcija su sklaidos parametru σ. Gauso funkcijos sklaida modeliuoja akustinių požymių erdvinį glodumą.

Požymių skaičiavimas yra imlus skaičiavimo ištekliams. Be to, apskaičiuojant požymį iš garso signalo analizės kadro, prarandama dalis jame esančios informacijos. Atsisakius CCFB požymio skaičiavimo, galima sumažinti metodui taikyti reikalingų didelių skaičiavimo išteklių. Chakrabarty, Habets (2019b) pasiūlė naudoti signalo kadro spektro fazės komponentę kaip sąsūkos DNT įėjimo požymį, tačiau apsiribojo kelių garso šaltinių lokalizavimu viename matmenyje (buvo skaičiuojamas azimuto kampas). Disertacijoje pasiūlytas metodo patobulinimas kelių garso šaltinių lokalizavimui dvimatėje erdvėje (skaičiuojant ne tik azimuto kampą, bet ir aukštį) naudojant mikrofonų gardelės signalų kadro spektro fazės komponentės požymį. Tyrimo metu naudojama keturių elementų tetraedrinė mikrofonų gardelė, ir 512 atskaitų signalo analizės kadras, todėl gaunamas spektro fazės požymis yra 4×257 elementų matrica. Naudojama ankstesniame skyriuje pasiūlyta sąsūkos DNT išėjimo sluoksnio dvimatė struktūra, atitinkanti garso šaltinio sklidimo krypties erdvines klases. Kaip ir ankstesniu atveju, DNT laukiamo atsako parametrai – dvimatės struktūros kampinė raiška Q ir Gauso funkcijos sklaida σ. DNT struktūra adaptuota iš Chakrabarty, Habets (2019b), naudojant pasiūlytą išėjimo sluoksnio dvimatę struktūrą.

Remiantis anksčiau pristatytais pasiūlytais metodais, buvo pasiūlytas papildomas metodo patobulinimas kelių garso šaltinių lokalizavimui trimatėje erdvėje taikant sąsūkos dirbtinį neuronų tinklą su spektro fazės komponentės įėjimo požymiu. Pasiūlyta DNT išėjimo sluoksnio trimatė struktūra, atitinkanti garso šaltinio sklidimo krypties erdvines klases stačiakampėje koordinačių sistemoje. Kiekvienas tokios struktūros elementas atitinka tam tikrą šaltinio vietą trimatėje erdvėje, o elemento išėjimo vertė atitinka tikimybę, kad įėjimo požymis priklauso jo reprezentuojamai erdvinei klasei, t. y. parodo, kad garso šaltinis yra aktyvus tam tikrame erdvės taške. Struktūros tūris atitinka pasirinktą tūrį patalpoje, kurioje yra ieškomi garso šaltiniai. Kaip ir anksčiau pristatyto metodo atveju, naudojamas mikrofonų gardelės signalų kadro spektro fazės komponento požymis. Naudota tinklo architektūra adaptuota iš anksčiau disertacijoje pristatyto metodo, tik pakeisti DNT elementų sluoksnyje skaičiai ir išėjimo sluoksnio struktūra. Pasiūlyti du būdai šaltinio koordinačių nustatymui iš DNT atsako. Vieno šaltinio atveju šaltinio koordinatės apskaičiuojamos pagal atsako maksimumo koordinates. Taip pat pasiūlytas originalus būdas, leidžiantis DNT atsake aptikti vieną ar daugiau šaltinių ir nustatyti jų koordinates. Naujas būdas pagrįstas slenksčio pritaikymu DNT atsakui pagal elementų vertes, atsake paliekant tik elementus su verte, didesne už slenkstinę. Atsake likę elementai grupuojami naudojant „k-means“ grupavimo algoritmą ir nustatant grupių centrų koordinates, kurios laikomos lokalizuotų garso šaltinių koordinatėmis.

3. Mokymu grįstų garso šaltinių lokalizavimo metodų eksperimentiniai tyrimai

Siekiant patikrinti ankstesniame skyriuje pristatyto metodo vienam garso šaltiniui lokalizuoti naudojant MLP su garso signalų kadrų galios požymiais veikimą, buvo atlikti eksperimentianiai tyrimai. DNT architektūra tyrimo metu buvo modifikuojama, siekiant išsiaiškinti, koks yra geriausias neuronų skaičius kiekviename iš paslėptųjų sluoksnių. Kiekvienas iš paslėptųjų sluoksnių tyrimo metu turėjo $N \in [1, 2, 5, 10]$ neuronų. Tyrimo rezultatai pateikti S3.1 lentelėje.

S3.1 lentelė. DNT grįsto vieno garso šaltinio lokalizavimo metodo tyrimo rezultatai

	Kompiuterinė imitacija			Eksperimentas su realiais įrašais		
h_1 , h_2	$E_{\text{rel}}^{h_{1,2}}$, %	$E_{\text{ang}}^{h_{1,2}}$, °	$E_{\text{dist}}^{h_{1,2}}$, m	$E_{\text{rel}}^{h_{1,2}}$, %	$E_{\text{ang}}^{h_{1,2}}$, °	$E_{\text{dist}}^{h_{1,2}}$, m
1, 1	5,244	2,710	2,209	5,219	1,871	1,874
1, 2	5,095	11,522	2,000	7,262	5,441	1,537
1, 5	5,908	8,216	1,756	22,686	3,877	0,770
1, 10	5,312	2,416	2,027	5,556	2,936	1,432
2, 1	5,963	7,105	1,736	9,760	1,050	0,556
2, 2	4,008	8,524	1,958	**1,884**	5,548	1,355
2, 5	6,069	7,292	1,799	7,896	**0,097**	**0,410**
2, 10	**3,993**	4,421	1,753	10,949	1,614	1,411
5, 1	4,592	14,644	2,089	6,887	6,378	1,453
5, 2	4,566	**0,120**	1,759	6,499	4,911	0,729
5, 5	6,190	2,109	**1,577**	4,417	3,329	0,389
5, 10	4,074	5,106	1,600	6,068	5,102	0,733
10, 1	8,000	1,483	1,752	11,319	3,014	0,639
10, 2	6,003	3,238	1,791	4,770	0,984	0,800
10, 5	6,353	4,149	1,913	10,261	7,815	0,446
10, 10	6,625	0,690	1,836	14,904	4,650	0,839

Tyrimo metu buvo nustatyta, kad lokalizuojant garso šaltinį su imituotais garso signalais, šaltinio lokalizavimo vidutinė paklaida buvo $1{,}58\,\mathrm{m}$ kai DNT turėjo $h_1 = 5$ ir $h_2 = 5$ neuronų paslėptuosiuose sluoksniuose. Atliekant eksperimentus su realiais garso įrašais, mažiausia šaltinio lokalizavimo vidutinė paklaida buvo $0{,}41\,\mathrm{m}$, kai $h_1 = 2$, $h_2 = 5$.

Ruošiant disertaciją, buvo surinktas duomenų rinkinys keturiems akustiniams scenarijams (dvi skirtingų apertūrų tetraedrinės mikrofonų gardelės; vienas arba du garso šaltiniai). Mikrofonų gardelių signalai buvo gauti garso šaltiniui esant vienoje iš 10 žymėtų šaltinio padėčių erdvėje. Dviejų šaltinių atveju, šaltinių padėčių deriniai buvo pasirinkti iš tų pačių 10 padėčių. Buvo naudojamos dvi tetraedrinės mikrofonų gardelės, kurių kraštinės ilgis buvo 30 cm (ARRAY30) ir 60 cm (ARRAY30).

Visiems garso įrašams gauti buvo naudojama Tascam US-20x20 USB garso sąsaja. Visi įrašai buvo atlikti naudojant f_s = 44 100 Hz diskretizavimo dažnį ir Q = 16 bitų kvantavimo raišką. Visi erdviniai matavimai buvo atlikti rankiniu būdu, naudojant matavimo juostą, kurios tikslumas $\pm$0,005 m. Duomenų rinkinį sudaro garso bylos „.wav“ formatu (4 kanalų garso įrašai mikrofono masyvo signalams ir monofoniniai garso failai atitinkamam šaltinio signalui), patalpos impulsinių atsakų matavimo duomenys su „Matlab“ suderinamu formatu („.mat“) ir „.wav“ formatu bei skaičiuoklės failas su atitinkamu informacija apie garso šaltinių, mikrofonų padėtis, patalpos geometriją bei garso šaltinių signalus. Duomenų rinkinys yra atviros prieigos ir paskelbtas publikacijoje recenzuojamame žurnale.

Duomenų rinkinys buvo surinktas stačiakampio gretasienio formos patalpoje Vilniaus Gedimino technikos universitete (VILNIUS TECH), „LinkMenų fabrike“. Patalpos matmenys buvo $[5{,}4 \times 5{,}86 \times 2{,}64]$ m. Trys iš keturių kambario sienų buvo pagamintos iš dažytos mūro, o ketvirtoji siena buvo gipso. Patalpos tūris buvo $89{,}869\,\mathrm{m}^3$ o bendras patalpą ribojančių paviršių plotas buvo $145{,}048\,\mathrm{m}^2$.

Žymėtos garso šaltinių padėtys buvo tolygiai paskirstytos patalpoje. Šešių kambario impulsų atsakų rinkinys buvo išmatuotas naudojant tris skirtingus šaltinio-mikrofono pozicijų derinius, naudojant du impulsinio atsako skaičiavimo metodus: maksimalios trukmės sekas (angl. *Maximum Length Sequence*, MLS) ir kintamo dažnio sinusoides (angl. *Swept Sine*). Reverberacijos trukmė T_{60} buvo apskaičiuota pagal išmatuotus impulsinius atsakus, naudojant Šrioderio (angl. *Schroeder*) metodą, o vidutinė reverberacijos trukmė T_{60} buvo $0{,}552\,\mathrm{s}$. Vidutinis paviršiaus absorbcijos koeficientas buvo apskaičiuotas iš reverberacijos trukmės ir patalpos geometrijos ir buvo $a = 0{,}206$. Patalpos Šrioderio dažnis buvo $F_c = 156{,}76\,\mathrm{Hz}$.

Šaltinių signalai buvo atkuriami naudojant kilnojamuis „JBL GO“ ir „Yamaha MSP3“ garsiakalbius. Šnekos signalai buvo paimti iš „AMI Corpus“ (Carletta *et al.* 2006) duomenų rinkinio.

Buvo atliktas kompiuterinis virtualios patalpos modeliavimas su ta pačia geometrija ir akustiniais parametrais, kaip ir realios patalpos. Palyginus rezultatus buvo nustatyta, kad realioje patalpoje išmatuotų ir imituotų patalpos impulsinių atsakų spektrai labai skiriasi tiek žemo, tiek aukšto dažnio diapazonuose, o modeliavimas yra gana tikslus tik dažnių diapazone nuo 60 Hz iki 500 Hz.

Taigi, kuriamo garso šaltinio lokalizavimo metodo ar algoritmo veikimas turi būti vertinamas pagal realių mikrofonų gardelių garso įrašus, nes imituoti garso signalai gali tiksliai neatspindėti realių situacijų.

Ankstesniame skyriuje pristatyto GRDNT grįsto metodo veikimas tiriamas naudojant realių mikrofonų gardelių signalų duomenų rinkinį, kuris buvo surinktas naudojant dvi apskritimines mikrofonų gardeles patalpoje, kurios apytiksliai matmenys yra $5\,\mathrm{m} \times 5\,\mathrm{m}$, aukštis $3{,}75\,\mathrm{m}$, o reverberacijos trukmė $T_{60} = 0{,}311\,\mathrm{s}$.

Mikrofonų signalų SRP-PHAT spektrai naudojami kaip įėjimo požymiai (720 elementų vektoriai) lokalizuojant vieną garso signalą dvimatėje erdvėje (GRDNT išėjimo sluoksnyje – du neuronai).

Pasiūlyta taikyti papildomą akustinių požymių atranką pagal kadro efektinę vertę ir keteros faktorių, siekiant atrinkti akustinius požymius, kuriose yra stiprus garso signalas.

Pasiūlyto metodo veikimas lyginamas su alternatyviais garso šaltinio lokalizavimo metodais: geometriniu šaltinio padėties nustatymu remiantis sklidimo kryptimis, nustatytomis iš SRP-PHAT spektrų, o taip pat šaltinio padėties nustatymu iš SRP-PHAT dvimačio žemėlapio (ieškant globalaus ar lokalaus maksimumo koordinačių). Tyrimo rezulatai apibendrinti S3.2 ir S3.3 lentelėse (VP - vidutinė paklaida, SN - standartinis nuokrypis).

S3.2 lentelė. Šaltinio lokalizavimo vidutinės absoliučios klaidos visiems taikytiems parametrų deriniams

Metodas	VP, m	SN, m	Pagerinimas, %
Geometrinis	1,95	1,62	80,1
SRP-PHAT žemėlapio global. maks. koord.	1,13	0,66	4,9
SRP-PHAT žemėlapio lokalaus maks. koord.	1,12	0,69	3,5
GRDNT	**1,08**	0,51	–

S3.3 lentelė. Šaltinio lokalizavimo vidutinės paklaidos geriausiai veikiančiam GRDNT parametrų deriniui

Metodas	VP, m	SN, m	Pagerinimas, %
Geometrinis	3,06	4,50	68,57
SRP-PHAT žemėlapio global. maks. koord.	1,17	0,74	17,97
SRP-PHAT žemėlapio lokalaus maks. koord.	1,14	0,75	15,94
GRDNT	**0,96**	0,62	–

Kelių garso šaltinių lokalizavimo dvimatėje erdvėje metodo, grįsto konvoliuciniu DNT su CCFB požymiais įėjime, veikimas patikrintas naudojant dvi DNT architektūras ir tris duomenų rinkinius, kuriuose skyrėsi aktyvių šaltinių skaičius (1, 2) ir Gauso funkcijos sklaida (1, 2). Eksperimentai atlikti su $Q = 10°$. Eksperimentų rezultatai pateikti S3.4 lentelėje (VP – vidutinė paklaida, SN – standartinis nuokrypis).

Iš eksperimentinio tyrimo rezultatų matyti, kad taikant pasiūlytą metodą garso šaltinio sklidimo krypties nustatymo vidutinė paklaida mažiausia buvo gauta vienam aktyviam

garso šaltiniui ir buvo lygi $22{,}67°$, o blogiausia buvo lygi $29{,}97°$. Dviejų aktyvių garso šaltinių atveju geriausia vidutinė paklaida buvo gauta lygi $25{,}22°$, o blogiausia – lygi $31{,}44°$.

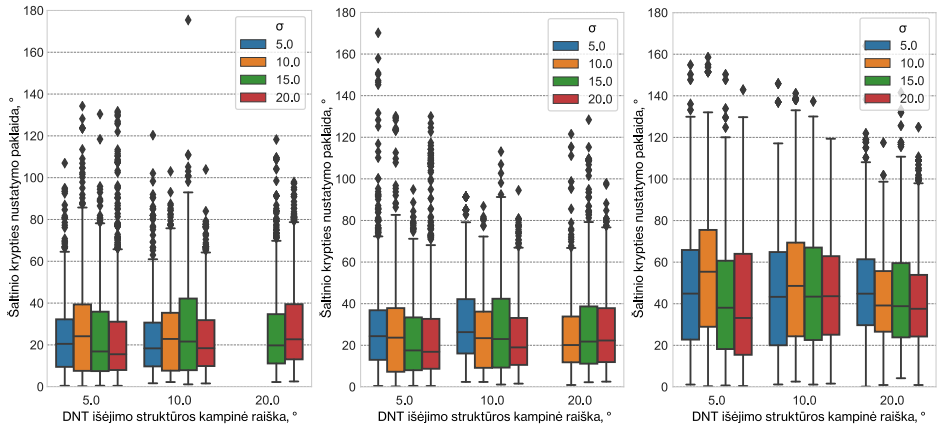
S3.4 lentelė. Vieno ir dviejų garso šaltinio lokalizavimo naudojant sąsūkos DNT su CCFB įėjimo požymiu eksperimentinio tyrimo rezultatai

Šaltinių sk.	σ	Mokymo žingsnis	Tinklo architektūra	VP, °	SN, °
1	2	0,001	CONV-WE-CCFB	29,97	57,64
1	2	0,001	CONV-CCFB-DOA	**22,67**	**48,02**
2	1	0,001	CONV-WE-CCFB	25,61	**19,58**
2	1	0,001	CONV-CCFB-DOA	27,17	36,75
2	2	0,001	CONV-CCFB-DOA	31,44	34,08
2	1	0,01	CONV-WE-CCFB	28,30	34,98
2	1	0,01	CONV-CCFB-DOA	**25,22**	29,29

Kelių garso šaltinių lokalizavimo dvimatėje erdvėje metodo, grįsto konvoliuciniu DNT su spektro fazės komponentės požymiais įėjime, veikimas eksperimentiškai ištirtas naudojant imitacinius triukšmo ir šnekos signalų duomenų rinkinius. DNT mokymui naudojamas 100 000 pavyzdžių duomenų rinkinys su spektro fazės komponentės požymiais, kai imitacinėje patalpoje yra trys triukšmo šaltiniai. DNT atsakas formuotas su $Q \in [5; 10; 20]°$ ir $\sigma \in [5; 10; 15; 20]°$. DNT mokomas su nepakeistais spektro fazės komponentės požymiais, ir su požymiais, kurių elementai buvo atsitiktine tvarka sumaišyti laiko ir dažnių srityse, kaip tai pasiūlė atlikti Chakrabarty, Habets tam, kad užtikrinti signalų ortogonalumą. Šaltinio sklidimo kryptis iš DNT atsako nustatoma priklausomai nuo garso šaltinių skaičiaus: vieno šaltinio atveju – randamos maksimumo koordinatės; dviejų šaltinių atveju – randamos lokalių maksimumų koordinatės. Pasiūlyto metodo veikimas testuotas su spektro fazės komponentės požymiais, gautais kai imitacinėje patalpoje yra trys triukšmo arba šnekos signalai. Gauti rezultatai lyginami su atskaitos garso šaltinio lokalizavimo metodu – SRP-PHAT, kuriuo gaunamas taip pat dvimatis atsakas, nurodantis tikėtino garso signalo, pasiekiančio mikrofonų gardelę tam tikra kryptimi, galią. Įvertinamos vidutinės šaltinių sklidimo krypties nustatymo paklaidos. Esant daugiau, nei vienam šaltiniui, lokalizavimo paklaidos skaičiuojamos tarp visų laukiamame atsake ir prognozuotame atsake aptiktų garso šaltinių sklidimo krypčių ir pasirenkama N_S mažiausių paklaidų, kur N_S – lokalizuojamų šaltinių skaičius. Tyrimo rezultatai pateikti S3.1 paveiksle.

Pasiūlytas metodas leidžia nustatyti trijų šnekos šaltinių dvimatę sklidimo kryptį su $16°$ vidutine paklaida, kai $Q = 5°$ ir $\sigma = 20°$; tai yra iki $36\,\%$ mažesnė vidutinė paklaida, nei naudojant alternatyvųjį SRP-PHAT metodą. Pasiūlytas metodas leido pasieki iki $29\,\%$ mažesnę šaltinių krypties nustatymo paklaidą nei SRP-PHAT metodas esant bet kurioms norimo atsako Q ir σ vertėms.

Sąsūkos DNT su spektro fazės komponentės požymiais grįsto kelių garso šaltinių lokazliavimo trimatėje erdvėje metodo veikimas eksperimentiškai patikrintas naudojant akustinius požymius, gautus iš imituotų tetraedrinių mikrofonų gardelių signalų. Tinklo veikimas patikrintas esant išėjimo struktūros raiškai $Q \in [0,25; 0,5; 1]$ m ir Gauso funkcijos sklaidai $\sigma \in [0,25; 0,5; 1]$ m. Šaltinių padėtys pasirinktos atsitiktinai $5{,}4\,\mathrm{m}$, $5{,}86\,\mathrm{m}$ and $2{,}84\,\mathrm{m}$ patalpos tūryje.



S3.1 pav. Trijų šnekos šaltinių sklidimo krypties nustatymo paklaidos; a) pasiūlyto metodo, mokyto su neapdorotais spektro fazės komponentės požymiais; b) pasiūlyto metodo, mokyto su laiko ir dažnio ašyse sumaišytais spektro fazės komponentės požymiais; c) SRP-PHAT metodo

Vieno ir dviejų garso šaltinių lokalizavimo vidutinės paklaidos (VP) prie anksčiau minėtų Q ir σ verčių, naudojant paminėtus koordinačių nustatymo iš DNT atsako metodus, pateiktos S3.5 lentelėje.

S3.5 lentelė. Vieno ir dviejų garso šaltinių lokalizavimo vidutinės paklaidos (VP) esant skirtingoms Q ir σ vertėms ir koordinačių nustatymo iš DNT atsako metodams

		Atsako max. koordinatės		k-means grupavimas			
		1 šaltinis				2 šaltiniai	
		Triukšmas	Šneka	Triukšmas	Šneka	Triukšmas	Šneka
Q, m	σ, m	VP, m	VP, m	VP, m	VP, m	VP, m	VP, m
0,25	0,25	2,51	2,60	0,79	0,94	2,74	2,74
0,25	0,50	1,26	1,39	**0,62**	**0,76**	1,10	1,09
0,25	1,00	**0,99**	**1,10**	0,81	0,91	1,18	1,17
0,50	0,25	2,18	2,32	0,67	0,86	1,19	1,17
0,50	0,50	2,29	2,35	0,69	0,82	**1,08**	**1,08**
0,50	1,00	1,05	1,14	0,84	0,94	1,18	1,18
1,00	0,25	2,73	2,73	0,97	1,10	1,41	1,40
1,00	0,50	1,92	2,00	0,81	0,91	1,17	1,16
1,00	1,00	1,11	1,22	0,89	0,99	1,20	1,20

Apibendrinant gautus rezultatus, daroma išvada, kad įmanoma lokalizuoti vieną ir du garso šaltinius trimatėje erdvėje taikant sąsūkos DNT su spektro fazės komponentės įėjimo požymį ir trimatę išėjimo sluoksnio struktūrą. Naudojant grupavimu grįstą garso šaltinių koordinačių nustatymo iš DNT atsako metodą galima lokalizuoti vieną garso šaltinį tri-

matėje erdvėje su vidutine lokalizavimo paklaida lygia 0,62 m triukšmo šaltiniui, 0,76 m šnekos šaltiniui, kai tinklas mokomas naudojant Q = 0,25 m ir σ = 0,5 m ir galima lokalizuoti du triukšmo arba šnekos šaltinius trimatėje erdvėje su 1,08 m vidutine lokalizavimo paklaida, kai tinklas mokomas naudojant Q = 0,5 m ir σ = 0,5 m.

Bendrosios išvados

Disertacijoje patvirtintos iškeltos hipotezės. Taip pat disertacijoje pasiūlyti trys mokymu grįsti garso šaltinių lokalizavimo metodai.

1. Naudojant hibridinį mokymą, grafu reguliarizuotu dirbtiniu neuronų tinklu su SRP-PHAT požymiais įėjime galima pasiekti iki penkių kartų mažesnę garso šaltinio lokalizavimo paklaidą nei naudojant geometrinį šaltinio vietos nustatymo iš SRP-PHAT duomenų metodą.

2. Naudojant koreliacijos dažnių juostose požymius sąsūkos dirbtinių neuronų tinklu dviejų garso šaltinių vidutinė lokalizavimo paklaida išlieka ne mažesnė nei 25 laipsniai.

3. Mikrofonų gardelių signalų spektrų fazės komponentės yra tinkamos naudoti kaip požymis dirbtiniais neuronų tinklais grįstiems garso šaltinio lokalizavimo dvimatėje ir trimatėje erdvėje metodams.

4. Taikant sąsūkos DNT su spektro fazės komponentės požymiais įėjime, trijų garso šaltinių dvimatės sklidimo krypties nustatymo vidutinė paklaida gali būti 16° su DNT išėjimo struktūros raiška $Q = 5^\circ$ ir Gauso funkcijos sklaida $\sigma = 20^\circ$ ($36\,\%$ mažesnė vidutinė paklaida nei taikant SRP-PHAT su tais pačiais parametrais)

5. Įmanoma lokalizuoti vieną ir du garso šaltinius trimatėje erdvėje taikant sąsūkos DNT su spektro fazės komponentės įėjimo požymiu ir trimate išėjimo sluoksnio struktūra. Grupavimu grįstas šaltinio koordinačių nustatymo iš dirbtinio neuronų tinklo atsako metodas, lyginant su atsako maksimumo koordinačių nustatymu grįstu metodu, leidžia sumažinti vieno garso signalo lokalizavimo trimatėje erdvėje vidutinę paklaidą mažiausiai 31 %.

Annexes[1]

Annex A. Declaration of Academic Integrity
Annex B. The Co-authors' Agreements to Present Publications Material in the Dissertation
Annex C. The Copies of Scientific Publications by the Author on the Topic of the Dissertation

[1] The annexes are supplied in the enclosed compact disc.

Saulius SAKAVIČIUS

IMPROVEMENT OF LEARNING-BASED METHODS FOR LOCALIZATION OF MULTIPLE SOUND SOURCES

Doctoral Dissertation

Technological Sciences,
Electrical and Electronic Engineering (T 001)

MOKYMU GRĮSTŲ METODŲ KELIEMS GARSO ŠALTINIAMS LOKALIZUOTI TOBULINIMAS

Daktaro disertacija

Technologijos mokslai,
elektros ir elektronikos inžinerija (T 001)

2021 11 09. 14,0 sp. l. Tiražas 20 egz.
Leidinio el. versija *https://doi.org/10.20334/2021-050-M*
Vilniaus Gedimino technikos universitetas
Saulėtekio al. 11, 10223 Vilnius,
Spausdino BĮ UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius