Geometry-aware sound source localization using neural networks

by
Eric Grinstein
M.Eng

A Thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy
of
Imperial College London

Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London
2025

Abstract

Sound Source Localization (SSL) is the topic within acoustic signal processing which studies methods for the estimation of the position of one or more active sound sources in space, such as human talkers, using signals captured by one or more microphone arrays. It has many applications, including robot orientation, speech enhancement and diarization. Although signal processing-based algorithms have been the standard choice for SSL over past decades, deep neural networks have recently achieved state-of-the-art performance for this task.

A drawback of most deep learning-based SSL methods consists of requiring the training and testing microphone and room geometry to be matched, restricting practical applications of available models. This is particularly relevant when using Distributed Microphone Arrays (DMAs), whose positions are usually set arbitrarily and may change with time. Flexibility to microphone geometry is also desirable for companies maintaining multiple types of microphone arrays in their line of products, and smaller companies or practicioners who wish to apply freely available pre-trained, off-the-shelf SSL models to their applications.

The main contributions of this thesis are the creation of a novel class of neural network models for the tasks of Positional Sound Source Localization (PSSL) and Direction-of-Arrival (DOA) estimation, named Neural-SRP. The method combines concepts from graph neural networks as well as from the classical Steered Response Power (SRP) localization method. Unlike current state of the art networks for SSL, the Neural-SRP method is able to function on microphone arrays and rooms of arbitrary geometry, while maintaining or improving localization performance.

Contents

Abstract **2**

Contents **3**

Statement of Originality **7**

Copyright Declaration **8**

Acknowledgment **9**

List of Figures **10**

List of Tables **13**

List of Abbreviations **14**

List of Symbols **17**

Chapter 1. Introduction **19**

1.1 Motivation and objectives 19
1.2 Publications 20
1.3 Thesis outline 21
1.4 Problem statement 21
1.5 Signal Model 23
1.6 Evaluation metrics 25
1.7 Reverberation, RIRs and simulation 26
1.8 Applications 27

Chapter 2. Sound Source Localization **29**

2.1 Delay-based methods 31
2.1.1 Measuring TDOA: Cross-correlation and GCC-PHAT 31
2.1.2 Time-domain Steered Response Power (SRP) formulation 34
2.1.3 Frequency-domain SRP formulation 35

2.1.4 Grid construction and search . . . 36
2.1.5 SRP complexity analysis . . . 37
2.1.6 Two-step SSL methods . . . 39
2.2 Energy-based methods . . . 41
2.3 Neural-based methods . . . 43
2.4 Subspace methods . . . 45
2.4.1 MUSIC . . . 45
2.5 Multi-source SRP approaches . . . 47
2.5.1 Modified SRP computation . . . 48
2.5.2 Source cancellation . . . 48
2.5.3 Grid refinement . . . 50
2.5.4 Clustering and distance analysis . . . 51
2.5.5 Sparsity assumptions . . . 52
2.6 SRP: practical considerations . . . 54
2.6.1 Tracking moving sources . . . 54
2.6.2 Directional sources and microphones . . . 55
2.6.3 Comparing SRP to other approaches . . . 56
2.6.4 Analyses of SRP . . . 58

Chapter 3. Dual-Input Neural Networks for SSL 60
3.1 Introduction . . . 60
3.2 Method . . . 62
3.2.1 Scope of this work . . . 62
3.2.2 Proposed method: Dual input neural network . . . 63
3.2.3 DI-NN-Embedding . . . 65
3.3 Experimentation . . . 65
3.3.1 Simulated anechoic rooms . . . 67
3.3.2 Simulated reverberant rooms . . . 68
3.3.3 Real recordings . . . 68
3.3.4 Metadata sensitivity study . . . 69
3.3.5 Metadata relevance study . . . 69
3.4 Results and discussions . . . 69
3.4.1 Results . . . 69
3.4.2 Limitations and extensions . . . 72
3.5 Conclusion . . . 73

Chapter 4. Graph neural networks for sound source localization 74
4.1 Introduction . . . 74
4.2 Problem statement . . . 76

4.3 Related work . . . 76
4.3.1 Classical SSL methods . . . 76
4.3.2 Neural network methods for DMA signal processing . . . 77
4.3.3 Relation Networks . . . 78
4.4 Method . . . 78
4.5 Experimentation . . . 80
4.5.1 Dataset . . . 80
4.5.2 Method hyperparameters . . . 81
4.6 Results . . . 82
4.7 Conclusion and future work . . . 82

Chapter 5. The Neural-SRP method for Sound Source Localization 84
5.1 Introduction . . . 84
5.2 Neural-SRP . . . 87
5.2.1 Input Feature Set . . . 87
5.2.2 Architecture . . . 88
5.2.3 Training for DOA estimation . . . 89
5.2.4 Training for PSSL . . . 90
5.3 Experimentation for DOA estimation . . . 92
5.3.1 Datasets . . . 92
5.3.2 Experiment 1: spatially white noise . . . 95
5.3.3 Experiment 2: directional noise . . . 96
5.3.4 Experiment 3: testing on an unseen geometry . . . 96
5.3.5 Experiment 4: multi-source tracking . . . 97
5.3.6 Complexity comparison . . . 98
5.4 Experimentation for PSSL . . . 98
5.4.1 Datasets . . . 98
5.4.2 Methods and baselines . . . 99
5.4.3 Experiment details . . . 100
5.5 Discussion and analysis . . . 101
5.5.1 DOA estimation experiments . . . 101
5.5.2 PSSL experiments . . . 103
5.6 Conclusions . . . 104

Chapter 6. A Generalized framework and review for the SRP method 105
6.1 Introduction . . . 105
6.2 Reducing SRP's complexity and computational time . . . 107
6.2.1 Coarse grids and Volumetric-SRP . . . 107
6.2.2 Iterative grid refinement . . . 109

6.2.3 Grids based on prior location estimates . . . 111
6.2.4 Incorporation of prior scene information . . . 112
6.2.5 Paralellization . . . 114
6.2.6 Other approaches . . . 115
6.3 Increasing robustness . . . 115
6.3.1 Modified GCC-PHAT functions . . . 115
6.3.2 Improving combination . . . 118
6.3.3 Pre/Post-processing . . . 118
6.3.4 Neural approaches . . . 119
6.3.5 Other approaches . . . 120
6.4 X-SRP . . . 122
6.5 Conclusion . . . 125

Chapter 7. On the use of complex-valued neural networks for SSL 126
7.1 Introduction . . . 126
7.2 Complex-valued CRNN . . . 128
7.3 Experimentation . . . 131
7.3.1 Dataset . . . 131
7.3.2 Neural network training and evaluation . . . 132
7.4 Results and discussion . . . 132
7.5 Conclusion . . . 134

Chapter 8. Conclusion 135

Appendix A. Energy-based SSL: proof of the locus of source candidates 159

Statement of Originality

I declare that this thesis has been solely composed and written by myself, Eric Zajler Grinstein, and presents my own original work unless otherwise stated. The main contribution of this thesis is the creation of Neural-SRP, a neural network model for the task of SSL. Neural-SRP improves the state-of-the-art of neural SSL in quantitative terms, by increasing localization performance in comparison to the competitive baselines, and qualitative, by allowing neural methods to function on any type of microphone array, unlike most past state-of-the-art methods. Neural-SRP is also the first application of Graph Neural Networks (GNNs) to the task of SSL, and is the first neural architecture to allow the introduction of relevant metadata such as the microphone geometry as a secondary input to the microphone signals into the neural network. Finally, a secondary contribution is a comprehensive review of the SRP method for SSL. These claims are substantiated by several conference and journal papers published in reputable international conferences, which are enumerated for convenience in Sec. 1.2.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgment

It has been an honor and a pleasure being supervised by Patrick Naylor and Mike Brookes, who have always given me all the technical support I needed, and also created a lovely environment in the SAP lab. I am grateful for the fruitful collaboration with my external supervisors Toon van Waterschoot and Christopher Hicks. I thank the members of the Speech and Audio Processing (SAP) lab at Imperial, as well as the members of the SOUNDS consortium for their partnership and friendship.

Outside academia, I thank my parents Ivo and Silvia, my brothers Alex and Marcela, and my grandmother Anitta for giving me the best kick-off in life: a loving family and quality education. I also remember family members who are no longer with us, Moisés, Ruth and Clara. Their memories bring me joy. I also thank my friends for the support during and before the PhD. Finally, I thank my wife Monique, you are the light of my life.

List of Figures

1.1 Depiction of the SSL problem, where source signals s_n propagate from positions $\mathbf{u}_n$ to be received at each microphone as signals x_m, which is affected by noise and reverberation. 23

2.1 Hyperbola branch of points with the same TDOA as a source located at $\mathbf{u}$ with respect to microphone positions $\mathbf{v}_1$ and $\mathbf{v}_2$. The axes represent the horizontal directions, in meters. 31

2.2 Example comparison between the normalized temporal cross-correlation and GCC-PHAT for a scenario containing two closely spaced microphones and a speech source producing a TDOA of -2 ms. 33

2.3 Pairwise SRP maps 33

2.4 Global SRP maps, obtained by summing the maps shown in Fig. 2.3 34

2.5 Example of an SRP map for the task of 2D DOA estimation. The true source location is located below the triangle. Note the likelihood is composed by a sum of correlation values and is therefore dimensionless. 37

2.6 Pairwise TDOA-based Least-squares maps 39

2.7 Global TDOA-based Least-squares maps, obtained by summing the maps shown in Fig. 2.6 40

2.8 Pairwise Energy-ratio Least-squares maps 41

2.9 Global Energy-ratio Least-squares maps, obtained by summing the maps shown in Fig. 2.8 42

2.10 Comparison of the Energy- and TDOA-based LS methods and SRP on a reverberant environment 43

2.11 Representation of the de-emphasis procedure described by Brutti et al. [99]. 49

3.1 Overview of the Dual-Input Neural Network (DI-NN) approach. Note that the Metadata embedding network is an optional block. 61
3.2 Detailed DI-NN architecture for the task of PSSL. 64
3.3 Experimental setup. (a) For the anechoic and reverberant simulations, each of the four microphones m_i is placed on a random point along the the coloured arrows, while the source s is randomly placed on a point within the rectangle defined by them. (b) The sampling procedure for Sec. 3.3.3, where positions of the microphones and source are randomly drawn from each differently coloured set of points. 67
3.4 (a) Mean localization error for DI-NNs and baselines on different datasets. (b) Normalized histogram comparison between the DI-NN and the CRNN baseline on the recorded dataset. (c) Cumulative version of (b). 70

4.1 (a): Example of a graph of distributed microphones. (b): Representation of the GNN-SLF model for three microphones. The computation of the heatmaps is described in Sec. 4.4. $\mathcal{F}$ represents the relation function which is computed for each pair of microphone features. 75
4.2 Localization error for the proposed methods and baselines. 82

5.1 Example of Neural-SRP's output when tracking two moving sources. The panel shows the target and predicted azimuth and elevations. 85
5.2 Neural-SRP network architecture. Left, green: pairwise network $\mathcal{P}$. Right, blue: Global decoder $\mathcal{D}$, exemplified for a 3-microphone input. Symbol "&" represents concatenation. 87
5.3 Example of the hyperbolic grid (bottom), used for training Neural-SRP, and the alternative Gaussian grid (top). 90
5.4 Detailed view of Neural-SRP architecture, where the numbers show the output dimension of each layer. The dotted line separates the pairwise network $\mathcal{P}$ from the global decoder $\mathcal{D}$, which receives the sum of pairwise features as its input. The input layer consists of T frames of 64 central Generalized Cross-Correlation with Phase Transform (GCC-PHAT) bins each. The mic. coords. layer is of shape $(T, 6)$ where the three coordinates for each of the microphone in the pair are replicated for all frames. 94

5.5 Localization error comparison between Neural-SRP, Cross3D and SRP for increasing levels of reverberation and Signal-to-Noise Ratio (SNR). The curves were smoothed using cubic interpolation. 102
5.6 Neural-SRP and classical SRP output for real recorded signals. 103

6.1 Comparison between SRP maps generated with (bottom) and without (top) volumetric techniques. 108
6.2 Low-pass version of the frequency-domain SRP, where only frequencies up to 200 Hz are considered. 111
6.3 Iterative Region contraction procedure, where different colours represent search regions and grids of points related to iterations i. The true source location is represented by the black star. 112
6.4 Neural-SRP+ (see chapter 5) conventional SRP map in a highly reverberant room. The source position is shown with a cross and the microphone positions with circles. 121
6.5 Flowchart of the generalized SRP algorithm. Parallelograms represent input data, rectangles represent functions, diamonds represent decisions and ellipses represent terminal states. 123

7.1 Example of features generated at the output of the neural network's convolutional layers. 127
7.2 Architecture of the proposed Convolutional Recurrent Neural Network (CRNN). 128
7.3 Representation of different types of convolutional kernels used on a multichannel spectrogram containing T time frames, F frequency bins and C channels. 130
7.4 Top and middle: training and validation errors at the end of every epoch for the proposed complex architecture as well as for the real baseline. Bottom: error histogram for the test set. 133

List of Tables

3.1 Hyperparameters 66
3.2 Metadata sensitivity analysis 71
3.3 Metadata relevance analysis 72

5.1 Functional comparison of the proposed model and baselines. 'Universal' refers to the method's capacity of working on any microphone array geometry. 86
5.2 Parameter ranges for simulated datasets. 93
5.3 Average localization error for experiment 1 (first and second columns) and 2 (third and fourth columns). All values are expressed in degrees. LOCATA (O) and (D) are the results following training using SimSW and SimDirect respectively. Both entries in the aforementioned columns show the same value for SRP, as it is not trained. 96
5.4 Average localization error for experiment 3 97
5.5 Average multi-source metrics and standard deviations of Neural-SRP and DOANet on the testing TAU-NIGENS dataset. Metrics and deviations were computed by averaging across the 3 training experiments. 97
5.6 Complexity analysis of Neural-SRP and baselines. The cells containing three numbers refer to 4, 8 and 12 microphones respectively. 98
5.7 Comparison of the mean and standard deviation of the localization error between multiple datasets and models. 104

List of Abbreviations

ADC Analogue-to-Digital Converter

AIV Augmented Intensity Vectors

CFRC Coarse-To-Fine Region Contraction

CMA Centralized Microphone Array

CNN Convolutional Neural Network

CPMA Coprime Microphone Array

CPU Central Processing Unit

CRNN Convolutional Recurrent Neural Network

CUDA Compute Unified Device Architecture

CVNN Complex-Valued Neural Network

DCASE Detection and Classification of Acoustic Scenes and Events

DFT Discrete Fourier Transform

DI-NN Dual-Input Neural Network

DMA Distributed Microphone Array

DOA Direction-of-Arrival

ESPRIT Estimation of Signal Parameters via Rotational Invariance Techniques

FDN Feedback Delay Network

FFT Fast Fourier Transform

FLOPS Floating-Point Operations

FPGA Field Programmable Gate Array

FC-NN Fully Connected Neural Network

GCC Generalized Cross-Correlation

GCC-PHAT Generalized Cross-Correlation with Phase Transform method of estimating TDOA

GCN Graph Convolutional Network

GNN Graph Neural Network

GPU Graphics Processing Unit

GRU Gated Recurrent Unit

IDFT Inverse Discrete Fourier Transform

IPP Integrated Performance Primitives

ISM Image Source Method
LS Least Squares
MAE Mean Absolute Error
MCCC Multichannel Cross-Correlation
MFCC Mel-frequency Cepstral Coefficients
ML Maximum Likelihood
ML-SSL Maximum Likelihood Sound Source Localization
MLP Multi-layer Perceptron
MUSIC Multiple Signal Classification
NMF Non-negative Matrix Factorization
PReLU Parametric Rectified Linear Unit
PSSL Positional Sound Source Localization
RBFN Radial Basis Function Network
RelNet Relation Network
ReLU Rectified Linear Unit
RIR Room Impulse Response
RMS Root Mean Square
RMSAE Root Mean Square Angular Error
RMSE Root Mean Square Error
RNN Recurrent Neural Network
SELD Sound Event Localization and Detection
SLF Spatial Likelihood Function
SNR Signal-to-Noise Ratio
SOF Spatial Observability Function
SRC Stochastic Region Contraction
SRP Steered Response Power
SRP-PHAT Steered Response Power with Phase Transform
SSL Sound Source Localization
STFT Short Time Fourier Transform
SVM Support Vector Machine
T/F Time/Frequency
TDOA Time-Difference-of-Arrival
TOF Time-of-Flight
UAV Unmanned Aerial Vehicle
ULA Uniform Linear Array
VAD Voice Activity Detector
WASN Wireless Acoustic Sensor Network

WGN White Gaussian Noise

XSRP eXtensible SRP

List of Symbols

Conventions

r	Scalars are represented as italic letters
$\mathbf{r}$	Vectors are represented in boldface
$\hat{r}, \tilde{r}$	Estimate or Approximation of variable r
$\bar{r}$	Frequency-domain quantities are marked with a top bar
$\mathbf{r}[k]$	k-th element of vector $\mathbf{r}$
$\mathbf{R}$	Matrices are represented using uppercase bold letters
$\mathcal{R}, \mathcal{R}(t)$	Sets and custom functions are represented using uppercase caligraphic letters

Acoustics

c	Speed of sound
N	Number of active sources
M	Microphone array size
P	Number of microphone pairs
$\mathbf{u}_n$	Cartesian coordinates of sound source n. In the single source problem, $\mathbf{u}$ is used.
$\mathbf{v}_m$	Cartesian coordinates of microphone m
$\mathcal{G}$	Set of candidate source positions forming the search grid
$a_m(\mathbf{u})$	Propagation attenuation from a source at $\mathbf{u}$ to microphone m.
$h_m(t; \mathbf{u})$	Room Impulse Response between a source at $\mathbf{u}$ and microphone m.
$\tau_m(\mathbf{u})$	Propagation delay from a source at $\mathbf{u}$ to microphone m.
$\tau_{lm}(\mathbf{u})$	TDOA between microphones l and m to a source at $\mathbf{u}$

Signal processing

f_s	Sampling rate
L	Signal frame size
$s_n(t)$	Sample of signal emitted by source n at time t

$x_m(t), \mathbf{x}_m(t)$ Received sample or frame at time t for microphone m

$\bar{x}_m(t, f)$ Received signal sample at time-frequency bin t, f for microphone m

$\bar{\boldsymbol{x}}_m(t)$ Received frequency-domain frame at time t for microphone m

$\mathbf{g}_t(t), \bar{\mathbf{g}}_f(t)$ Time or frequency GCC-PHAT vector at time t

$\mathcal{X}(t)$ The set of microphone signal frames at time t

Special functions

$\mathrm{DFT}(\mathbf{x}), \mathrm{IDFT}(\bar{\mathbf{x}})$ Forward and inverse Discrete Fourier Transform operation

$\mathrm{STFT}(\mathbf{x}), \mathrm{ISTFT}(\bar{\mathbf{x}})$ Forward and inverse Short-time Fourier Transform operation

$\mathrm{CC}(\tau; \mathbf{x}_l, \mathbf{x}_m)$ Temporal cross-correlation function of signals $\mathbf{x}_l$ and $\mathbf{x}_m$ at time lag τ

$\mathrm{GCC}(\tau; \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m)$ GCC-PHAT function between signals $\bar{\mathbf{x}}_l$ and $\bar{\mathbf{x}}_m$ at time lag τ

$\mathrm{SRP}(\mathbf{u}\,;\ \mathcal{X})$ Temporal SRP-PHAT function for set of signals $\mathcal{X}$ evaluated at candidate location $\mathbf{u}$

Chapter 1

Introduction

1.1 Motivation and objectives

MICROPHONE ARRAY SIGNAL PROCESSING [1], [2] is an established field within the acoustic signal processing community. A microphone array is an electronic device containing two or more microphones. A little known fact is that most of today's cell phones, laptops, voice assistants and conference systems have multiple microphones, and internally implement array processing techniques for tasks such as speech enhancement and noise reduction to improve the acoustic experience on conference calls and recordings. Microphone arrays can also be created by wirelessly connecting multiple devices, creating what is known as a Wireless Acoustic Sensor Networks (WASNs) or Distributed Microphone Arrays (DMAs).

This thesis focuses on the task of Sound Source Localization (SSL) using microphone arrays. SSL analyzes the signals received by each microphone along with their positional information to estimate the location of one or more sources of interest. Some of SSL applications are robot orientation [3], noise reduction, speech enhancement [4], and diarization [5]. SSL is used by mainstream products, such as the Amazon Echo voice assistant [6].

SSL has been widely studied over several decades and its origin can be traced to radar processing techniques, which have played an instrumental role in World War II. Since then, multiple signal processing methods have been developed although state of

the art localization performance is nowadays obtained though the use of deep learning techniques. This thesis is focused on an important shortcoming of most current Deep-SSL methods – their dependence on microphone and/or room geometry, i.e., the requirement of the number and positions of microphones to be matched during training and testing, restricting practical applications of the available methods.

The main contributions of this work are the creation of a novel class of neural network models named Neural-SRP. Unlike current state-of-the-art networks for SSL, the novel Neural-SRP approach is able to function on microphone arrays and rooms of arbitrary geometry, allowing them to be applied to DMAs as well on centralized microphone arrays.

1.2 Publications

During the course of this work, the following have been published or submitted:

1. E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes and P. A. Naylor, "The Neural-SRP Method for Universal Robust Multi-Source Tracking," in IEEE Open Journal of Signal Processing, vol. 5, pp. 19-28, 2024 [7]

2. E. Grinstein, V. W. Neo, and P. A. Naylor, "Dual input neural networks for positional sound source localization," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, no. 1, p. 32, 2023 [8]

3. E. Grinstein, M. Brookes and P. A. Naylor, "Graph Neural Networks for Sound Source Localization on Distributed Microphone Networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5 [9]

4. E. Grinstein, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP method for positional sound source localization," in Asilomar Conf. Signals Syst. Computers, 2023 [10]

5. E. Grinstein and P. A. Naylor, "Deep Complex-Valued Convolutional-Recurrent Networks for Single Source DOA Estimation," International Workshop on Acoustic Sig-

nal Enhancement (IWAENC), Bamberg, Germany, 2022, pp. 1-5 [11]

6. E. Grinstein, E. Tengan, B. Çakmak, T. Dietzen, L. Nunes, T. van Waterschoot, M. Brookes, and P. A. Naylor "Steered Response Power for Sound Source Localization: a tutorial review," submitted to EURASIP Journal on Audio, Music and Signal Processing, 2024 [12].

1.3 Thesis outline

The remaining chapters in this thesis are organized as follows.

Chapter 2 provides a literature review on source localization, describing most relevant methods and trends.

Chapter 3 presents the Dual-Input Neural Network (DI-NN) architecture, which was used on the proposed Neural-SRP method.

Chapter 4 formulates the problem of SSL using a Graph Neural Network (GNN) paradigm, which is adopted on the proposed Neural-SRP SSL method.

Chapter 5 presents Neural-SRP, the main contribution of this work which is based on the last two previous chapters.

Chapter 6 provides an in depth review of the Steered Response Power (SRP) method, as well as presents a generalized framework for describing and extending variations.

Chapter 7 discusses experiments using Complex-Valued Neural Networks (CVNNs) and their advantages to conventional networks for the task of SSL.

Chapter 8 presents a conclusion to this thesis.

1.4 Problem statement

The problem of SSL can be characterized according to:

- Array properties: the type of microphone array used, namely, a Centralized Microphone Array (CMA) or a DMA. This also includes the number of microphones used

as well as their relative positions.

- Coordinate system properties: the type of output produced, namely, Cartesian coordinates or an angular output. These outputs can also be viewed as distinct tasks respectively named Positional Sound Source Localization (PSSL) and Direction-of-Arrival (DOA) estimation
- Source properties: specifically, the number of maximum active sources, and if sources move or remain static.
- Environmental properties: if localization is performed indoors or outdoors, as well its noise properties, namely, directional or difuse sources and reverberation caused by reflective surfaces and objects
- Interconnection and interoperability properties: if cooperation between sources and microphones exist (active SSL) or not (passive SSL)

Near- versus Far-field localization

This subsection discusses the different types of localization which are frequently encountered in the literature, namely, Positional Sound Source Localization (PSSL) and DOA estimation. PSSL consists of fully estimating the source's position, and is usually employed when the distances between microphones in the array is similar to the distance between the microphones and the source. This is equivalent to saying the source is located in the *near-field* of the array. This configuration is referred to as a *distributed* array, which can be constituted for example of multiple network-connected devices such as laptops, cell phones or voice assistants. In this case, as each device has their own Analogue-to-Digital Converter (ADC), they must be synchronized to a common sampling frequency f_s, or a compensation algorithm must be applied to the signals to prevent synchronization issues [13].

Conversely, when employing a centralized microphone array such as a single voice assistant, the distance between microphones is usually significantly smaller than the distance between the sources of interest and the array itself. This is equivalent to saying

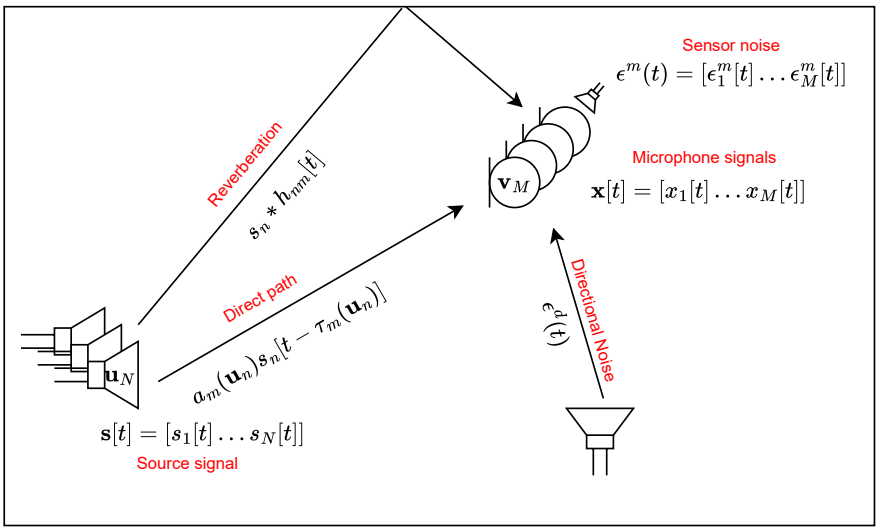

Figure 1.1: Depiction of the SSL problem, where source signals s_n propagate from positions $\mathbf{u}_n$ to be received at each microphone as signals x_m, which is affected by noise and reverberation.

the source is located in the array's *far-field*. In this case, the spherical wave leaving the source is observed as a plane wave which has no defined origin: an infinite set of sources may produce a plane wave with the same incident angle to the array. For this reason, the range ρ is usually not estimated when using compact arrays. The task of estimating the azimuth, ψ, and elevation, θ, is referred to as DOA estimation.

1.5 Signal Model

This chapter is finished with the definition of the signal model used throughout this work. This model encapsulates propagation from a maximum of N sources to each of the M microphones in the array. A single model is described for both PSSL and DOA estimation, although different interpretations for the variables are used for each task. These definitions are based on [7].

The N sources of interest and M microphones in the array are located in a 3-dimensional Cartesian system of coordinates. The M known microphone positions at time t are $\mathbf{v}_m(t) = [v_m^{(1)}(t)\, v_m^{(2)}(t)\, v_m^{(3)}(t)]^T$ for $1 \leq m \leq M$. Conversely, the $N(t)$ active sources at time t are represented using the set of positions $\mathcal{U}(t) = \{\mathbf{u}_1(t) \dots \mathbf{u}_N(t)\}$, where

$\mathbf{u}_n$ is defined analogously to $\mathbf{v}_m$. The sources may also be expressed in spherical coordinates $\mathbf{u} = [\psi\, \theta\, \rho]^T$ with respect to a reference point, typically the centre of a microphone array. Variables ψ, θ and ρ respectively represent the source's *azimuth*, *elevation* and *range*.

Although microphones discretize the continuous signal emitted by the source, it is advantageous to represent the propagation effects in continuous time. The continuous signal arriving at microphone m from source n is equal to

$$x_m(t) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} h_m(r; \mathbf{u}_n) s_n(t-r) \mathrm{d}r + \epsilon_m(t), \tag{1.1}$$

that is, a sum of convolutions between the Room Impulse Responses (RIRs) $h_m(t; \mathbf{u}_n)$ between the sources and microphone, which models the effects of propagation delay, attenuation and reverberation, plus a noise term which models sensor noise. In the case of Gaussian sensor noise, $\epsilon_m(t) \sim \mathcal{N}(\mathbf{0}, \sigma_m^2 I)$, where σ_m controls the Signal-to-Noise Ratio (SNR). In the case of a directional noise source, such as a fan, the noise term is defined as

$$\epsilon_m(t) = \sigma_m\big(\mathbf{h}_{m\epsilon} * \varepsilon(t)\big), \tag{1.2}$$

the impulse response $\mathbf{h}_{m\epsilon}$ convolved with a random signal $\varepsilon \sim \mathcal{N}(\mathbf{0}, I)$ scaled by a factor σ_m. Note the noise impulse response is not time-dependent, as we assume directional noise sources to remain spatially stationary at unknown position $\mathbf{u}_\epsilon$.

Many systems are modeled using a simplified model, where reverberation is modeled using the noise term $\epsilon_m(t)$. It is described as

$$x_m(t) = \sum_{n=1}^{N} a_m(\mathbf{u}_n) s(t - \tau_m(\mathbf{u}_n)) + \epsilon_m(t), \tag{1.3}$$

that is, the signal emitted by source n is received at microphone m attenuated by a factor $a_m(\mathbf{u})$, delayed by $\tau_m(\mathbf{u}_n)$ samples and corrupted by noise term $\epsilon_m(t)$. This is equivalent to adopting the RIR in (1.1) as a pure impulse $h_m(t; \mathbf{u}) = a_m(\mathbf{u})\delta(t - \tau_m(\mathbf{u}))$. Note that this model assumes attenuation to be frequency-independent. In the next chapter, energy-based methods, which exploit the attenuation $a_m(\mathbf{u}_n)$ to estimate $\mathbf{u}_n$ will be presented,

as well as delay based methods, which instead exploit $\tau_m(\mathbf{u}_n)$ for the same aim.

Alternatively, it is often advantageous to define (1.3) in the frequency domain, by assuming the source signal to be a sum of complex-valued sinusoids $\bar{s}(t, f)$ of frequencies f. In practice, such a signal can be obtained by applying the Short Time Fourier Transform (STFT) on $s(t)$. $x_m(t, f)$ is then defined for each time-frequency pair t, f as

$$\bar{x}_m(t, f) = \sum_{n=1}^{N} \bar{s}(t, f) a_m(\mathbf{u}_n, f) e^{-jf\tau_m(\mathbf{u}_n)} + \epsilon_m(t, f). \tag{1.4}$$

The advantage of (1.4) in comparison to (1.3) is that delay τ_m and atenuation $a_m(f)$ effects can be jointly represented by multiplication with a complex-valued scalar.

Although the above definitions are conceptually useful, in practice, the algorithms presented in this thesis are typically computed using a *frame* or vector of dimension L discrete samples for each microphone. A frame $\mathbf{x}_m(t)$ is defined in the time domain as

$$\mathbf{x}_m(t) = [x_m(t)\, x_m(t - T_s) \ldots x_m(t - (L-1)T_s)]^T, \tag{1.5}$$

where $T_s = 1/f_s$. Furthermore, a frequency domain frame $\bar{\mathbf{x}}_m(t)$ is defined as

$$\bar{\mathbf{x}}_m(t) = \mathrm{DFT}(\mathbf{x}_m(t)), \tag{1.6}$$

that is, the application of the Discrete Fourier Transform (DFT) to temporal frame $\mathbf{x}_m(t)$. $\bar{\mathbf{x}}_m(t)$, where each of its entries represents a time-frequency bin $\bar{x}_m(t, f)$ with $f \in \mathcal{F}$, where

$$\mathcal{F} = \{f | f = -f_s/2 + kf_s/L,\, k = 0, \ldots, L-1\}, \tag{1.7}$$

typically constitutes the set of analysis frequency components used.

1.6 Evaluation metrics

The main metric used for single source DOA estimation is the Root Mean Square Angular Error (RMSAE) [14], defined for a pair of positions $(\mathbf{u}, \hat{\mathbf{u}})$ each with azimuth and elevations (θ, ψ) and $(\hat{\theta}, \hat{\psi})$ respectively, as

$$\mathcal{E}_l(\mathbf{u}, \hat{\mathbf{u}}) = \arccos^2(\cos\theta \cos\hat{\theta} + \sin\theta \sin\hat{\theta} \cos(\psi - \hat{\psi})), \tag{1.8}$$

where (1.8) is usually averaged for all frames in the evaluation dataset. Conversely, for PSSL, the localization error or Root Mean Square Error (RMSE) is defined as

$$\mathcal{E}_l(\mathbf{u}, \hat{\mathbf{u}}) = \|\mathbf{u} - \hat{\mathbf{u}}\|. \tag{1.9}$$

Another important metric is the percentage of anomalies. An anomalous estimate is defined as the percentage of samples in the testing dataset which surpass a given localization error threshold γ. In mathematical terms, it is defined for a dataset of K samples as

$$\mathcal{E}_a = \sum_{k=1}^{K} [\mathcal{E}_l(\mathbf{u}_k, \hat{\mathbf{u}}_k) > \gamma], \tag{1.10}$$

where $[\cdot]$ represents the Iverson bracket [15], which outputs 1 if the condition is true, or 0 otherwise. In practice, the choice of γ depends on the application.

For multiple sources, the localization error is defined for each correctly detected source. For multi-source experiments, the detection metrics of precision, recall and the F1 scores are also used, as defined in [16], [17]. These metrics are computed for each frame, based on the number of true and estimated sources $|\mathcal{U}|$ and $|\hat{\mathcal{U}}|$. $|\mathcal{U}|$ and $|\hat{\mathcal{U}}|$ are first used to compute the number of true positive $\mathrm{TP} = \min(|\mathcal{U}|, |\hat{\mathcal{U}}|)$, false positive $\mathrm{FP} = \max(0, |\hat{\mathcal{U}}| - |\mathcal{U}|)$ and false negative $\mathrm{FN} = \max(0, |\mathcal{U}| - |\hat{\mathcal{U}}|)$ detections. Finally, the Precision (PR), Recall (RE) and F1 metrics are computed as:

$$\begin{aligned} \mathrm{PR} &= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \\ \mathrm{RE} &= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \\ \mathrm{F1} &= 2(\mathrm{PR} \times \mathrm{RE})/(\mathrm{PR} + \mathrm{RE}). \end{aligned} \tag{1.11}$$

As in the single source metrics, the final metrics are usually obtained by averaging all frame metrics in the evaluation dataset.

1.7 Reverberation, RIRs and simulation

As described in (1.1), the acoustic channel between the two points $(\mathbf{u}_n, \mathbf{v}_m)$ where a source and a microphone are located is typically characterized using a RIR, which models the propagation and reverberation effects. A RIR may be divided in three segments,

namely, the direct path, early and late reflections. The RIR's delay until the first impulse encodes the Time-of-Flight (TOF) $\tau_m(\mathbf{u}_n)$ between source and microphone, while the level of the impulse encodes the attenuation coefficient $a_m(\mathbf{u}_n)$. In addition, the early and late reverberation zone of the RIR encode the effects caused by the room surfaces such as walls and furniture. Surfaces are characterized by their absorption/reflection coefficients, which are known to be frequency-dependent [18]. A room is frequently classified by its RT60 (or T60), defined as the time taken for the impulse responsez to decay by 60 dB from its peak level. In practice, each combination of source and microphone position produces a different RT60, so an average of multiple points may be computed.

Many methods exist for measuring RIRs. Although the most straightforward method is to have a microphone record an impulsive sound (such as a gunshot) played by a loudspeaker, sine-sweeping techniques were found to produce more robust measurements [19]. These RIRs may then be convolved with anechoic source signals to produced an *auralized* signal. As recording a RIR may be a time-consuming activity, many SSL algorithms are evaluated using *simulated* RIRs generated using a room acoustics simulator.

A standard method of simulating RIRs is the Image Source Method (ISM) developed by Allen and Berkley [20]. This method models sound to travel on straight lines or rays. Each time the ray encounters a surface, a virtual, or *image* source is created using the surface as a mirror. The number of times this process is recursively applied to the rays associated with the image sources is defined as the method's *order*. The realistic RIRs produced by the ISM comes at the cost of its exponential complexity to its order [21]. For this reason, many alternatives and variations have been studied on recent decades. For example, a low-order ISM can be used to simulate the direct path and early reflections, while the late reflections can be efficiently simulated using an Feedback Delay Network (FDN) [22].

1.8 Applications

SSL is a foundational task which has been applied in many domains, having been used as an input feature for speech enhancement/beamforming tasks [23]–[25], voice activity detection [26], [27], speaker diarization [5], [28]–[31], sound source separation [32]–[34] and

array calibration [35]. This section focuses on SSL applications using the SRP method, which is of particular interest to this thesis.

Although SSL can be used to localize any type of sound source, many applications focus on a specific sound event. A prominent application is that of surveillance and defence. SRP can be used to localize irregular Unmanned Aerial Vehicles (UAVs) activity [36], [37], as well as using an UAV with an embedded microphone array to localize sources of interest itself [38], [39]. Other applications in security include intrusion detection [40], [41], and gunshot localization [42], [43].

Another category of interest is that of scene understanding in large and/or outdoor environments, such as the detection of indoor and outdoor sources of noise pollution [44]–[46] and the detection of underground seismic events [47]. SRP was also applied for commercial and environmental purposes, such as the localization of sound-emitting fish using an underwater hydrophone [48], and to detect faulty equipment within electrical power stations [49]. Furthermore, with the increased interest in smart and self-driving vehicles sensors, localization of horns and crashes using SRP [50], [51] can also be performed, or localizing talkers inside the vehicle itself [52].

Turning to indoor environments, SRP can be applied to the medical domain, being used to localize and analyze footsteps with the goal of early detection of dementia [53], as well as for fall detection of elderly people [54]. SRP can also be used to improve human-robot interactions [55]–[57], as well as for camera steering corporate meetings [58] and smart rooms [59], [60]. SRP was also applied to a helmet-mounted microphone array [61], which can be used for increasing acoustic awareness on industrial sites, for example.

Chapter 2

Sound Source Localization

This section presents the current state-of-the-art on Sound Source Localization (SSL) methods, which are classified as signal processing- and neural-based. Signal processing-based methods often have the advantage of being explainable, and having a lower computational complexity than neural methods. By directly processing relevant environmental metadata such as the microphone array geometry, the former methods can be directly applied to any environment. The main advantages of neural SSL methods are in their superior localization perfomance in adverse environments, and straightforward formulation, where acoustic knowledge is intrinsically embedded in the training data instead of requiring domain expertise. Most signal processing-based methods use formulation (1.3) as their signal model, which models the propagation effects as a delay $\tau_m(\mathbf{u})$ and attenuation $a_m(\mathbf{u})$ of the source signal. $\tau_m(\mathbf{u})$ is also known as the Time-of-Flight (TOF) between the microphone and source, can therefore be expressed, in seconds, as

$$\tau_m(\mathbf{u}) = \frac{\|\mathbf{u} - \mathbf{v}_m\|}{c}, \tag{2.1}$$

where c is the speed of sound. In turn, the attenuation $a_m(\mathbf{u})$ is defined as a combination of the individual microphone gains and attenuation caused by propagation, which is accepted to follow an inverse-square law. By assuming the microphones are calibrated at unit gain, a_m becomes

$$a_m(\mathbf{u}) = \frac{1}{\|\mathbf{u} - \mathbf{v}_m\|^2}. \tag{2.2}$$

For a single source, if $\tau_m(\mathbf{u}$ or $a_m(\mathbf{u})$ can be accurately estimated, a system of equa-

tions can be solved to obtain the source coordinates. However, this is only possible in active localization systems, which transmit power and transmission start/end timestamps to the SSL system. However, this thesis focuses on *passive* localization systems, where such information is unavailable, such as in the scenario of localizing human speakers. Passive systems often focus on estimating relative measures between *pairs* of microphones. An important metric is the Time-Difference-of-Arrival (TDOA) between a pair of microphones l and m, defined as $\tau_{lm}(\mathbf{u}) = \tau_l(\mathbf{u}) - \tau_m(\mathbf{u})$, or

$$\tau_{lm}(\mathbf{u}) = \frac{\|\mathbf{u} - \mathbf{v}_l\| - \|\mathbf{u} - \mathbf{v}_m\|}{c}. \tag{2.3}$$

The TDOA for a pair of microphones can be interpreted as how much earlier/later a signal arrives at the first microphone in comparison to the second microphone. Multiple positions $\hat{\mathbf{u}}$ produce the same delay τ_{lm} for a pair of microphones fixed at $(\mathbf{v}_l, \mathbf{v}_m)$. These positions lie along a hyperbola/hyperboloid branch in 2D/3D, as it is shown in [62] and can be viewed in Fig. 2.1. The maximum possible TDOAs for a microphone pair occurs when the source and microphones are colinear and the source is not located between the microphones. The maximum TDOA is then obtained as

$$\tau_{lm}^{\text{lim}} = \|\mathbf{v}_l - \mathbf{v}_m\|/c. \tag{2.4}$$

By determining the intersection of the hyperbolas produced by multiple microphone pairs, the source position can be estimated. Approaches utilizing this strategy are often called triangulation, TDOA-based, *indirect* or two-step approaches [63], since they require a first step of estimating the TDOA before estimating the source's location. Although these approaches are computationally efficient, their over-reliance on the estimated TDOAs make them unsuitable under adverse noisy or reverberant scenarios, a weakness which is mitigated at the cost of higher computational complexity by methods such as Steered Response Power (SRP). The following section provides a discussion on energy, neural and subspace-based localization methods.

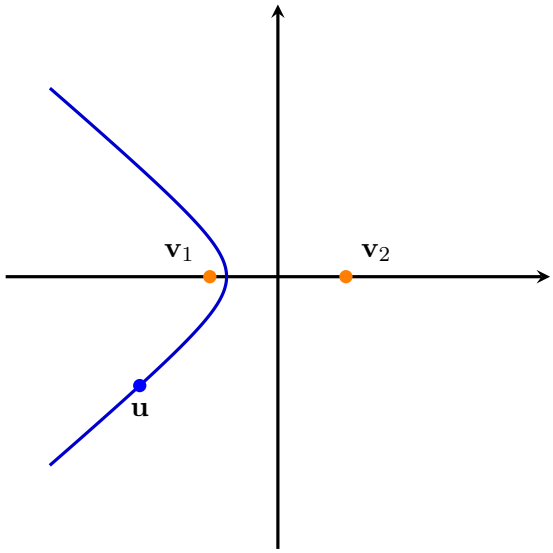

Figure 2.1: Hyperbola branch of points with the same TDOA as a source located at $\mathbf{u}$ with respect to microphone positions $\mathbf{v}_1$ and $\mathbf{v}_2$. The axes represent the horizontal directions, in meters.

2.1 Delay-based methods

2.1.1 Measuring TDOA: Cross-correlation and GCC-PHAT

The TDOA τ_{lm} between two microphones can be estimated as the argument of the peak of the cross-correlation between microphone signal frames $\mathbf{x}_l(t)$ and $\mathbf{x}_m(t)$. The discrete cross-correlation (CC) function is defined as

$$\mathrm{CC}(\tau \mid \mathbf{x}_l, \mathbf{x}_m) = \mathbf{x}_l^T(t)\mathbf{x}_m(t-\tau), \tag{2.5}$$

where τ must be a multiple of the sampling period T_s and appropriate zero padding is applied.

Despite its straightforward formulation, (2.5) is seldom used in practice for localizing speech sources in reverberant and noisy environments, as the non-flat spectrum of the source signal reduces the selectivity of the function. Instead, the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) function [64], [65] is usually adopted. 'Generalized' comes from the fact that a cross-correlation value is produced for every frequency component of the signals after a pre-filtering operation. This operation is typically the 'Phase Transform' weighting, which whitens the frequency components, thus sharpen-

ing the correlation peak. The GCC-PHAT function is defined for each time-frequency bin as

$$\text{GCC-PHAT}(f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m) = \frac{\bar{x}_l(t,f)\bar{x}_m^*(t,f)}{|\bar{x}_l(t,f)|\,|\bar{x}_m(t,f)|}. \tag{2.6}$$

The denominator of (2.6) is often referred to as the Phase Transform (PHAT). In practice, (2.6) is computed for a set of analysis frequencies $\mathcal{F}$ to generate a GCC frame $\bar{\mathbf{g}}$. In practice, the Fast Fourier Transform (FFT) algorithm of size L is used, where L is typically chosen to be a power of 2, to obtain their frequency-domain representation. In those cases, $\mathcal{F}$ is thus implicitly defined as $\lfloor (L/2) \rfloor + 1$ uniformly spaced non-negative frequencies up to the Nyquist rate, where $\lfloor \cdot \rfloor$ represents the floor operation. Conversely, by applying the Inverse Discrete Fourier Transform (IDFT) to $\bar{\mathbf{g}}$, a time-domain vector $\mathbf{g}$ can obtained,

$$\mathbf{g} = \text{IDFT}(\bar{\mathbf{g}}), \tag{2.7}$$

where each entry $\mathbf{g}[k]$ represents a temporal correlation value between $\mathbf{x}_l$ and $\mathbf{x}_m$ at sample k. A frame can be built in a similar manner using the temporal CC in (2.5). The magnitude normalization in (2.6) improves the resolution of (2.7) by giving equal weight to all frequency components and focusing on phase information only. An example comparison between two frames computed using (2.7) and temporal CC is shown in Fig. 2.2, where it can be observed that the peak produced by GCC-PHAT is much sharper than by CC. This can be explained because the PHAT operation makes the source signal white, which results in a single peak in its autocorrelation function.

In an ideal scenario, the temporal CC or GCC-PHAT function exhibits a sharp peak at τ_{lm}, which can be used for two-step methods. However, under reverberant or noisy scenarios, the cross-correlation function can exhibit multiple peaks, rendering the TDOA estimates and the subsequent triangulation-based approaches unreliable. As will be shown in the following section, more reliable methods such as Steered Response Power (SRP) apply the principle of least commitment [66], [67]; instead of estimating τ_{lm} early on and discarding all other values and peaks of the cross-correlation function, SRP associates each cross-correlation value with a candidate locus in space using (2.3).

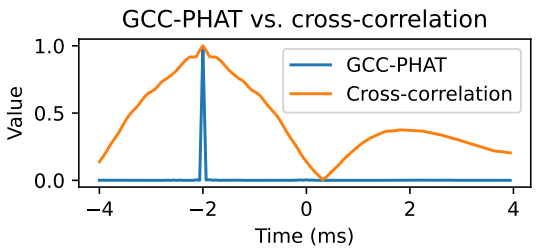

Figure 2.2: Example comparison between the normalized temporal cross-correlation and GCC-PHAT for a scenario containing two closely spaced microphones and a speech source producing a TDOA of -2 ms.

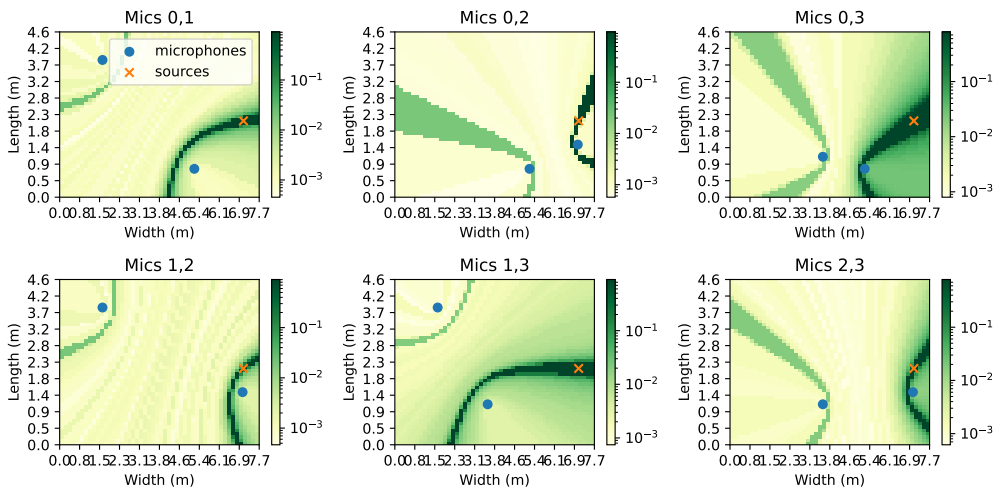

Figure 2.3: Pairwise SRP maps

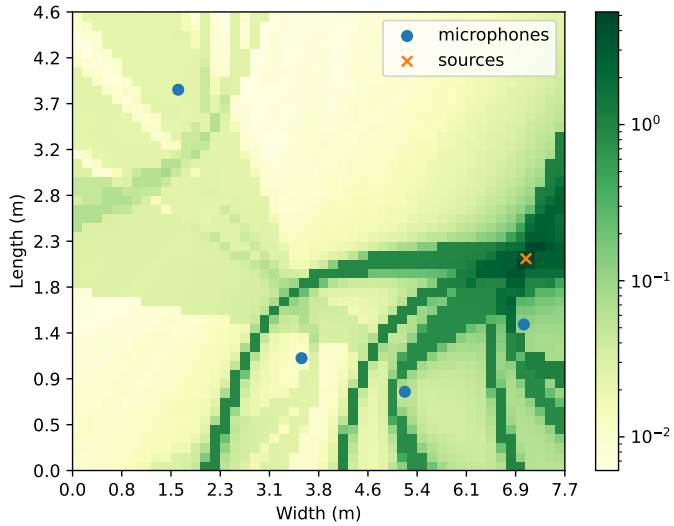

Figure 2.4: Global SRP maps, obtained by summing the maps shown in Fig. 2.3

2.1.2 Time-domain Steered Response Power (SRP) formulation

The conventional SRP for a candidate source location $\mathbf{u}$ and a pair of microphones (l, m) is defined as [68], [69]

$$\mathrm{SRP}_{lm}(\mathbf{u} \mid \mathbf{x}_l, \mathbf{x}_m) = \mathrm{CC}(\lfloor \tau_{lm}(\mathbf{u}) \rceil \mid \mathbf{x}_l, \mathbf{x}_m), \tag{2.8}$$

that is, the cross-correlation function between signal frames $\mathbf{x}$ and $\mathbf{x}_m$, evaluated at delay $\lfloor \tau_{lm}(\mathbf{u}) \rceil$, where $\lfloor \cdot \rceil$ represents rounding to the nearest multiple of T_s. Note that the time index t is hereafter omitted for clarity, and that time-domain GCC-PHAT defined in (2.7) is usually preferred to (2.5) in practice for its improved performance on realistic scenarios. The time-domain SRP formulation is presented for the sake of clarity and completeness. Finally, the global SRP is defined as the sum of all pairwise SRPs,

$$\mathrm{SRP}(\mathbf{u} \mid \mathcal{X}) = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \mathrm{SRP}_{lm}(\mathbf{u} \mid \mathbf{x}_l, \mathbf{x}_m), \tag{2.9}$$

where $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ is the set of M L-dimensional frames pertaining all microphones. This value is related to the likelihood of a source being located at a candidate point $\mathbf{u}$.

The complete SRP method consists of evaluating (2.9) for a set of candidate locations and selecting the location maximizing (2.9) as the estimated location. The set of candidate locations typically consists of a regularly sampled spatial grid. The grid construction procedure will be defined in Sec. 2.1.4.

2.1.3 Frequency-domain SRP formulation

This formulation decomposes the microphone signals into frequency bands, which are independently analysed using GCC-PHAT in (2.6) as

$$\begin{aligned}\mathrm{SRP}_{lm}(\mathbf{u}, f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m) = \\ \mathrm{GCC\text{-}PHAT}(f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m) e^{jf\tau_{lm}(\mathbf{u})}.\end{aligned} \tag{2.10}$$

Equation (2.10) can be interpreted as steering, or shifting, the microphone signal $x_m(f)$ by a phase $jf\tau_{lm}(\mathbf{u})$. Note that although (2.10) produces a complex value, its imaginary part is typically discarded as irrelevant [70]. Finally, the global SRP is represented in the frequency domain in a similar way to the time-domain formulation (2.9), after summing over the set $\mathcal{F}$ of frequencies being analyzed,

$$\mathrm{SRP}(\mathbf{u} \mid \bar{\mathcal{X}}) = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \sum_{f \in \mathcal{F}} \mathrm{SRP}_{lm}(\mathbf{u}, f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m), \tag{2.11}$$

where $\bar{\mathcal{X}}$ is the frequency-domain representation of $\mathcal{X}$. In practice, the frequencies used may be lower than the Nyquist rate to prevent a phenomenon called spatial aliasing [71]. Note the time and frequency definitions of SRP are not equivalent due to this aforementioned low-pass filtering, as well as the rounding operator required when constructing a temporal CC or GCC vector makes (2.9) operate using integer delays, which may not correspond to the true source's TDOA. The error due to rounding may be reduced by using distributed arrays or mitigated by applying interpolation [72]. However, significant errors may be produced for compact arrays, where the TDOA range defined by (2.4) is typically only a few samples [73].

2.1.4 Grid construction and search

To estimate the location of the source, (2.9) or (2.11) are evaluated over a set, $\mathcal{G}$, of G candidate positions relative to a reference point in the room, typically one of its corners. The elements of $\mathcal{G}$ are usually defined by creating a uniform spatial grid. For performing SSL in a cuboid-shaped room, a cuboid-shaped grid is typically used. For example, when performing planar or 2D localization $|\mathcal{G}| = |\mathcal{G}^{(1)}| \times |\mathcal{G}^{(2)}|$, where $|\mathcal{G}^{(1)}|$ and $|\mathcal{G}^{(2)}|$ are respectively the number of points used for the width and length dimension. $\mathcal{G}$ becomes

$$\begin{aligned}\mathcal{G}_{2D} = \{ & [g^{(1)}R^{(1)} \; g^{(2)}R^{(2)}]^T \mid \\ & g^{(1)} \in \{1, \ldots, |\mathcal{G}_{2D}^{(1)}|\} \\ & g^{(2)} \in \{1, \ldots, |\mathcal{G}_{2D}^{(2)}|\}\},\end{aligned} \tag{2.12}$$

where $R^{(1)} = D^{(1)}/|\mathcal{G}_{2D}^{(1)}|$ and $R^{(2)} = D^{(2)}/|\mathcal{G}_{2D}^{(2)}|$ are the width and length *resolution* for a room of width $D^{(1)}$ and length $D^{(2)}$. Conversely, when performing planar or 2D Direction-of-Arrival (DOA) estimation, the grid can be made by setting the origin to the microphone array centre, and a circular grid is created,

$$\begin{aligned}\mathcal{G}_{\text{DOA2D}} = & \\ & \{ [\cos(\psi) \; \sin(\psi)])]^T \mid \psi \in \\ & \{R^{(\psi)}, 2R^{(\psi)} \ldots, 2\pi\}\}.\end{aligned} \tag{2.13}$$

In (2.13), each point represents a distinct candidate source direction. Furthermore, neighboring points are separated by the angular resolution $R^{(\psi)}$, where ψ is the candidate source's *azimuth*. In 3D DOA estimation, the *elevation*, defined as the angle between the segment connecting the source and array centre and the horizontal plane is also estimated.

For both DOA estimation and Positional Sound Source Localization (PSSL), the complete SRP map consists of evaluating the SRP function for all candidate locations in the grid $\mathcal{G}$, and selecting the location producing the maximum SRP value as the estimated position,

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u} \in \mathcal{G}} \text{SRP}(\mathbf{u}). \tag{2.14}$$

An example of an SRP map for a simulated environment of low reverberation is shown

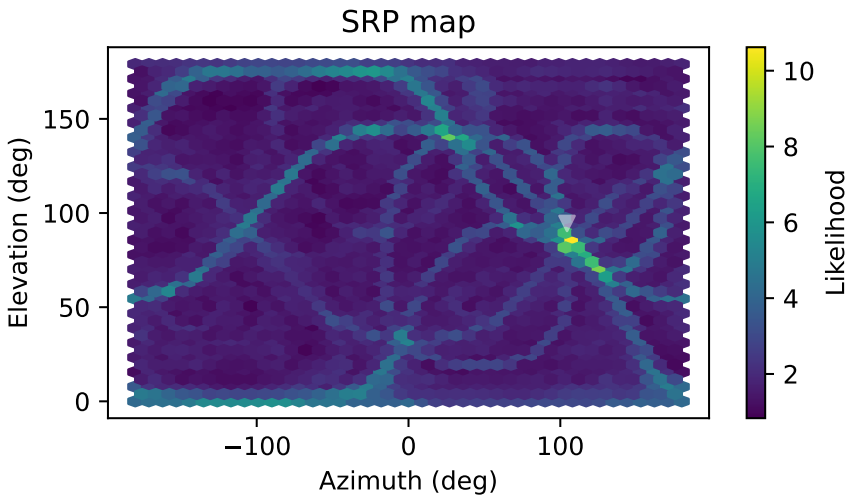

Figure 2.5: Example of an SRP map for the task of 2D DOA estimation. The true source location is located below the triangle. Note the likelihood is composed by a sum of correlation values and is therefore dimensionless.

in Fig. 2.5.

2.1.5 SRP complexity analysis

As SSL algorithms are frequently deployed on embedded devices, analyzing and reducing the algorithms' computational complexity is of high interest. This subsection analyzes the frequency-domain, conventional SRP method as defined in (2.11). This analysis will be revisited in chapter 6. Here, complexity is measured by the number of real multiplications and divisions performed by the algorithm, ignoring the additions, as commonly done. The Bachmann–Landau (or big-O) notation is adopted, which measures asymptotic behaviour of algorithm complexity as input sizes grow.

The method can be divided into four sequential operations. The first two operations consist of extracting the Discrete Fourier Transform (DFT) for each frame of the M microphones followed by computing the GCC-PHAT for all P microphone pairs, where $P = M(M-1)/2$. In practice, the FFT algorithm [74] is used to implement the DFT. The FFT has a complexity of $\mathrm{O}(L \log L)$. The FFT operation is assumed to convert a time-domain frame of size L into a frequency-domain frame of same size. Since GCC-PHAT consists of an element-wise multiplication of the vectors $\bar{\mathbf{x}}_l$ and $\bar{\mathbf{x}}_m$ divided by their respective magnitudes, its complexity is therefore $\mathrm{O}(L)$.

The third step is the creation of the P pairwise SRP likelihood grids of size $G = |\mathcal{G}|$, for all L frequencies, followed by their sum to create a global SRP grid. As this operation consists of multiplying the GCC-PHATs by an exponential $e^{jf\tau_{lm}(\mathbf{u})}$, its complexity is $\mathrm{O}(GPL)$. The final step consists of comparing all grid points to obtain the argument of its maximum, which is the estimated source location. As comparisons are often assumed to offer a lower complexity, this last step is ignored. The number of operations performed by SRP is thus obtained as

$$\begin{aligned} \mathrm{O}_{\overline{\mathrm{SRP}}} &= \mathrm{O}\big(ML\log L + PL + GPL\big), \\ &\simeq \mathrm{O}\big(ML\log L + GPL\big), \end{aligned} \tag{2.15}$$

where the three terms in the first line represent each of the sequential operations discussed above. The simplification on the bottom line is obtained by removing the second term, as $G \gg 1$. It can be seen from (2.15) that straightforward strategies can be followed to reduce the complexity of SRP. One is to use only a subset of microphones $M' < M$ or subselecting $P' < M(M-1)/2$ pairs instead of evaluating all pair combinations. Another is to employ a smaller frame size L and reducing the frequency range in which the SRP map is computed. Finally, a coarser grid can be employed. All these strategies come, however, with a reduction in localization performance. Most of the research presented in this section proposes strategies to reduce the grid size G, or modify the functionality of the conventional SRP method while minimizing the loss in localization performance.

In turn, the computational complexity of time-domain SRP in (2.9) is smaller than in (2.15), as a single map is computed in the time domain instead of L frequency domain maps, i.e., it uses one less nested 'for each' loop. The complexity of (2.9) is therefore expressed as

$$\mathrm{O}_{\mathrm{SRP}} = \mathrm{O}\big(ML\log L + PL + GP\big), \tag{2.16}$$

which disregards the negligible inverse DFT used to obtain the temporal GCC vector (2.7). Furthermore, projection of the cross-correlation function is achieved in (2.9) by accessing an element in the cross-correlation vector, which is more computationally efficient, albeit less precise, than the multiplication by a complex exponential used in the frequency-domain version.

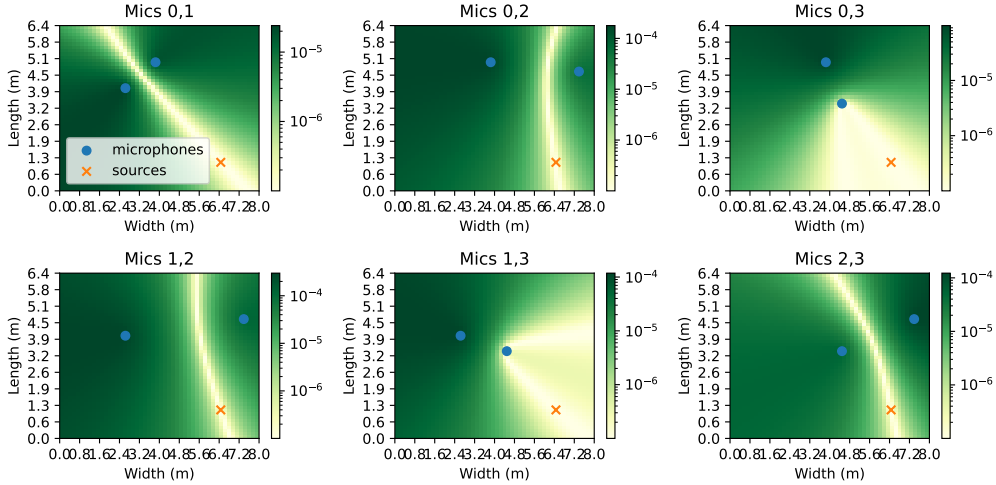

Figure 2.6: Pairwise TDOA-based Least-squares maps

2.1.6 Two-step SSL methods

Delay-based SSL methods usually rely on computing the TDOA between each microphone pair within the system, which corresponds to the difference in time taken for the source signal to propagate to different microphones. The locus of candidate source positions with the same TDOA with respect to a microphone pair is, when considering planar coordinates, a hyperbola [62], [63]. The source is located at the intersection of the hyperbolae defined by all microphone pairs. The multiple TDOAs can be combined using a Least Squares (LS) framework [75], or using a Maximum Likelihood (ML) approach if some noise properties of the system are known or assumed [63]. In general, TDOAs are estimated using cross-correlation based methods such as GCC-PHAT [64], which are shown to be robust to reflections produced in the room due to, for example, the walls, ceiling and furniture, i.e. reverberation [76]. However, these methods are still known to offer less robustness than SRP. The remainder of this subsection presents the Least-Squares method for PSSL, which is popular due to its straightforward formulation and performance on free-field or anechoic environments.

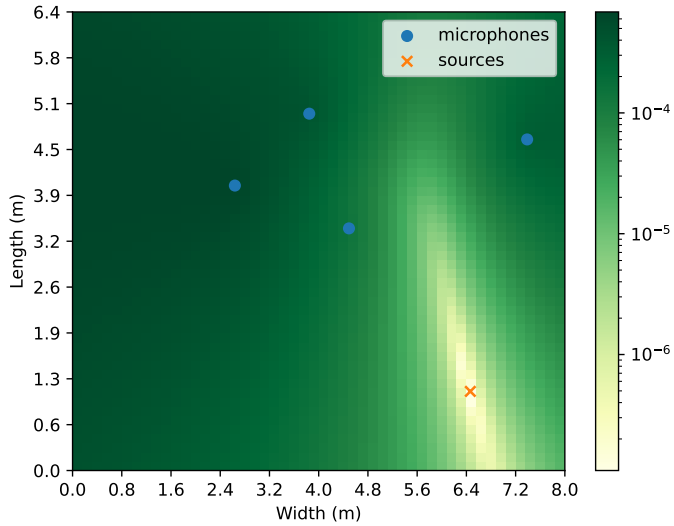

Figure 2.7: Global TDOA-based Least-squares maps, obtained by summing the maps shown in Fig. 2.6

Least-squares approach

The first step in the LS method consists of defining the measured/estimated TDOA $\hat{\tau}_{lm}$ between each microphone pair (l, m) as the location of the cross-correlation peak between the received microphone signals. GCC-PHAT is often preferred to classical cross-correlation, leading to the equation

$$\hat{\tau}_{lm} = \arg\max_{\tau}(\mathbf{g}_{lm}(\tau)/f_s \ , \tag{2.17}$$

A local error function $E_{lm}(\mathbf{u}) = \|\tau_{lm}(\mathbf{u}) - \hat{\tau}_{lm}\|^2$, can then be defined as the squared difference between the theoretical (2.3) and measured (2.17) TDOA of each microphone pair. A global error can then be defined by aggregating local errors for a candidate location $\mathbf{u}$ as

$$E(\mathbf{u}) = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \|\tau_{ij}(\mathbf{u}) - \hat{\tau}_{ij}\|^2. \tag{2.18}$$

To estimate the source's location, $E(\mathbf{u})$ is computed for a set of candidate locations $\mathbf{u}$ within the room, from which the location producing the minimum error is selected as the estimated source location,

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} E(\mathbf{u}) \ . \tag{2.19}$$

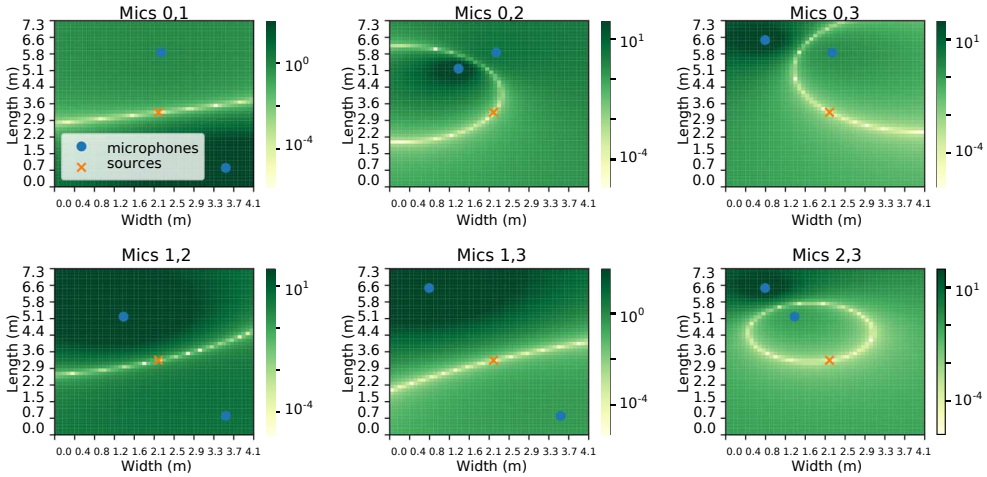

Figure 2.8: Pairwise Energy-ratio Least-squares maps

In the absence of noise and reverberation, this estimate corresponds to the true position of the source [63]. Pairwise and global TDOA-based LS maps are respectively shown in Fig. 2.6 and Fig. 2.7 for a simulated anechoic room. Conversely, the global maps in Fig. 2.10 shows the global maps for energy- and TDOA-based LS methods for SSL, where only SRP is able to correctly localize the source, indicating its robustness to reverberation.

2.2 Energy-based methods

Unlike triangulation methods and SRP, which exploit the relative delay between microphones, energy-based methods localize sounds by exploring the attenuation of acoustic waves as they propagate through open spaces. As these methods are not suited for indoor scenarios where reverberation is present, they are not further explored in this thesis, and were included here for the sake of completeness. By considering a scenario where two microphones and a source are placed in an outdoor environment, the received signals are expected to be delayed and attenuated versions of each other. The *energy* of a signal frame $\mathbf{x}_m(t)$ of size L is defined as

$$E(\mathbf{x}_m^2(t)) = \frac{1}{L}\sum_{k=1}^{L}\mathbf{x}_m(t)[k]^2. \tag{2.20}$$

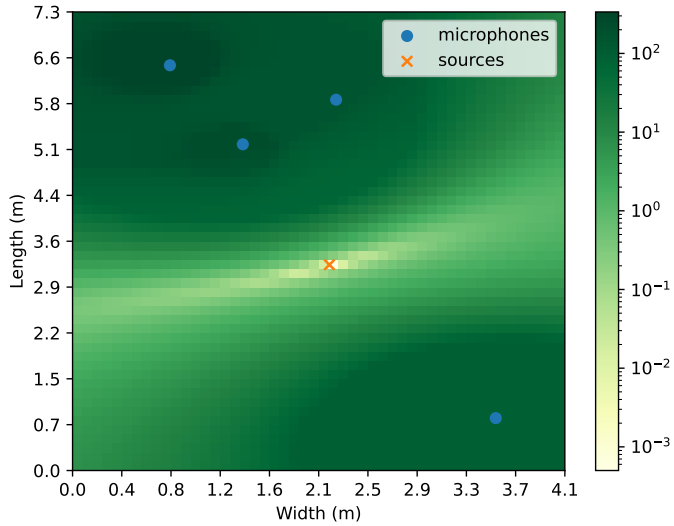

Figure 2.9: Global Energy-ratio Least-squares maps, obtained by summing the maps shown in Fig. 2.8

The energy (2.20) is related to the TOF between signals defined in (2.1): in active localization systems where the initial energy of the signal leaving the source is known, this can be used to estimate the attenuation caused by propagation, and therefore the distance between source and microphone. This information is, however, unknown in many practical applications. In this case, a relative energy measure, similar to the TDOA between microphone pairs, can be developed, named the *energy ratio* between signals, equal to

$$k_{lm} = \frac{E(\mathbf{x}_l^2(t))}{E(\mathbf{x}_m^2(t))} = \frac{\|\mathbf{u}_1 - \mathbf{v}_l\|}{\|\mathbf{u} - \mathbf{v}_m\|} \tag{2.21}$$

for a single source located at $\mathbf{u}_1$. While the left side of (2.21) is measured, the right side is theoretically defined, allowing for an estimator to be defined analogously to the LS delay-based method defined in Sec. 2.1.6. Similarly to Sec. 2.1.6, an error metric between estimated and theoretical energy ratios can be used for each pair of microphones, followed by summation across all pairs to create a global error metric.

Finally, in contrast to TDOAs defining a hyperboloid of candidate locations for each pair of microphones, an energy ratio defines a spheroid of candidate locations in space, or a circle on a plane, as shown by the pairwise error grids in Fig. 2.8. The circle's centre

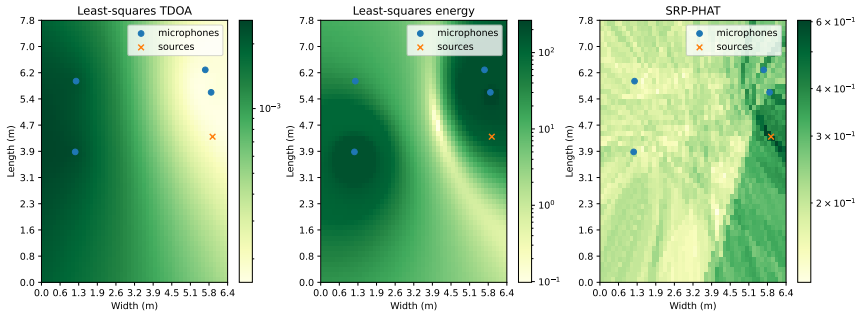

Figure 2.10: Comparison of the Energy- and TDOA-based LS methods and SRP on a reverberant environment

and radius are theoretically defined as [77], [78]

$$c_{lm} = \frac{\mathbf{u}_l - k_{lm}^2 \mathbf{u}_m}{1 - k_{lm}^2} \tag{2.22}$$

and

$$r_{lm} = \frac{k_{lm} \|\mathbf{u}_l - \mathbf{u}_m\|^2}{1 - k_{lm}^2}. \tag{2.23}$$

Proofs for (2.22) and (2.23) are however not presented in the literature. A minor contribution of this thesis is the deriation of these equations, shown in Appendix A. The resulting sum of the error grids shown in Fig. 2.8 is shown in Fig. 2.9. Note that these maps were generated in an anechoic environemnt. Energy-based methods become unreliable in the presence of reverberation, as shown in Fig. 2.10.

2.3 Neural-based methods

In recent years, deep neural networks have been widely adopted for the task of sound source localization [79]. The various approaches differ in the input features used, the network architectures and output strategies. Most studies focus on the task of DOA estimation, i.e., estimating the angle between the propagation direction of the acoustic wavefront due to the source and a reference axis of the array.

Possibilities for the input features include the raw audio samples of the microphone signals [80], their frequency-domain representation through the Short Time Fourier Trans-

form (STFT) [11], [81]–[83], or cross-correlation [84]. Multiple architectures have been also tested, including the Multi-layer Perceptron (MLP) [84], Convolutional Neural Networks (CNNs) [81] and residual networks [85]. A prominent architecture is the Convolutional Recurrent Neural Network (CRNN), which has received widespread adoption in the field [83], [86], [87], and was frequently adopted in the original contributions of this thesis. Finally, approaches differ in terms of the network's output strategy. While regression-based approaches directly estimate the source's coordinates, classification based-approaches discretize the source locations to a grid of available positions. We refer to [88] for a discussion on the merits of both approaches. We also refer the reader to a substantial survey of neural SSL papers [79]. If the input feature consists of the output of a classical signal processing method, such as the SRP maps shown in Fig. 2.5, the network we classify it as *hybrid*. Othervise, we shall classify it as *Time/Frequency (T/F)*.

A particularly relevant baseline for the work presented on this thesis is the Cross3D method proposed by Diaz-Guerra *et al.* [89] for the application of single-source DOA tracking. Their method can be interpreted as an image processing network, where its input is the 2D power map produced by the SRP method. The model's name is due to its architecture being a 3-dimensional causal CNN, where the three dimensions are azimuth, elevation and time. The authors show that the model can be trained on simulated data generated using the image source method [20] and tested on a realistic dataset of real recordings. Recent work by the authors [90] modified the neural network architecture and feature extractor, significantly reducing the computational cost of Cross3D. Multi-source capabilities were also recently introduced in [91].

A second relevant baseline is the DOANet method proposed by Adavanne *et al.* [17] for tracking up to two simultaneous sound events. The main model used is a bidirectional CRNN [92]. The authors show that including tracking metrics defined in [16] significantly improved the model's performance. The output of the network consists of a vector of size 8, where the first 6 elements refer to the estimated source positions, and the last two represent the activity of each track, similar to a Voice Activity Detector (VAD).

2.4 Subspace methods

The final class of relevant localization methods operates by applying eigenanalyses to the received microphone signals to estimate the source. The most important methods are Multiple Signal Classification (MUSIC) [93] and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [94]. The remaining parts of this section describe the former method, and provides a small discussion on why it is not considered further in this thesis, and was only included here for the sake on completeness.

2.4.1 MUSIC

MUSIC was introduced by Schmidt [93] and is a widely adopted algorithm for SSL among other tasks such as source counting and spectral estimation [93]. MUSIC assumes the signals to be *narrowband* with a carrier frequency w, although a broadband modification of MUSIC was proposed in [95]. Narrowband signals can be obtained from a broadband speech signal through the application of the STFT to each microphone signal and treating each frequency band independently. The signal $x_m(t)$ received at each microphone m is modeled as as a sum of the N active sources,

$$x_m(t) = \sum_{n=1}^{N} s_n(t) a_m(\mathbf{u}_n) e^{-jw\tau_m(\mathbf{u}_n)} + \epsilon_m(t). \tag{2.24}$$

The same noise variance σ^2 is assumed for all microphones. The signal vector $\mathbf{x}(t)$ of size M received by all microphones can be modeled in matrix form as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \epsilon(t), \tag{2.25}$$

where $\mathbf{A}$ is a *steering matrix* of dimensions $M \times N$. Each source has an associated column $\mathbf{a}_n$ in $\mathbf{A}$ defined as the *steering vector* which maps the signal $s_n(t)$ sent from source n to all microphones. The resulting signal sensed at the microphones is a combination of the steered signals plus the sensor noise $\epsilon(t)$. Ignoring the noise, the received signal vector can be interpreted as a *linear combination* between the columns of $\mathbf{A}$. In other words, the source signals are mapped to a vector subspace defined by the columns of $\mathbf{A}$. MUSIC explores the *covariance matrix* between the microphone channels, which is expressed as

$\mathbf{R}_x = \mathbb{E}(\mathbf{x}\mathbf{x}^H)$, where $\mathbb{E}$ is the expected value operator and H is the Herimtian transpose operator. Substituting x by its definition in (2.25), expanding the results, and assuming independence between signal and noise, $\mathbf{R}_x$ can be expressed as

$$\mathbf{R}_x = \mathbf{A}\mathbf{R}_s\mathbf{A}^H + \sigma^2\mathbf{I}, \tag{2.26}$$

where $\mathbf{R}_s$ is the covariance matrix of $\mathbf{s}$ and $\sigma^2\mathbf{I}$ is the identity matrix of size $M \times M$. $\mathbf{A}\mathbf{R}_s\mathbf{A}^H$ and $\mathbf{I}$ are Hermitian matrices as they satisfy $\mathbf{R}_x^{ij} = [\mathbf{R}_x^{ji}]^* \ \forall(i,j)$, where $[]^*$ represents complex conjugation. Hermitian matrices can be decomposed into a diagonal form using an orthonormal basis, that is,

$$\mathbf{R}_x = \mathbf{E}^H\mathbf{\Lambda}\mathbf{E}, \tag{2.27}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing real non-negative numbers, corresponding to the eigenvalues which we assume to be in ascending order. Conversely, the columns of $\mathbf{u}$ correspond to the associated eigenvectors. As $\mathbf{A}\mathbf{R}_s\mathbf{A}^H$ is generated by matrix $\mathbf{A}$ of dimensions $M \times N$, it has N non-negative eigenvalues, and $M - N$ zero-valued eigenvalues. By summing it to $\sigma^2\mathbf{I}$, which has M eigenvalues equal to σ^2, $\mathbf{R}_x$ has $M - N$ eigenvalues equal to σ^2.

Given the eigenvectors $\mathbf{e}_m$, i.e. the columns of $\mathbf{E}$, and ordering them in decreasing order with respect to their associated eigenvalues, we have the first $M - N$ eigenvalues associated with sensor noise, which we know to be independent from the signal. MUSIC is able to estimate the source locations by searching for N candidate steering vectors which minimize the sum of the dot products with the noise-associated eigenvalues:

$$\arg\min_{\mathbf{u}} \sum_{i=1}^{M-N} |\mathbf{a}(\mathbf{u})\mathbf{e}_i|, \tag{2.28}$$

where $\mathbf{a}(\mathbf{u})$ is a candidate steering vector for a direction $\mathbf{u}$.

Although MUSIC has been shown to offer competitive performance to other established SSL methods such as SRP in ideal conditions, it requires significantly more computational power due to its eigenanalysis procedure. Also, MUSIC assumes the speech signal is more powerful than noise at each frequency bin in the spectrogram, an assump-

tion which is seldom guaranteed in practical scenarios [96]. Finally, MUSIC requires noise statistics to be assumed or measured, which are often also difficult to achieve in practice. For those reasons, MUSIC is not further explored in this work.

2.5 Multi-source SRP approaches

This section starts by revisiting the problem statement described in Sec. 2.1.2. Instead of defining the target output of our system as a single source position vector $\mathbf{u}$, it is extended to be a matrix $\mathbf{U}$ of dimensions $3 \times N$, defined as

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_N \end{bmatrix}, \tag{2.29}$$

where N is the number of active sources. Note that N is usually unknown in practice, and must also be estimated on such cases. Updating the model for the signal received at each microphone is also required, as it becomes a weighted sum of all active sources. In the frequency domain, the received signal at microphone m can be described as

$$\bar{x}_m(t,f) = \sum_{n=1}^{N} s_n(t,f) a_m(\mathbf{u}_n, f) e^{-jf\tau_m(\mathbf{u}_n)} + \epsilon_m(t,f). \tag{2.30}$$

Despite the modified signal model, the analysis of the CC function between two microphone signal frames $\mathbf{x}_l$ and $\mathbf{x}_m$ in the presence of N simultaneous talkers usually presents one peak related to each source. Although this would allow the conventional SRP method to be used directly, the function may also exhibit 'ghost peaks' related to the reflections caused by the room's surfaces, hindering the estimation procedure. Also, the relative amplitude of peaks may vary considerably, especially in cases where the sources have different power levels, hindering the application of simple thresholding methods. Finally, the interfering sources reduce the correlation amplitudes at delays $\tau_m(\mathbf{u}_n)$ are reduced in comparison to the single source case, hindering the analysis of the SRP map.

Due to the aforementioned limitations of using the conventional SRP method for multi-source localization, different SRP-based alternatives have been proposed. These alternative methods, while presenting their own particularities in terms of implementation, target scenario and performance, are categorized in the following subsections based on their core modification when compared to the conventional SRP method.

2.5.1 Modified SRP computation

Different strategies have been proposed where the computation of the SRP map is modified in order to allow for better localization of multiple sources. For instance, in [97], an alternative to the conventional PHAT-weighting function is proposed, aiming to achieve flexibility in combining different narrowband components. Simulation results, obtained for both single and multi-source cases, indicate that the use of the modified PHAT-weighting function can improve localization performance for both narrowband and broadband signals.

In [98], similarly to the efforts aimed at achieving an improved combination of pair-wise information for increasing localization robustness outlined in Sec. 6.3.2, the use of harmonic and geometric means of the GCC functions over all available microphone pairs was explored to build an acoustic map. When compared to the conventional summation of pair-wise functions, as previously expressed in (2.9), results show that the use of geometric and harmonic means contributes to removing undesired sidelobes and improving source level estimation.

2.5.2 Source cancellation

Many multi-source, SRP-based methods rely on schemes that reduce the influence of a previously located and dominant source on newly computed SRP maps. For instance, in [99], the localization of two sources is performed in a two-step manner. First, the position of the source with the highest correlation peak is estimated as in the conventional SRP method. To estimate the second source, the first source is de-emphasized from the CC function through the use of a TDOA-domain notch filter. This process is illustrated in Fig. 2.11. Although this approach can be further applied for the localization of three sources, the authors state that the noise in the correlation function with three sources would be prohibitive, and that tracking approaches should be applied instead.

The removal of a previously located source's contribution from an SRP map can also be achieved through the projection of the observed GCCs onto a subspace that is orthogonal to the source position, as described in [100]. Results obtained with both simulated and experimental data indicate that such an approach can outperform the de-

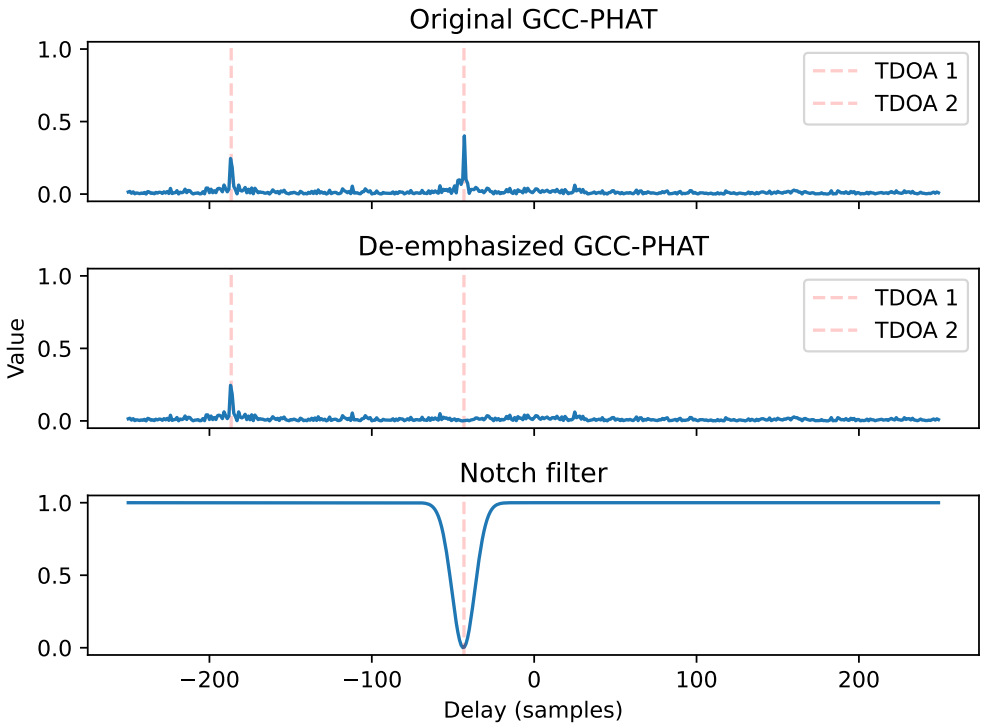

Figure 2.11: Representation of the de-emphasis procedure described by Brutti et al. [99].

emphasis method from [99], especially in cases of sources with different power levels. Moreover, the use of a truncated formulation of the proposed source cancellation scheme allows for a reduction in computational cost while performing comparably to [99], without requiring parameter tuning associated to the TDOA-domain notch filter design.

Subspace processing for source cancellation within an SRP-based framework has also been proposed in [101], where the SVD-PHAT method [102] is extended to address the case of multiple sources. Thererin, the contribution of a previously located source (obtained by means of a k-d tree search) is removed from the observed projections of the GCCs onto a reduced-dimensional subspace. The proposed multi-source SVD-PHAT approach was compared to a source cancellation scheme, similar to the de-emphasis method from [99], where a source's contribution is removed from the observed GCCs and a new SRP-PHAT map is computed for locating the next source. Simulation results indicate that the multi-source SVD-PHAT can outperform the successive recomputation of the SRP-PHAT map.

As an alternative to employing a source cancellation procedure to the observed GCCs, the spatial gradient SRP-PHAT method proposed in [103] involves successively removing the influence of the current, most dominant source directly in the observed SRP map by means of a negative spatial gradient function. Experimental results for two-speaker scenarios show that the spatial gradient SRP-PHAT can be an effective localization method in scenarios with a diffuse noise field.

In [104], an approximate analytical formulation of an SRP map using a Gaussian Mixture Model is proposed, such that probability density functions can be used to estimate the location of multiple sources while removing their corresponding contributions from the probabilistic SRP map. Experimental results with scenarios involving up to three speakers indicate that while this approach can effectively locate multiple sources, its performance degrades when sources differ greatly in power.

2.5.3 Grid refinement

In [105], grid refinement is indirectly used to localize multiple sources by identifying different zones of interest, defined in terms of TDOA intervals, as those where acoustic sources are dominant in terms of a cumulative SRP function. In this way, a conventional grid

search step for source localization can be performed over a reduced search space with the desired spatial resolution. The localization of multiple sources can then be achieved by iteratively removing the influence of the dominant sources [104]. Experimental results show that such approach can improve localization performance in multi-source scenarios at a lower computational cost than the authors' previous work [104].

In [106], a hierarchical search-grid refinement method is proposed, where a probability measure of a sound source's presence in different regions, formulated as a spatially averaged SRP map, is used to identify the limited set of steering directions for which the search grid resolution can then be improved for localizing multiple sources. This approach is shown to lower the computational cost while performing similarly to the conventional SRP method that employs the highest resolution level over the entire search space.

2.5.4 Clustering and distance analysis

Another concept often exploited in multi-source localization methods relates to data clustering and analyzing distances between multiple source location estimates. For instance, the sources' preliminary location estimates can be obtained through the conventional SRP method. Then, spatial clustering can be employed to track the estimated locations of multiple sources over different time frames [107]. Alternatively, a narrowband SRP formulation can be employed to obtain location estimates per frequency bin and time frame, while Gaussian mixture modeling can then be used to cluster the location estimates [108]. Furthermore, both the location and activity of multiple sources can be tracked [108].

In [109], source location estimates are obtained by using SRP-PHAT combined with agglomerative spatial clustering and SRC (cf. Sec. 6.2.2). Experimental results show that the localization performance of the proposed approach degrades when the peaks to be identified have widely different amplitudes or are closely located in the CC function. Accordingly, the proposed approach is further extended in [110], by replacing the agglomerative clustering step from [109] with Gaussian mixture modeling of the observed SRP map, or by identifying the peaks in the SRP map while assuming a minimal distance between sources. The performance limitations first demonstrated in [109] are also addressed in [111], where the localization of multiple speech sources is achieved by computing subband SRP maps, estimating the dominant source's position for each subband, and employing

agglomerative clustering across all subbands to obtain the final set of source location estimates. In [112], a method named Multi-Stage Rejection Sampling (MSRS) is proposed, which involves spatially clustering probability density points, derived as a function of the observed SRP-PHAT map, to identify regions of interest. Then, volume contraction is used in the identified regions for localizing multiple sources.

In [113], a three-step framework is proposed for multiple source localization. It relies on: step 1) partitioning the search region into cubic volumes, clustering such volumes and, based on equivalent TDOA bounds; step 2) computing a delay density map to find in which clusters it is more likely to have a sound source; step 3) further analyzing the chosen clusters with conventional SRP to obtain the final source location estimates.

Finally, the approach proposed in [114] for a specific microphone setup of central and lateral microphone arrays, involves finding the intersection between the source positions estimated with the central array's SRP map and the ones estimated with the lateral arrays through an adaptive subband generalized eigenvalue decomposition (GEVD) scheme, in order to obtain the final 3D location estimates of multiple sources. Simulation results with up to three speakers demonstrate that the proposed method outperforms other state-of-the-art methods under varying levels of noise and reverberation.

2.5.5 Sparsity assumptions

Sparsity-based modifications to the conventional SRP method have also been proposed for the task of multiple source localization. For instance, by assuming a limited number of active sources with respect to the search grid of candidate locations, localization can be performed by employing a sparse-regularized generative model that fits the observed SRP map, combined with a subspace filtering step that compensates for what is not directly accounted for by the fitted model [115]. Experimental results show that although the use of this approach can outperform the conventional SRP-PHAT in the multi-source scenarios tested, its overall performance highly depends on the choice of the hyperparameters used in the proposed problem formulation.

Alternatively, in [116], it is shown that group sparsity can be exploited when modeling an observed broadband SRP map as a linear function of power spectral densities (PSDs), related to an overcomplete set of candidate locations. Hence, multi-source localization

can be achieved by solving a group-sparse optimization problem and identifying peaks in the estimated PSDs. Simulation results obtained for two-speaker scenarios show that the proposed method performs better than or similar to the conventional SRP-PHAT method for varying levels of noise and reverberation, while overall outperforming the frequency-domain sparse iterative covariance-based estimation (SPICE) [117], [118] method.

In [119], the authors exploit time-frequency sparsity, by assuming that only one speech source is dominant in a given time-frequency bin, i.e., they are assumed to be W-disjoint [120]. By analyzing each frequency bin and performing single-source localization, histograms with all individual DOA estimates can be generated and used in a matching-pursuit-based step to perform multiple source localization. Simulation and experimental results indicate that this approach can outperform other state-of-the-art multi-source localization methods, at a lower computational cost. The sparsity of speech signals in the time-frequency domain is similarly exploited in [121], where a weighted, wideband histogram of source locations is computed based on narrowband DOA estimates, obtained with SRP-PHAT applied to different frequencies and observation frames. The weighted histogram is then used to perform multiple source localization through peak detection, and simulation results indicate the advantage of the proposed method when compared to the wideband SRP-PHAT for two-speaker scenarios in reverberant environments.

In [122], sparse modeling of the GCCs observed by a microphone array is employed in the task of localizing sound sources and their corresponding acoustic reflections. A linear inverse problem is proposed to be solved, with its formulation depending on a time-domain propagation matrix. The authors present two implementations of the proposed method, with the first based on orthogonal match pursuit (OMP) [123], and the second on the truncated Newton interior-point method [124]. It is demonstrated through an experimental study that the use of sparsity constraints in the solution of the proposed linear inverse problem contributes to better location estimates when compared to the direct use of a time-domain SRP map. The choice of propagation matrix used for formulating the linear inverse problem presented in [122] was further investigated in [125], where the influence of the temporal width threshold, associated to the determination of propagation matrix coefficients, is demonstrated. Additionally, when assuming the GCC coefficients to be PHAT-weighted, an alternative formulation of the propagation matrix circumventing such temporal width threshold is proposed, with experimental results indicating the advantage

of using such alternative formulation in terms of computational time.

Finally, in [126], an SRP-based method is proposed for simultaneous multiple source localization that employs Non-negative Matrix Factorization (NMF) [127] to decompose the time-frequency signal into a weighted sum of broadband atoms, which are time-frame-dependent and correspond to different groupings of frequency bands related to distinct sources. This method, named SRP-NMF, attempts to combine the advantages of both narrowband and broadband approaches that exploit sparsity in their corresponding domains, and experimental results indicate it performs better than or similarly to state-of-the-art methods based on fully broadband or narrowband signal formulations.

2.6 SRP: practical considerations

This section further analyzes practical aspects of the SRP method, such as tracking moving sources and incorporating source/microphone directivity.

2.6.1 Tracking moving sources

Although a source may remain mostly stationary in many scenarios such as conference calls, the same cannot be said for many situations in surveillance, robotics and healthcare. It is therefore reasonable to reformulate the source position $\mathbf{u}$ to be time-dependent, i.e., $\mathbf{u}(t)$. The task of estimating a source's position at multiple time instants is hereafter referred to as tracking.

A straightforward way to achieve tracking using conventional SRP is to compute an SRP map and estimate the source position independently for successive frames at times t_{i-1} and t_i. This estimate can be often improved through the incorporation of a state-space dynamic model as well as previous estimates $\{\hat{\mathbf{u}}(t_{i-1}) \quad \hat{\mathbf{u}}(t_{i-2})\ldots\}$. Such a state-space model provides source tracking by introducing dynamic constraints into the source localization procedure, modeling for instance the speed of the source. This allows for smoother position estimates to be produced and for unreliable observations, such as those caused by reverberation and noise, to be properly identified and handled.

The most common approaches for source tracking using SRP are Kalman filters [107], [128]–[130], particle filters [129], [131]–[137] and deep neural networks [89], [90], [138],

[139]. Unlike in neural methods, the state-space model is explicitly defined in Kalman and particle filters.

Particle filters are frequently preferred over Kalman filters due to their simpler formulation and ability to model non-linear systems. Particle filters model the source location with the help of Q candidate positions known as particles, each having an associated likelihood or weight π_q, $q = 1, \ldots, Q$. The estimated source location is obtained as a weighted sum of the particles, where the weights are their respective likelihood. At each iteration, the particles are updated according to a given kinematic model. Optionally, a resampling process may be also applied to reduce the variance of the particles.

The movement of a source at consecutive time steps is commonly modeled using Langevin dynamics [131], [133]–[136], [140], which assume that the source moves independently in each direction. The relationship between the source's position at times t_{i-1} and t_i is equal to [140]

$$\mathbf{u}(t_i) = \mathbf{u}(t_{i-1}) + \dot{\mathbf{u}}(t_i)\Delta t, \tag{2.31}$$

where $\Delta t = (t_{i-1} - t_i)/f_s$, and $\dot{\mathbf{u}}(t_i)$ is the source's velocity, modelled as

$$\dot{\mathbf{u}}(t_i) = a^{(1)}\dot{\mathbf{u}}(t_{i-1}) + b^{(1)}F^{(1)}. \tag{2.32}$$

In (2.32), $F^{(1)} = N(0,1)$, $a^{(1)} = e^{-\alpha^{(1)}\Delta t}$ and $b^{(1)} = \beta^{(1)}\sqrt{1 - a^{(1)}}$ are known as the damping and excitation parameters, respectively responsible for controlling the inertia and innovation of the movement in each direction. $\alpha^{(1)}$, $\beta^{(1)}$ are hyperparameters, to be chosen or tuned.

2.6.2 Directional sources and microphones

The SRP signal model can be modified for the case where sources and/or microphones exhibit directional acoustic behaviour, that is, the amplitudes of the microphone signals are dependent on the orientation of microphones and sources. The directivity profile for microphone m is defined as a function $0 < d_m^{(1)}(\theta_m) \leq 1$, where θ_m is an angle. An analogous function can also be defined for the source's directivity $d^{(2)}(\theta_s)$. Finally, the angles θ_1, θ_2, θ_3 and θ_4 are defined as the angles of departure, the source direction, the angle of arrival and microphone direction respectively. The attenuation term defined in

(1.3) can then be specified as [141]

$$a_m = d_m^{(1)}(\theta_1 - \theta_2)d^{(2)}(\theta_3 - \theta_4)\frac{k_d}{\|\mathbf{u} - \mathbf{v}_m\|}, \tag{2.33}$$

where $\frac{k_d}{\|\mathbf{u}-\mathbf{v}_m\|}$ represents the attenuation caused by propagation, which generally follows an inverse law. In practice, this attenuation can be incorporated into SRP by including the source's candidate orientation as another search dimension [141]. Note that the gains between microphones must be assumed to be calibrated, and that the source and microphone directivity patterns, as well as the microphone orientations, must be known or assumed. Microphone directivity can also be exploited to reduce the number of microphone pairs and region size used for SRP [130], [142].

When operating with distributed microphone arrays, source directivity can be estimated in two steps, firstly by estimating the source position, followed by the creation of a spherical grid around the source. The point with the highest SRP value on the sphere is selected to represent the source's orientation [143], [144]. A similar approach is applied in [145]–[147], which assumes that the arrays directly facing the speaker will exhibit an SRP map with a sharp peak. The sharpness is measured using the map's kurtosis, which is then used to estimate the talkers orientation. If the microphone gains are calibrated, the GCC-PHAT's peak values can be used for comparison instead of the kurtosis [148].

2.6.3 Comparing SRP to other approaches

In [76], a theoretical comparison between the SRP and the Maximum Likelihood Sound Source Localization (ML-SSL) method is made. The functioning of the ML-SSL method is similar to SRP, as the source location is also estimated as the maximum argument of a likelihood function. However, the ML-SSL method explicitly models the noise received at each microphone as well as its correlation with other microphones. Although this can be advantageous, allowing microphone signals exhibiting large noise to be ignored, it requires noise statistics to be assumed or measured. The aforementioned paper starts by defining the ML-SSL signal model similarly to (1.1) directly in the frequency domain. The authors assume that reverberation is independent across microphone signals, and that microphones boast a high Signal-to-Noise Ratio (SNR). The paper shows that, under these assumptions, the ML-SSL method becomes independent of noise and reverberation,

and equal to the SRP formulation, and uses this proof to justify why SRP works well under low-noise, reverberant environments.

In [149], the authors conducted a performance analysis of several GCC-PHAT-based algorithms for a large-aperture microphone array. They presented a real-time source localization algorithm based on TDOAs derived from a phase transform applied to the generalized cross-power spectrum. The algorithm is then enhanced by preprocessing the data using local beamformers. A comparison was made by testing these two algorithms and the SRP-PHAT in an environment in which the microphone signal-to-reverberation ratio was in the range [-2 dB, -12 dB] and the signal-to-background noise ratio with flat frequency weighting in the band 80 Hz to 10 kHz was in the range [-6 dB, -16 dB]. They found that SRP-PHAT provides reliable location estimates under adverse conditions but has a larger computational cost.

In [150], the authors developed two GCC methods based on time delay estimation using classical CC and smoothed coherence transform algorithms. They analyzed the performance of the aforementioned GCC-based algorithms for multi-source, point-based localization by comparing them with the existing FASTTDE, GCC-PHAT and FAST SRP-PHAT [151]. The methods were evaluated in terms of 1) localization accuracy, 2) detection accuracy, and 3) computational cost. The localization accuracy of the FAST SRP-PHAT was much higher than that of the other three methods. In terms of detection, the other three methods exhibited higher localization performance. Finally, it was shown that SRP-PHAT had a higher computational cost than other methods.

In [152], a comparison between the ROOT-MUSIC algorithm and SRP-PHAT is made. The root mean square error (RMSE) is used as a performance metric and the results show that even though ROOT-MUSIC is more computationally efficient, SRP-PHAT exhibits superior performance in challenging conditions, such as environments with reverberation and low SNR.

In [153], [154], the authors evaluate the performance of broadband spatio-spectral estimators, including SRP and two-step localization methods. They perform an eigen-analysis of the parameterized spatial correlation matrix and show that the attenuation can be estimated from this matrix. They propose a DOA estimator based on MCCC and show that this method yields a higher resolution than the conventional SRP. DOA estima-

tion performance is similar in anechoic environments and environments with reverberation time of 300 ms, however it is worse in environments with reverberation time of 600 ms compared to SRP.

In [155], the authors evaluate the performance of a multiple source localization method based on Augmented Intensity Vectors (AIV) using spherical microphone arrays. Their simulations comprising various angular separation and reverberation times show that AIV has an average accuracy between 5 and 10 degrees for sources with angular separation of 30 degrees or more and it performs better than the methods using Pseudo-Intensity Vectors and SRP. Another finding is that a plane wave decomposition-based SRP method cannot localize all sources if the number of sources is three or more and if they are separated less than 45 degrees.

In [26], SRP-PHAT is used along with Hidden Markov Models (HMMs) and face tracking for voice activity detection and localization. Results show that using HMMs along with SRP-PHAT increases the accuracy and utilising face tracking in addition yields even better results.

In [156], the authors compared two algorithms [157], [158], which are also explained in Sec. 6.2, in both simulated and real-world scenarios in which speakers were recorded by eight microphones spread out on the wall and ceiling, and concluded that hybrid localization is more robust than hierarchical localization and is computationally faster. Especially when the reverberation time is greater than 300 ms, the localization error in hierarchical localization increases at a much higher rate.

In [159], a general framework for the integration of microphone signals for SSL is presented. A Spatial Observability Function (SOF), which is the mean square difference between the Spatial Likelihood Function (SLF) and the true probability of an object, is used as an indication of the accuracy of the map. SRP-PHAT is a special case of this method in which the SLF from each array is integrated without taking the SOFs into account.

2.6.4 Analyses of SRP

The authors of [160] propose an analytical model based on sound propagation and its interaction with the environment that predicts SRP maps in both anechoic and non-

anechoic conditions, and under both far- and near-field assumptions. They investigate how and to what extent the signal bandwidth, array topology, room geometry and spectral content of the signal affect SRP maps. The findings show that SRP functions depend on the array topology, room geometry and signal bandwidth but not on the spectral content of the signal. They validate their model by comparing it with the true SRP maps.

In [161], the authors investigated the geometrical sensor calibration errors in a Uniform Linear Array (ULA) used in far-field human speech source localization. They observed that the errors in configuration of the endpoint sensors result in larger localization errors compared with same configuration errors of the inner sensors. In addition, they show that the localization errors increase when the total configuration error is above a threshold related to the propagation distance and the system's sampling rate.

In [162], the authors proposed an SRP constraint to suppress local extrema. They weighted the SRP function using a coherence factor, determined by observing the signs of the GCCs between all possible microphone pairs. If the sign was the same for all microphone pairs, this indicated a high coherence and the coherence factor is 1. If half of them were negative and half were positive, then were deemed as incoherent, and assigned a coherence factor of 0. The method was shown to operate without loss of localization accuracy with respect to conventional SRP.

Chapter 3

Dual-Input Neural Networks for SSL

3.1 Introduction

Most signals, such as audio and images, contain metadata. Metadata can be signal-based, which describes quantitative properties of the signal, such as its sampling rate, as well as semantic, which describes, for example, contextual properties. In speech processing, semantic metadata could consist of the speaker's language or gender. Whether signal-based or semantic, including metadata as a secondary input into neural network models may provide relevant information which would translate into an economy of training time, model parameters and flexibility. However, metadata typically has a different rank than the input signals, making its incorporation into those models not trivial.

The main focus of this chapter is to study the effectiveness of schemes to process signals and exploit metadata jointly using neural network models. This chapter focuses on the task of Sound Source Localization (SSL) using distributed microphone arrays to demonstrate the effectiveness of the proposed approach. In the context of SSL, relevant metadata which is exploited by classical methods is the microphone positions, which can be acquired by manual measurement or using self-calibration [163] methods. Other relevant metadata is the room dimensions and its reverberation time.

SSL approaches may be divided into classical signal processing-based and data-driven

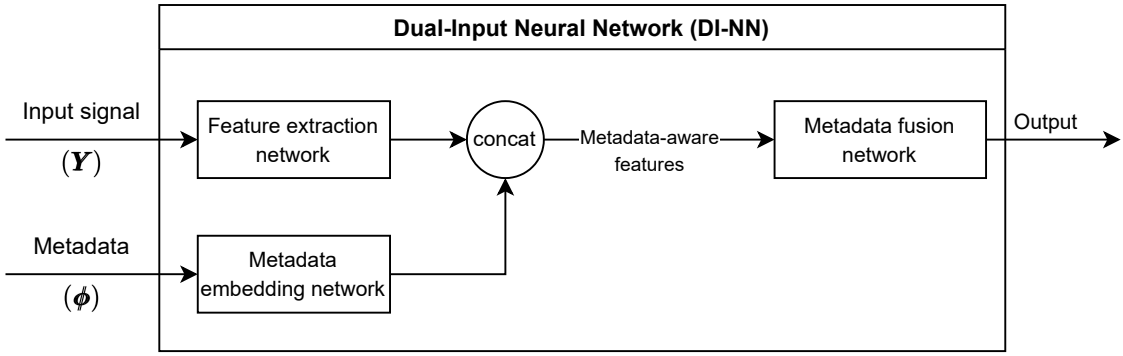

Figure 3.1: Overview of the DI-NN approach. Note that the Metadata embedding network is an optional block.

neural network-based methods. By explicitly exploiting metadata describing microphone positions and room dimensions, classical approaches may be applied to different rooms and microphone configurations. Conversely, neural network approaches have recently achieved state of the art results for source localization [80], [84], [86], at the expense of requiring one network to be trained for every microphone topology. One reason current neural approaches do not incorporate the microphones' positional information is that the microphones' signal and positional data are very different from one another in nature and rank.

Previous work which discusses the joint processing of signals and metadata is [164], where a single input neural network is used to process metadata in conjunction with a low-rank physical signal. However, unlike the proposed approach, the method of [164] is restricted to multilayer perceptron architectures and vector input and metadata, limiting its application in practical scenarios.

Another related field is multimodal fusion [165], [166], although this is usually concerned with learning representations using two types of signals, such as audio-visual data. Simultaneously processing signals and metadata have also been explored using non-neural models for sound source separation [167], where metadata consists of information about the type of sound (speech, music) and how the sources were mixed. However, none of the existing work discusses effective schemes for incorporating and evaluating signals and metadata of different rank.

The main contribution of this chapter is the DI-NN neural network architecture, which is capable of processing high-dimensional signals, namely spectrograms, along with a relevant metadata vector of lower rank (i.e., 1). An overview diagram of the proposed approach is shown in Fig. 3.1, which will be discussed in Sec. 3.2.2. The proposed method is compared to three baselines for the task of Positional Sound Source Localization (PSSL), namely, a metadata-unaware Convolutional Recurrent Neural Network (CRNN), a metadata-aware classical signal processing approach, as well as an alternate metadata-aware neural network. The proposed method is able to outperform all baselines by a large margin in realistic scenarios. In contrast to previous approaches [80], [168], DI-NN dispenses with the need for training a network for each scenario, broadening the method's applicability.

This chapter continues as follows. The approach for training the proposed DI-NN for SSL is described together with several baseline methods in Sec. 3.2. In Sec. 3.3, the experiments comparing DI-NN approach with the baselines using multiple datasets are described. Finally, results and conclusions are drawn in Sec. 3.4.

3.2 Method

3.2.1 Scope of this work

This work is restricted to the localization of a static source at the planar coordinates $\mathbf{u} = [u^{(1)}, u^{(2)}]^T$. The source emits an intermittent signal $s(t)$ at time t. In the experiments, $s(t)$ may consist of White Gaussian Noise (WGN) as well as of speech utterances. Also, M static microphones with known positions are present in the room, each placed at coordinates $\mathbf{v}_m = [v_m^{(1)}, v_m^{(2)}]^T$. Both source and microphones are enclosed in a room of planar dimensions $\mathbf{d} = [d^{(1)}, d^{(2)}]^T$. The amount of reverberation in the room is modeled by its reverberation time r, a measure of the amount of time it takes for a sound to decay by 60 dB from its original level. The gains between the microphones are assumed to be approximately calibrated, although experiments in Sec. 3.3.3 show that DI-NN is robust to uncalibrated microphones of the same kind.

Finally, the *metadata vector* $\phi \in \mathbb{R}^{N_\phi}$ is the concatenation of the coordinates of the microphones, the room dimensions and reverberation time, as shown in Fig. 3.2. The

aforementioned types of metadata are chosen as they are typically explicitly exploited in classical localization methods such as the Least Squares (LS).

3.2.2 Proposed method: Dual input neural network

The proposed DI-NN architecture is comprised of two neural networks, a **feature extraction network** and a **metadata fusion network** as can be seen in Fig. 3.1. An additional third network, called the **metadata embedding network** is also used in the alternative DI-NN-Embedding network, which will be presented in Sec. 3.2.3 .

The input of the network consists of the Short Time Fourier Transform (STFT) of the microphone signals as defined in Sec. 3.2.1. Instead of using the complex representation generated by the STFT, the real and imaginary parts of the STFT $\mathbf{Y}$ are used as separate channels as in [83], giving rise to $2M$ input channels. The role of the feature extraction network is to transform this high rank tensor into a feature vector which compactly represents relevant information for the task in hand. In the experiments, a CRNN [92] is adopted as the feature extraction network, due to its wide adoption for SSL [86], [87], [169].

This metadata-unaware vector is then concatenated to the available metadata, thus creating a metadata-aware feature vector. For the considered application, the metadata is a vector consisting of the positions of the microphones, the dimensions of the room, and its reverberation time. This metadata-aware feature vector is then fed to a metadata fusion network, whose role is to merge the metadata and feature vector to produce the result. In the experiments, a two-layer Fully Connected Neural Network (FC-NN) which maps the metadata-aware features to a two dimensional vector corresponding to the estimated coordinates of the source is adopted.

The feature extractor CRNN is divided into two sequential sub-networks: a Convolutional Neural Network (CNN) block, responsible for extracting local patterns from the input data and a Recurrent Neural Network (RNN), responsible for combining these pattens into global, time-independent features. A diagram representing the components of the DI-NN network is shown in Fig. 3.2.

The convolutional block receives a tensor of shape (M, L, F) representing a multichannel complex STFT, where M represents the number of audio channels, L represents

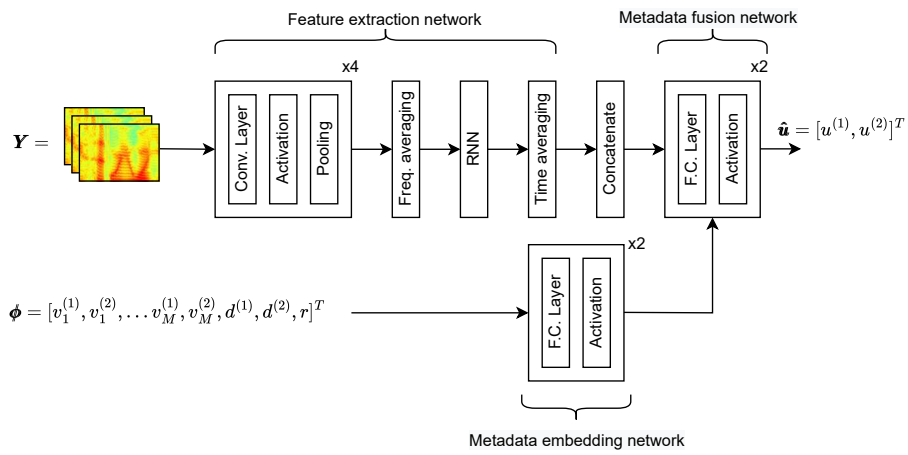

Figure 3.2: Detailed DI-NN architecture for the task of PSSL.

the number of time frames generated by the STFT, and F is the number of frequency bins used. The role of this block is two-fold: firstly, to combine local information across all microphone channels, and secondly to reduce the rank of the data to make it more tractable for the RNN layer.

The convolutional block consists of four sequential layers, where each performs three sequential operations. Firstly, a set of K convolutional filters is applied to the input signal, resulting in K output channels. Secondly, a non-linear activation function is applied to the result. Finally, an average pooling operation is applied to the width and height of the activations, generating an output of reduced size. After passing the input through the four convolutional layers, global average pooling is performed across all frequencies, generating a matrix.

After the convolutional block, the resulting matrix serves as input to a bidirectional, gated recurrent unit neural network (GRU-RNN) [170]. As sound may not be present throughout the whole duration of the audio signal, such as during speech pauses, the RNN is important for propagating location information to silent time-steps. After this network, rank of the features are once again reduced by performing average pooling on the temporal axis, resulting in a vector of time-independent features.

The outputs of the feature extraction network are then concatenated to the available metadata and serve as input to the metadata fusion network. This network consists of a set

of two fully connected layers which map the metadata-aware features to a two-dimensional vector corresponding to the estimated cartesian coordinates of the active source. Both networks are jointly trained using the same loss function, defined as the L_2-norm or the sum of the absolute error between the network's estimate of the source coordinates $\hat{\mathbf{u}}$ and the target $\mathbf{u}$, given by

$$\mathcal{L}(\mathbf{u}, \hat{\mathbf{u}}) = \|\mathbf{u} - \hat{\mathbf{u}}\| \,. \tag{3.1}$$

The loss function was chosen empirically by comparing it to other functions, such as the L_1-norm, on the validation dataset. The more common squared error loss was also considered. Although both losses yielded similar results in the experiments, the absolute error was chosen for its easier interpretability, since it corresponds to the distance in metres between target and estimated coordinates.

3.2.3 DI-NN-Embedding

To test whether it is advantageous to process the metadata before combining it with the microphone features, a variant of the DI-NN model is also proposed, where the metadata ϕ is processed by a *metadata embedding network* to produce an embedding, which is then concatenated to the microphone features. This network is represented by the *metadata embedding network* block in Fig. 3.1.

3.3 Experimentation

This section describes the experiments with DI-NNs with three SSL datasets representing scenarios of varying difficulties. For each dataset, DI-NN is compared to two other methods. The first method is a CRNN with the same architecture but without using the available metadata, i.e., without the "Concatenate" block in Fig. 3.2. The gains of the proposed DI-NN are shown when compared to this baseline. The second comparative method is the classical LS source localization method described in Sec. 2.1.6. The experiments will be described below.

All of the experiments consisted of randomly placing one source and four microphones within a room. The height of the microphones, source and room were fixed for all experiments. For each experiment, the goal of the proposed method and baselines

was to estimate the planar coordinates of the source within the room using a one-second multichannel audio signal as well as the positions of the microphones. The training and testing samples do not overlap, and hence demonstrate the method's effectiveness for handling unseen scenes and metadata. To simulate sound propagation in a reverberant room, the image source method [20] implemented by the Pyroomacoustics Python library (MIT license) [171], was used. The neural networks were trained using PyTorch (BSD license) [172] along with the PyTorch Lightning (Apache 2.0 license) library [173]. The models were trained using a single NVIDIA P100 GPU with 16 GB of RAM memory. The configuration of the experiments is managed using the Hydra (MIT license) library [174]. Code used for generating the data and training the networks is released on GitHub [1], as well as a Kaggle notebook [2] to allow reproduction of the experiments without the need for any local software installation. The hyperparameters used for training the proposed method and baselines are shown in Table 3.1.

Table 3.1: Hyperparameters

Parameter	Value
Num. parameters (DI-NN)	3.5M
Num. conv. kernels	64, 128, 256, 512
Conv. kernel size	2x2
Conv. layer pooling size	2x2
GRU output size	256
Metadata fusion net. layer out. sizes	$512 + N_\phi$, 2
Metadata embedding layer out. sizes	$2N_\phi$, N_ϕ
Activation func. last layer	None
Activation func. other layers	Rectified Linear Unit (ReLU)
Num. Discrete Fourier Transform (DFT) bins (for STFT)	1024
DFT hop length (for STFT)	512
Input duration	0.5 secs.
Sampling rate	16kHz
Grid resolution of LS method	2 cm
Learning rate	0.0005
Batch size	32
Num. epochs	40
Batch normalization [175]	Only after conv. layers
Optimizer	Adam [176]

[1] Code: https://github.com/egrinstein/di_nn
[2] Demo notebook:
https://kaggle.com/code/egrinstein/di-nn-training-notebook

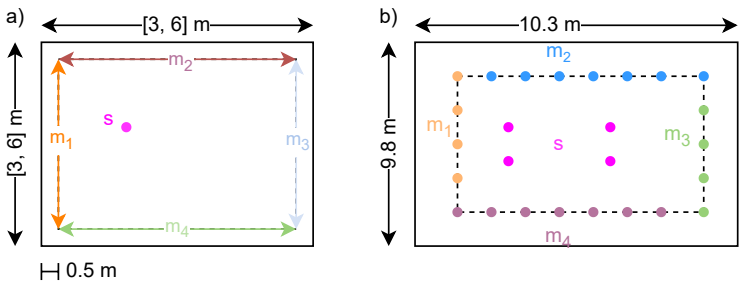

Figure 3.3: Experimental setup. (a) For the anechoic and reverberant simulations, each of the four microphones m_i is placed on a random point along the the coloured arrows, while the source s is randomly placed on a point within the rectangle defined by them. (b) The sampling procedure for Sec. 3.3.3, where positions of the microphones and source are randomly drawn from each differently coloured set of points.

3.3.1 Simulated anechoic rooms

The goal of this experiment is to evaluate the performance of the DI-NN and baselines in multiple rooms and microphone positions in the absence of reverberation. The dataset generation procedure is shown in Fig. 3.3a. For each dataset sample, two numbers are randomly selected from a uniform distribution in the interval [3, 6] m representing the room's width and length. The height of the rooms is fixed at 3 m. Next, one microphone is randomly placed along a line segment 0.5 m away and parallel to each room's walls. Microphones were placed close to the wall as a simplified localization scenario, as the main goal is to test the effectiveness of the metadata fusion procedure. Nonetheless, this scenario is realistic in the context of smart rooms, where the microphones are usually placed in or near the room's walls.

Finally, the source is randomly placed in the room, following a uniform distribution while respecting a minimum margin of 0.5 m from the walls. In this experiment, the source signal is WGN, and 30 dB Signal-to-Noise Ratio (SNR) sensor noise, simulated using WGN, is also added to each microphone. A dataset of 15,000 samples is generated, from which 10,000 samples are used for training, 2,500 for validation, and 2,500 for testing.

3.3.2 Simulated reverberant rooms

The data for the simulated reverberant rooms experiment is generated similarly to the anechoic experiment. However, instead of simulating sound propagation in an anechoic environment, each dataset sample is randomly assigned a reverberation time value for its corresponding room from a uniform distribution within the range of $[0.3 - 0.6]$ s. This value is used to simulate reverberation using the image source method [20]. For the source signal, speech recordings from the Voice Cloning Toolkit (VCTK) corpus [177] were used. The number of training, testing and validation samples is same as in the above section.

3.3.3 Real recordings

For this experiment, instead of simulations, measurements from the LibriAdhoc40 dataset [178] (GPL3 license) were used. The signals were recorded in a highly reverberant room containing a grid of forty microphones and a single loudspeaker, which was placed in one of four available locations. The microphones recorded speech sentences taken from the Librispeech [179] corpus, which were played back through the loudspeaker. The reverberation time measured by the dataset authors was of approximately 900 ms.

To generate each dataset sample, four of the forty available microphones were randomly selected. The microphone selection was restricted to the outermost microphones of the grid, where one microphone per side is selected. A visual explanation of the microphone selection procedure is provided in Fig. 3.3b. There are four available positions for the microphones near each of the west and east walls and seven positions near each of the north and south walls. Furthermore, there are four available source positions. There are, therefore, $4 \times 4 \times 7 \times 7 \times 4 = 3{,}136$ source/microphone combinations available for selection. Finally, four speech utterances were selected for each combination, resulting in a dataset of 12,544 samples. 50% of those combinations was used for training, 25% for validation and 25% for testing. To create the training dataset for this experiment, the aforementioned training split was augmented with the training data of the reverberant dataset described in Sec. 3.3.2, resulting in a dataset consisting of $10{,}000 + 6{,}272 = 16{,}272$ signals.

3.3.4 Metadata sensitivity study

In practical scenarios, the metadata, e.g., microphone coordinates and room reverberation time in PSSL, are uncertain because they are typically estimated or measured. To investigate the robustness of the proposed approach to such uncertainties, a sensitivity study was conducted using the test dataset in Sec. 3.3.2. The dataset was modified by introducing different levels of perturbations to the input metadata, followed by a computation of the mean localization error for each level using the model trained in Sec. 3.3.2.

The first three studies consist of perturbing the microphone coordinates of the testing dataset with increasing levels of random Gaussian noise. The reported precision of microphone coordinates measured optically is under a millimeter [180]. Conversely, when these are estimated using self-localization algorithms, the reported errors are under 7 cm [181], [182]. The standard deviation levels of the introduced noise were set to to 1, 10 and 50 cm. In the fourth study, random Gaussian noise was introduced to the reverberation time with a standard deviation of 200 ms, based on reported errors obtained on reverberation estimation procedures [183], [184].

3.3.5 Metadata relevance study

To quantify the contribution of each metadata category to the improvement in localization performance, a metadata relevance study was conducted, where the DI-NN network was trained using six different combinations of the microphone positions, room dimensions and reverberation time. The results are summarized in Table 3.3.

3.4 Results and discussions

3.4.1 Results

Figure 3.4a compares the average error of the proposed DI-NN and DI-NN-Embedding methods to the CRNN and LS baselines. DI-NN, DI-NN-Embedding and CRNN models were trained four times independently for each experiment using random initial network parameters. The results shown in Fig. 3.4 are averaged across the four times on simulated data, with error bars showing the standard deviation across the runs. Conversely, as the

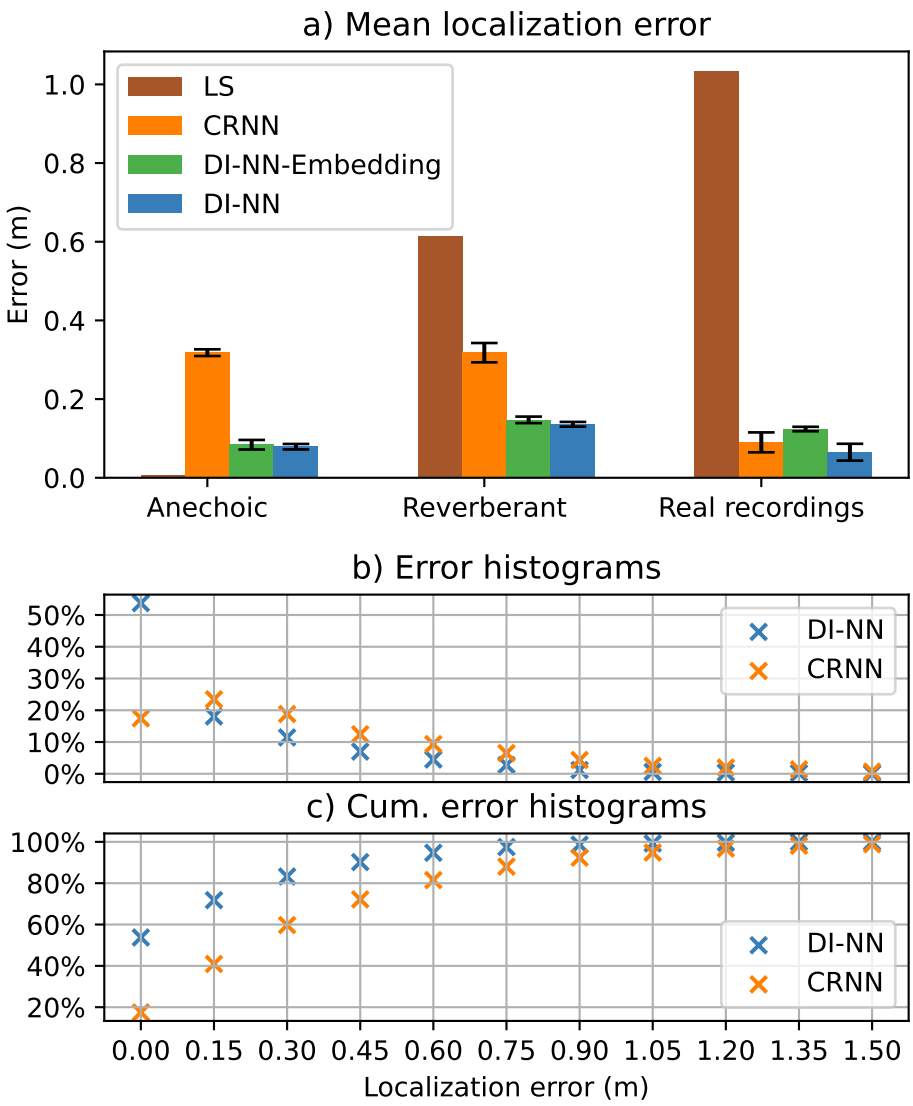

Figure 3.4: (a) Mean localization error for DI-NNs and baselines on different datasets. (b) Normalized histogram comparison between the DI-NN and the CRNN baseline on the recorded dataset. (c) Cumulative version of (b).

LS method is deterministic, it does not require multiple runs.

A first remark is that although the LS approach is very effective in the anechoic scenario, its performance is degraded on the other datasets, indicating its sensitivity to reverberation. The CRNN outperforms the LS method in reverberant scenarios without knowledge of the microphone's coordinates. Interestingly, the CRNN baseline also obtains good localization performance on the recorded dataset, indicating that the network is able to infer the metadata to an extent when trained on a single room.

However, by exploiting the microphone coordinates, the DI-NN is shown to significantly improve the performance compared to the CRNN. The most significant difference is observed in the anechoic case, where an improvement close to three times is obtained. In this case, the microphone coordinates are more useful as this information cannot be derived from the signals. In a reverberant room, however, the network might be able to use reflections to its advantage, as discussed in [185], to infer the microphone coordinates and making the metadata less useful. Figure 3.4a also shows the errors obtained using the alternative DI-NN-Embedding architecture were similar to the DI-NN in all scenarios, indicating no advantage in the proposed embedding, although it still allows the network to exploit the metadata.

In turn, Fig. 3.4b compares the normalized error histograms between the approach and the CRNN baseline on the real recordings test dataset. The mode of the DI-NN's error is centred on the 0-15 cm bin compared to the 15-30 cm bin for CRNN's error. The cumulative distribution for the same data is shown in Fig. 3.4c. While the DI-NN is shown to locate over 50% and 80% of the dataset samples with less than 15 and 45 cm error, the CRNN achieves the same errors for less than 20% and 60% of the data, respectively.

Table 3.2: Metadata sensitivity analysis

Coord. std. (m)	Reverb. std. (ms)	Err. increase (%)
0.01	0	0.05
0.1	0	1.02
0.5	0	32.9
0	200	0.4

The results of the sensitivity study conducted in Sec. 3.3.4 are displayed in Table 3.2. The last column refers to the relative error increase between the perturbed case and

the noiseless experiment conducted in Sec. 3.3.2. The results show that the proposed approach is robust to the uncertainty inherent in practical measurements of the microphone coordinates and reverberation time estimates. The case where the microphone coordinates are disturbed by an extreme error of 0.5 m (more than five times above typical errors) has been included to demonstrate the impact of including microphone coordinates for PSSL, reiterating the importance and improved performance of metadata in the proposed fusion approach.

Table 3.3: Metadata relevance analysis

Mic. coords.	Room dims.	RT60	% performance
✓	✓	✓	100
✓	✓	✗	102
✓	✗	✓	100
✗	✓	✓	61
✓	✗	✗	104
✗	✓	✗	60
✗	✗	✓	47

Finally, the results of the metadata relevance analysis study described in Sec. 3.3.5 are displayed in Table 3.3. Each line represents a version of the DI-NN model trained on the reverberant dataset. The first three columns describe which metadata types are used in the model, and the last column shows the model performance relative to the model using all metadata, represented in the first line. The results show that the microphone coordinates are the most relevant for the model. In fact, using the microphone coordinates alone provides the best results. The results also indicate that the room dimensions are more relevant than the reverberation time in the absence of the microphone coordinates.

3.4.2 Limitations and extensions

DI-NN exploits the metadata, such as the microphone coordinates and reverberation time and therefore this data must be known a priori or somehow measured. This additional information is however shown to be justified by a significant improvement in performance. While the gains of the microphones are assumed to be calibrated in the experiments, which may not be verifiable in practical scenarios, it is shown in Sec. 3.3.3 that the model can perform well even when using uncalibrated microphones of the same kind. If calibration

cannot be ensured, extracting gain invariant features from the signal pairs such as the cross spectra [186] may be used as a preprocessing step.

The scope was limited to the localization of one static sound source using static microphones to focus on metadata fusion. However, extensions to moving sources and microphones could be possible by using smaller processing frames, for example. Another extension would be to estimate the three dimensional coordinates of the source. Finally, a possible extension for multiple source localization is expanding the output of DI-NN to a vector of size $2N$, where N is the number of maximum sources, and performing Permutation Invariant Training (PIT) [187].

3.5 Conclusion

In this work, DI-NN was proposed, a simple yet effective way of jointly processing signals and relevant metadata using neural networks. Results for the task of SSL on multiple simulated and recorded scenarios indicate that the DI-NN is able to exploit successfully the metadata, as its inclusion reduced the mean localization error by a factor of at least two compared to the CRNN baseline, as well as significantly improving localization results in comparison with the classical LS algorithm in reverberant environments. Additional relevance and sensitivity studies revealed that the microphone coordinates were the most important metadata, and that the DI-NN is robust to realistic noise in the metadata.

Chapter 4

Graph neural networks for sound source localization

4.1 Introduction

Distributed Microphone Array (DMA) signal processing [188] is an active field in the acoustic signal processing community, with important applications in speech enhancement, noise reduction and Sound Source Localization (SSL) [188]–[190]. In contrast to centralized microphone arrays [69], DMAs may be created through the wireless connection of multiple distributed devices such as cell phones, laptops and virtual assistants. In this context, they are also frequently referred to as Ad-hoc microphone arrays, or Wireless Acoustic Sensor Networks (WASNs).

Although DMAs bring advantages in terms of acoustic coverage with respect to centralized arrays, they also bring challenges. One such challenge forms the focus of this chapter, namely, having a dynamic number of input microphone channels, as a DMA may be created using the devices present in a dynamic scene. This number may change in runtime due to many reasons, including software or hardware failures of individual devices, battery depletion, or the device being removed from the scene. This restricts the application of many of the deep learning methods that have been successfully applied to centralized microphone networks such as [81], [86], which require a static input size. Conversely, classical SSL approaches such as [159] are able to function on an arbitrary number

of microphones.

In this work, Graph Neural Networks (GNNs) [191]–[193] are proposed as a suitable way of processing DMA signals for the task of SSL. A GNN variant called the Relation Network (RelNet) [193] is adopted. The approach is validated for the task of localizing a single static source in multiple scenarios, showing it to outperform the baselines. The main contribution of this work is the first application of GNNs for the task of SSL, the proposed method is designed to handle a variable number of microphone channels. Furthermore, this approach can work on unseen microphone coordinates and room dimensions through a metadata fusion procedure.

This chapter continues by providing a problem statement in Sec. 4.2. Sec. 4.3 includes a review of related work on DMAs using deep learning, as well as a review of classical SSL methods and the RelNet GNN, which serve as building blocks for the proposed model. In Sec. 4.4, the proposed approach is described. Sec. 4.5 describes the experimental validation, and Sec. 4.6 presents the results and Sec. 4.7 concludes the chapter.

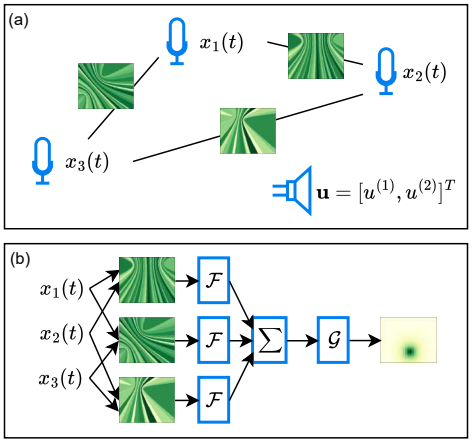

Figure 4.1: (a): Example of a graph of distributed microphones. (b): Representation of the GNN-SLF model for three microphones. The computation of the heatmaps is described in Sec. 4.4. $\mathcal{F}$ represents the relation function which is computed for each pair of microphone features.

4.2 Problem statement

Our goal is to estimate the 2D coordinates $\hat{\mathbf{u}}$ of a sound source located at $\mathbf{u} = [u^{(1)}\, u^{(2)}]^T$ within a reverberant room of known dimensions $\mathbf{d} = [d^{(1)}\, d^{(2)}\, d^{(3)}]^T$. The source emits a speech signal $s(t)$ at instant t. In addition to the source, M microphones are present in the room, where the microphone m has a known position $\mathbf{v}_m = [v_m^{(1)}\, v_m^{(2)}\, v_m^{(3)}]^T$. A metadata vector ϕ is also defined as

$$\phi = [v_1^{(1)}\, v_1^{(2)} ... d^{(2)}\, d^{(3)}]^T, \tag{4.1}$$

which serves as a secondary input to the proposed method, allowing it to function on any room dimensions and microphone coordinates.

This microphone network can be viewed as a graph where microphones and their respective positions are represented by nodes, and edges are defined as the relation between pairs of nodes. This relation will be described in the following sections.

4.3 Related work

4.3.1 Classical SSL methods

Our proposed method can be seen as a generalization of classical grid-based SSL methods such as the Time-Difference-of-Arrival (TDOA) [62], [63], Spatial Likelihood Function (SLF) [159], [194] and energy-based [195] approaches. These approaches share many similarities, which are summarized by their shared behaviour described in Alg. 1.

Algorithm 1 Classical SSL methods

```
function ESTIMATE_SOURCE_LOCATION(X, φ)
    g ← 0
    for each i ≠ j ∈ [1..M] do
        g += F(x_i, x_j; φ(i, j))
    return G(g)
```

Alg. 1 starts with the creation of an empty grid $\mathbf{g}$, which is assumed to be flattened in 2D. The next step consists of computing a *relation* $\mathcal{F}$ between each pair of microphones (i, j) available, using their signals $(\mathbf{x}_i, \mathbf{x}_j)$ as well as the *metadata* available ϕ, consisting of the microphone and room dimensions and the speed of sound. These relations consist

of assigning, for each cell within the grid, a value expressing how likely a source is to be in a particular grid cell.

The relations between all pairs are aggregated through summation (or multiplication, see [194]) to generate a heatmap gathering all pairwise information. Depending on whether the problem is formulated using a LS or Maximum Likelihood (ML) approach, the minimum or maximum value of the grid will respectively correspond to the location of the source [63]. $\mathcal{G}$ is therefore a peak-picking function, whose goal is to select the grid cell where the source is located.

The TDOA, SLF and energy-based methods differ mainly by the function $\mathcal{F}$ computed. Each cell within the grid represents a candidate source location which has a theoretical TDOA between the two microphones. In the TDOA method, each grid cell is assigned the distance between its theoretical TDOA and the measured TDOA, computed by picking the peak of the generalized cross-correlation function between the microphones' signals, typically computed using the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [64].

In the SLF method, each cell receives the cross-correlation value at the lag corresponding to its TDOA. SLF is a generalization of the Steered Response Power (SRP) method [69]. Finally, the energy-based method uses a metric based on the ratio of the two microphone signals' energies. In Fig. 4.1a, the edges of the graph represent maps computed using the SLF method.

4.3.2 Neural network methods for DMA signal processing

Classical SSL methods normally do not account for room reverberation, which may divert the heatmap's peak from the true source location, or reduce its sharpness. Neural networks can become robust to reverberation if trained on suitable scenarios. Here, works on neural networks for DMAs are reviewed.

In [196], an attention-based neural network capable of handling connection failures is proposed for the task of speech enhancement. Unlike the proposed method, this network is limited to a maximum number of input microphones channels. In [197] and [198], variable-input processing is achieved through a global average pooling scheme.

Two works have explored GNNs for acoustic signal processing. In [199], a GNN is used to profile noise within a railway setting. However, their work requires the source signal to be known beforehand, limiting its application in many scenarios. This restriction is not present in the proposed approach. In [189], a Graph Convolutional Network (GCN) [200] is used in conjunction with an encoder-decoder network for the task of speech enhancement. Conversely, a Relation Network GNN is used in favour of an encoder-decoder GCN in our proposed method, which is shown to be well suited for the task of SSL.

4.3.3 Relation Networks

The Relation Network (RelNet) [193] is chosen as the proposed graph network architecture due its conceptual similarities to classical SSL methods. RelNets were introduced in the context of visual question answering. The input of the network consists of a set of *nodes*, represented by feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$. The network $\mathcal{RN}$ may be summarized as

$$\hat{\mathbf{y}} = \mathcal{RN}(\mathbf{X}) = \mathcal{G}\left(\sum_{i \neq j} \mathcal{F}(\mathbf{x}_i, \mathbf{x}_j)\right), \tag{4.2}$$

where (4.2), $\mathcal{F}$ generates a *relation* between nodes (i, j). These relations are summed together, and this sum is the input to $\mathcal{G}$, which produces the answer $\hat{\mathbf{y}}$ to the target question. The nodes $\mathbf{x}_i$ and the relations $\mathcal{F}(\mathbf{x}_i, \mathbf{x}_j)$ can be seen as a complete undirected graph, where all nodes are connected by a relation function $\mathcal{F}$, that is, $\mathbf{G} = (\{\mathbf{x}_i\}, \{\mathcal{F}(\mathbf{x}_i, \mathbf{x}_j)\})$. As in [193], both $\mathcal{F}$ and $\mathcal{G}$ are implemented as Multi-layer Perceptrons (MLPs), trained jointly using backpropagation.

4.4 Method

A diagram of the proposed network is shown in Fig. 4.1b. Using a RelNet allows the proposed approach to first process pairs of microphone signals into features, and later combine them through summation. This allows it to function on a variable number of input microphones. Furthermore, the proposed method is designed to operate on multiple room dimensions and microphone coordinates by combining this metadata ϕ before estimating

the source location.

The input to the proposed method consists of the set of M microphone signal frames $\{\mathbf{x}_m\}$, where $\mathbf{x}_m$ is a vector of size L representing a signal frame, and a metadata vector ϕ containing relevant information such as the microphone coordinates and room dimensions. The relation function $\mathcal{F}$ is defined as

$$\mathcal{F}(\mathbf{x}_i, \mathbf{x}_j; \phi) = \mathrm{MLP}(\mathcal{H}(\mathbf{x}_i, \mathbf{x}_j; \phi)), \tag{4.3}$$

where MLP is a multi-layer perceptron and $\mathcal{H}$ is a preprocessing or feature extraction function. The inclusion of a preprocessing function allows the use of classical features such as GCC-PHAT or SLF. Conversely, post-processing these functions using a MLP allows these features to be improved by introducing learned rules, as will be shown for the application of SSL.

In turn, the relation fusion function is chosen as $\mathcal{G}(\mathbf{u}) = \mathrm{MLP}(\mathbf{u})$, where $\mathbf{u}$ represents the sum of all pairs of relations as in Alg. 1. This function is a substitution of the peak-picking algorithm in Alg. 1, expanding its functionality for other possible applications. Note that the relation function was not trained to be symmetric.

As in [193], weights $\mathbf{w}_{\mathcal{F}}$ and $\mathbf{w}_{\mathcal{G}}$ of the MLPs in $\mathcal{F}$ and $\mathcal{G}$ are jointly trained through a gradient-based procedure by minimizing an application-specific loss function $\mathcal{L}(y, \hat{y})$ between the network output $\hat{y}$ and target y

$$\begin{aligned} \mathbf{w}_{\mathcal{F}} &= \mathbf{w}_{\mathcal{F}} - \lambda_{\mathcal{F}} \frac{\partial \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{w}_{\mathcal{F}}} \\ \mathbf{w}_{\mathcal{G}} &= \mathbf{w}_{\mathcal{G}} - \lambda_{\mathcal{G}} \frac{\partial \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{w}_{\mathcal{G}}}, \end{aligned} \tag{4.4}$$

where $(\lambda_{\mathcal{F}}, \lambda_{\mathcal{G}})$ are the learning rates, usually defined by the optimizer used, such as Adam [176].

Two preprocessing functions $\mathcal{H}$ have been investigated experimentally for the proposed relation function $\mathcal{F}$. The first is the cross-correlation between pairs of two microphone signals, computed using the GCC-PHAT method. In this case, the network needs to learn to map time lags into space. As an alternative, the cross-correlation is projected into space using the SLF method. The output of this method is a flattened $N \times N$ grid

or a N^2 vector. In this case, the network needs to learn to denoise the maps which may have been corrupted by reverberation.

A final step in the feature extraction step is concatenating the microphone coordinates of the pair as well as its room dimensions into the features. This is especially important for the GCC-PHAT feature extractor, as the network must learn how to project the temporal information into space.

The target of the MLP of function $\mathcal{G}$ is to further enhance the summed maps produced by $\mathcal{F}$. Its output has the same size as $\mathcal{F}$, representing a flattened $N \times N$ grid of cells centered at coordinates $\{\mathbf{u}_{u,v}\}$ within the room. The target value $y(u, v)$ of each grid cell (u, v) is computed as

$$y(u, v) = e^{-\|\mathbf{u}_{u,v} - \mathbf{u}\|_2}, \tag{4.5}$$

where $\mathbf{u}$ is the target source location. Note the maximum value of 1 occurs when $\mathbf{u}_{u,v} = \mathbf{u}$ and approaches 0 exponentially as the distance between $\mathbf{u}_{u,v}$ and $\mathbf{u}$ increases. The mean absolute error between the network output and target is used as the loss function. This formulation can be extended for detection of multiple sources, which can be extracted through peak-picking. However, this work focuses on the detection of a single source.

4.5 Experimentation

This section describes the experiments with the proposed network for SSL described in the previous section. The proposed methods are referred to as GNN-GCC for the network using the GCC-PHAT feature extractor, and GNN-SLF for the one using the SLF extractor. This approach is compared with two baselines, the classical Time-Difference-of-Arrival (TDOA)-based and Spatial Likelihood Function (SLF)-based approaches, as described in Sec. 4.3. A public repository containing all methods is provided on Github [1]

4.5.1 Dataset

The proposed network was tested using synthetically generated data using the image source method [201], generated using the Pyroomacoustics library [171]. To demonstrate that the proposed approach is able to operate with a different number of microphones

[1] https://github.com/egrinstein/gnn_ssl

than it was trained on, the training set for the GNN uses training examples containing $\{5, 7\}$ microphones, while the test set examples contain $\{4, 5, 6, 7\}$ microphones.

For each dataset sample, two numbers are randomly selected from a uniform distribution in the interval [3, 6] m representing the room's width and length. The room's height is uniformly selected from the interval [2, 4] m. The room's reverberation time is sampled uniformly from the interval [0.3, 0.6] s using Eyring's formula [202]. The microphones and source are randomly placed within the room, with the restriction of each device being at least 0.5 m from each other and the room's boundaries. Each source is set to emit a speech sample from the VCTK corpus [203]. The SNR in each microphone is set at 30 dB, simulated by adding WGN independently to each channel to the auralizations generated using the image source method. The training, validation and test datasets contain respectively 15,000, 5000 and 10,000 examples.

4.5.2 Method hyperparameters

The networks were trained for a maximum of 100 epochs with early stopping if the validation loss stops increasing after 3 epochs. A learning rate of 0.0005 using the Adam optimizer [176] was employed. A batch size of 32 was used. These parameters were chosen empirically. All grids used are of dimensions 25×25. The input frame size used was L=500 ms. For the GCC-PHAT method, a DFT of $1,024$ samples was used. Since the maximum TDOA value was bounded by the room's diagonal, only the central 200 correlation bins are selected, similar to [84]. In the proposed method, the relation function's MLP contains 3 layers, each of output size 625. The function $\mathcal{G}$'s MLP consists of 3 layers, all with an output size of 625 neurons. ReLU activation function was used for all layers except for the output, which uses no activation.

The grids computed in the SLF and TDOA baselines as well as the feature extractor in the GNN-SLF method have a size of 25×25. The source estimation procedure in the baselines and proposed methods consists of picking the location of the highest value in the SLF method, and the lowest on in the SLF method.

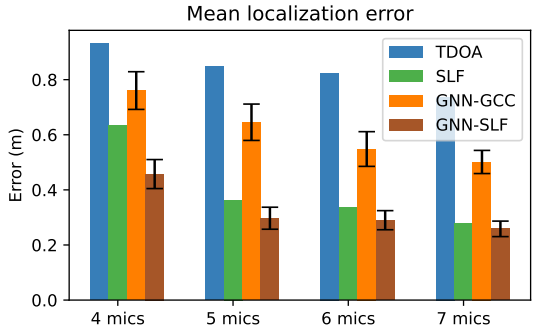

Figure 4.2: Localization error for the proposed methods and baselines.

4.6 Results

The metric used to evaluate the methods consists of the mean euclidean distance between the estimated and true source location on the test set. The results are shown in Fig. 4.2. Note that although all methods are tested on unseen simulations containing $\{4, 5, 6, 7\}$ microphones, the method was only trained using examples containing $\{5, 7\}$ microphones. To ensure a fair comparison, the networks were trained multiple times. The black bars show their standard deviation.

It can be seen that the GNN-SLF method outperforms all others, demonstrating the effectiveness of the approach. The biggest relative improvement of 29% with respect to classical SLF is observed for four microphones. An explanation is that, when there are fewer measurements available, improving or discarding them becomes crucial, which may be the operation being performed by the network. Is can be also seen that GNN-GCC performed poorly, only surpassing the TDOA baseline. This indicates that requiring the network to learn to map time delays to spatial position is a possibly more demanding task than dealing with the already spatialized information.

4.7 Conclusion and future work

The RelNet, a type of GNN, is proposed for the task of SSL on distributed microphone arrays. Results show that it can significantly improve the localization performance over classical localization algorithms, achieving a 29% improvement in the case of 4 microphones.

It is also shown that the method generalizes to an unseen number of microphones. Future directions include testing approach for localizing multiple sources and learning graph topologies different than the complete graph.

Chapter 5

The Neural-SRP method for Sound Source Localization

This chapter contains the major contribution of this work, a neural network-based method for sound source localization which is able to function on any microphone geometry. As its functioning is based on the classical SRP method, it is named *Neural-SRP* method.

5.1 Introduction

Direction-of-Arrival (DOA) estimation uses the signals from a microphone array to estimate the angular position of one or more active sound sources relative to the array. Applications include event detection [42], [54], [204], camera steering [58] and sound source separation [32], [34], [205]. Although many classical, signal processing based methods such as Multiple Signal Classification (MUSIC) [93], Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [94] and SRP [68], [206] have been extensively explored over the last decades, state-of-the-art localization performance is usually currently obtained using deep learning methods [79], where a neural network model is trained to estimate the location of the desired sources using a feature representation of the multi-channel microphone signals.

Neural DOA estimators can be classified according to their input features as *Time/Frequency (T/F)* or *hybrid*. T/F networks (e.g. DoaNet [17]) typically process

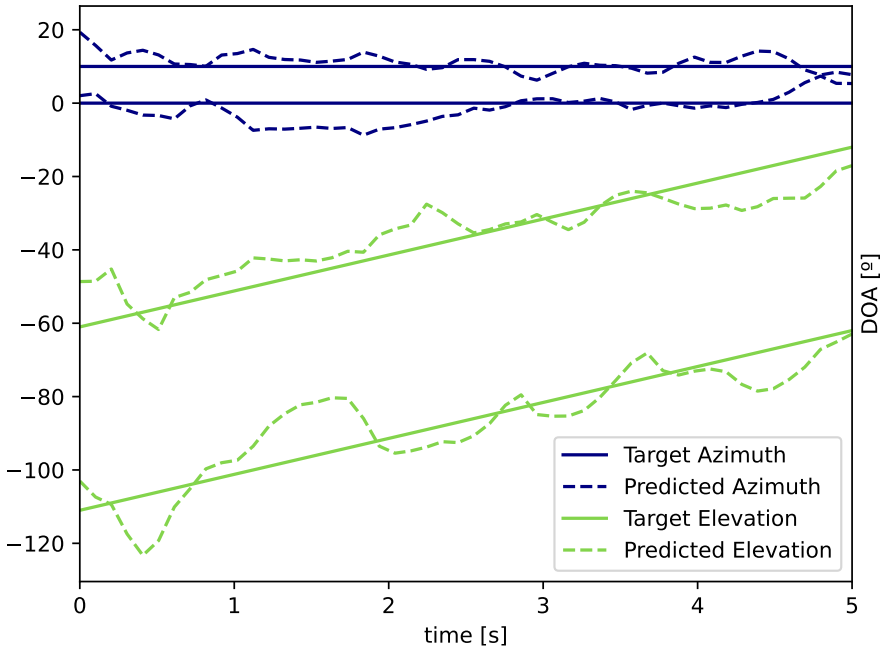

Figure 5.1: Example of Neural-SRP's output when tracking two moving sources. The panel shows the target and predicted azimuth and elevations.

features such as the multichannel STFT, GCC-PHAT or the raw audio signal. A disadvantage of these networks is inflexibility to the microphone geometry, i.e., the number of microphones and respective positions of the array. This requires retraining for each array geometry, a cumbersome task which limits their off-the-shelf usage as a general tool. This also requires companies providing multiple array geometries within their line of products, such as voice assistants, to maintain multiple training pipelines. In contrast, current hybrid networks (e.g. Cross3D [89]) overcome this limitation by processing an input feature set that is independent of the number of microphone channels and their geometry, typically obtained using a classical signal processing DOA estimator such as the SRP method described in Sec. 2.1.2. A limitation of this approach is that it inherits the limitations of the underlying DOA estimator, such as an assumption of anechoic propagation and the lack of robustness to directional noise sources.

The main contribution of this work is Neural-SRP, a T/F neural localization method which overcomes the limitations of previous models. Table 5.1 shows a qualitative comparison of Neural-SRP with respect to Cross3D [89] and DOANet [17], arguably the literature's most established single and multi-source DOA estimation models. Unlike the DOANet, Neural-SRP is causal, therefore applicable to real-time applications, and uni-

Model	Causal	Universal	Multi-source
DOANet [17]			✓
Cross3D [89]	✓	✓	
Neural-SRP	✓	✓	✓

Table 5.1: Functional comparison of the proposed model and baselines. 'Universal' refers to the method's capacity of working on any microphone array geometry.

versal, therefore applicable to arbitrary microphone geometries. In addition, unlike the Cross3D method, Neural-SRP is able to localize multiple sources simultaneously, as illustrated in Fig. 5.1. Finally, The proposed network is significantly smaller than the baselines. Code for the Neural-SRP architecture that can reproduce the experiments in this chapter is available on Github [1].

Geometric independence is achieved by the introduction of two concepts, *pairwise processing* and *metadata fusion*. The former is inspired by the conventional SRP method, where a local feature is extracted between all microphone pairs, such local features then being summed to create a global feature. By providing the network with the microphone positions using a *metadata fusion* procedure, it is able to produce an *encoded pairwise spatial likelihood map*. After summation, the global feature is then decoded to estimate the sources' locations.

This chapter presents a continuation of the work in chapter 3, where the concept of a dual-input neural network capable of jointly processing signals and metadata, such as the microphone positions, room dimensions and reverberation time was introduced for the task of PSSL. This allowed a T/F neural model to operate on distributed microphone arrays of unseen geometries, but with a fixed number of microphones. This constraint is removed in chapter 4, where a spatial approach involving GNNs is applied to the enhancement of SRP maps. The remainder of this section focuses on the Cross3D [89] and DOANet [17] methods, which are respectively state-of-the-art hybrid and T/F models which serve as comparison baselines to this work.

This chapter continues as follows. Sec. 5.2 describes the proposed model, followed by the experimental validation in Sec. 5.3. The results are discussed in Sec. 5.5.

[1] https://github.com/egrinstein/neural_srp

5.2 Neural-SRP

5.2.1 Input Feature Set

The input feature of Neural-SRP consists of the GCC-PHAT$_t$, $\mathbf{g}_{ij}$, of all pairs of microphone signal frames $(\mathbf{x}_i, \mathbf{x}_j)$, as defined in (2.6). The input feature consists of the GCC-PHAT between all microphone pairs, thus generating an input of shape $(M(M-1)/2, T, G)$, where T is the number of time-frames and G is the number of central GCC delays used. This selection has the advantage of reducing the input size and removing delays which are bigger than the maximum theoretical TDOA for the microphone array, computed as

$$G = 2 \max\left\{ \frac{\|\mathbf{v}_i - \mathbf{v}_j\| f_s}{c} \right\} + 2G_0 \tag{5.1}$$

where $1 \leq i < j \leq M$ and $G_0 \geq 0$ is a parameter to increase the feature size to values beyond the maximum theoretical TDOA, which increases performance in practice [17]. This input feature is also used by the DOANet model. However, while the DOANet model jointly processed all input features using a single network, the proposed model processes each pairwise feature independently to create a summable encoded likelihood map, allowing the network to accept any number of microphone pairs as its input.

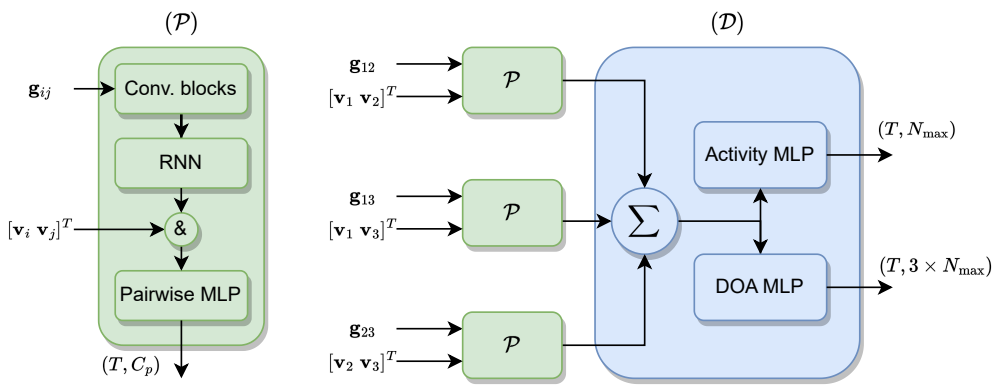

Figure 5.2: Neural-SRP network architecture. Left, green: pairwise network $\mathscr{P}$. Right, blue: Global decoder $\mathscr{D}$, exemplified for a 3-microphone input. Symbol "&" represents concatenation.

5.2.2 Architecture

The Neural-SRP network is divided into two sub-networks, namely, a pairwise network $\mathscr{P}$ and a global decoder $\mathscr{D}$. The architecture is shown in Fig. 5.2 and is summarized as

$$\hat{\mathcal{U}} = \mathscr{D}\left(\sum_{i=1}^{M}\sum_{j=i+1}^{M} \mathscr{P}(\mathbf{g}_{ij}, \mathbf{v}_i, \mathbf{v}_j)\right). \tag{5.2}$$

The goal of $\mathscr{P}$ is to create an encoded and summable spatial likelihood feature for each signal pair, using GCC $\mathbf{g}_{ij}$ along with its respective microphone coordinates $(\mathbf{v}_i, \mathbf{v}_j)$. These features are them summed together, creating a global feature which is then decoded by $\mathscr{D}$ to estimate a set of locations $\hat{\mathcal{U}}$. The proposed method's name derives from the structural similarity between (5.2) and (2.9).

The pairwise network consists of a modified Convolutional Recurrent Neural Network (CRNN) architecture. The parameters of the pairwise network are shared across all pairs. Each pairwise GCC is first processed by a sequence of 2D convolutional blocks. To maintain causality, the kernel size in the time dimension is set to 1 and no pooling is applied in that dimension. Unit strides were used on convolutional layers. The resulting feature of shape $(T, C_c^{(0)}, C_c^{(1)})$ is transformed into shape (T, C_c) by flattening the last two dimensions of size $C_c^{(0)}$, the number of convolutional output kernels, and $C_c^{(1)}$, the number of GCC bins after pooling.

To improve tracking performance, the resulting feature is then processed by a one-directional RNN of type Gated Recurrent Unit (GRU) [170]. To produce a spatially-aware feature (i.e., which incorporates microphone position information), the microphone coordinates of each microphone in the pair are concatenated to each channel, followed by transforming this feature into an encoded likelihood map of shape C_p through the application of another MLP. An interpretation of this step is 'steering' the feature produced by the RNN according to the direction of the segment connecting the microphone pair's positions. The reader is referred to chapter 3 for a detailed discussion on methods for incorporation of metadata, namely microphone position information, for the improvement of SSL methods.

The decoder $\mathscr{D}$ consists of two independent MLPs as in the DOANet model. The first is an activity detector similar to a multichannel Voice Activity Detector (VAD) [207],

while the second outputs the $\hat{N}$ estimated locations. These outputs are implicitly related, in the sense that if the nth activity detector indicates no activity, the values of the nth estimated DOA should be ignored.

5.2.3 Training for DOA estimation

Both pairwise and global networks are jointly optimized using the network's output. In the following, the loss function is defined for each temporal instant t and will therefore omit this index. $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_{\hat{N}}]$ and $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1 \dots \hat{\mathbf{u}}_{\hat{N}}]$ are defined as the target and output DOA matrices respectively, where each column is a unit vector representing a true or estimated DOA. $\mathbf{z}$ and $\hat{\mathbf{z}}$ are also defined as $\hat{N}$-dimensional binary vectors which refer to the target and output activities. In the case where only a single source exists, the loss function is defined as

$$\mathcal{L}(\mathbf{U}, \hat{\mathbf{U}}, \mathbf{z}, \hat{\mathbf{z}}) = \alpha\, z_1 \|\mathbf{u}_1 - \hat{\mathbf{u}}_1\| + \beta\, \mathrm{BCE}(z_1, \hat{z}_1) \tag{5.3}$$

where the first term is the *Euclidean localization error* between the true and estimated DOA, weighted by the true activity, so as to ignore silent frames. The Euclidean error is employed in favour of the more interpretable angular error as previous works [88], [89] found it to yield better training results. The weighting factors α and β are hyperparameters. The second term is the binary cross-entropy $\mathrm{BCE}(z_1, \hat{z}_1) = (z_1 \log \hat{z}_1 + (1 - z_1) \log(1 - \hat{z}_1))$ between the true and target activity.

To prevent the loss function from diverging to $-\infty$, the maximum value of $\log(\cdot)$ is clamped to a constant B. When two or more sources are active, the training must take the *assignment problem* into account, so as not to penalize equivalent target and true permutations [208]. This problem can be defined as finding the association matrix $\mathbf{A}$, a permutation of the rows of the identity matrix of size $\hat{N}$. The optimal $\mathbf{A}$ minimizes the multi-source localization error, defined as

$$\mathcal{L}^{doa}(\mathbf{U}, \hat{\mathbf{U}}, \mathbf{z}) = |\mathbf{D} \odot \mathbf{A}|/|\mathbf{z}|, \tag{5.4}$$

where $[D]_{ij} = \|\mathbf{u}_i - \hat{\mathbf{u}}_j\|$ is the distance matrix between all target and output combinations, $\odot$ is the element-wise product, $|\cdot|$ is the matrix norm and $|\mathbf{z}|$ is the number of active sources.

Although $\mathbf{A}$ can be deterministically computed using the Hungarian algorithm [209], the latter is not differentiable, hindering its application for training the neural network using a backpropagation procedure. This is solved in the same manner as [17], where a neural network is used to approximate the Hungarian method, and then used for training. The association matrix is also used for aligning the target and output activities $\mathbf{z}$ and $\hat{\mathbf{z}}$, after which the binary cross-entropy function is applied for each entry.

5.2.4 Training for PSSL

A frequent choice to be taken when designing a Neural SSL model is whether to choose the regression or classificaiton output encoding, as each shows advantages and disadvantages [88]. The aforementioned formulation uses the regression paradigm for the task of DOA estimation.

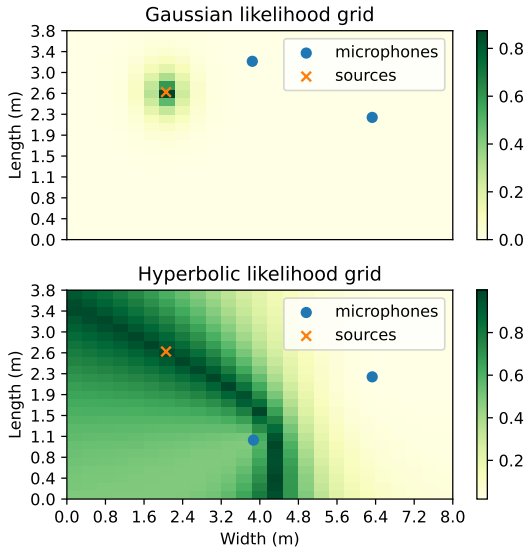

Figure 5.3: Example of the hyperbolic grid (bottom), used for training Neural-SRP, and the alternative Gaussian grid (top).

In that case, the network's desired output can be formulated as a likelihood grid, i.e., a matrix when performing 2D localization, where each candidate grid cell's value is proportional to its distance to the source, that is, $\text{NSRP}(\mathbf{u}) \propto |\mathbf{u} - \mathbf{u}_s|$. Is is desirable to bound the value in the interval $[0, 1]$, so it can be normalized and interpreted as a

probability. In [84], a Pseudo-Gaussian target equal to

$$y_g(\mathbf{u}) = e^{-(\|\mathbf{u}-\mathbf{u}_s\|/\sigma)^2} \tag{5.5}$$

is used, which satisfies the aforementioned properties. However, this target is unrealistic for training on a single microphone pair, as source locations share the same cross-correlation function along a hyperbola [62] in anechoic scenarios. The target output is modelled using a *hyperbolic target grid*, defined for each grid cell as

$$y_h(\mathbf{u}) = e^{-(\|\tau_{ij}(\mathbf{u})-\tau_{ij}(\mathbf{u}_s)\|/\sigma)^2}, \tag{5.6}$$

where σ controls the function's decay. $\tau_{ij}(\mathbf{u})$ represents the theoretical TDOA between microphones i and j from a source at $\mathbf{u}$, as defined in (2.3). A comparison between both grids can be seen in Fig. 5.3. Note that (5.6) is maximized to a value of 1 at the source's location, and decays to 0 as the distance to the source increases. This grid assigns a high likelihood value along a hyperbola branch, hence its name. An interpretation for this target is that the network must learn to estimate a likelihood grid similar to one that would be produced by SRP in an anechoic environment. In other words, the network has to learn to jointly dereverberate and locate the source.

Finally, the training loss function is defined as

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = |\boldsymbol{Y} - \hat{\boldsymbol{Y}}|, \tag{5.7}$$

the Mean Absolute Error (MAE) between the target grid $\boldsymbol{Y}$ computed using (5.6) and the network output $\hat{\boldsymbol{Y}}$. The training procedure for Neural-SRP consists of applying a gradient based method search for the network parameters or weights which minimize (7.2).

Anechoic to reverberant transfer learning

Experiments showed that training directly on highly reverberant data led to the network getting stuck on local optima values. This was overcome by applying a two-stage training procedure. Firstly, the network was trained using anechoic data. Then, training was

resumed using a reverberant dataset, which allowed the network to reach substantially lower error values. This procedure is referred to as transfer learning or fine-tuning [210].

5.3 Experimentation for DOA estimation

Experiments were performed consisting of training/evaluating Neural-SRP and baselines on datasets of different complexities and characteristics, each serving the purpose of evaluating the method's performance in different conditions. Five datasets were used, three simulated and two recorded, which are described below. The Cross3D and DOANet baselines use the same architectural parameters and training procedures described in their respective original papers [17], [211].

The network parameters for Neural-SRP are summarized in Fig. 5.4, where the tensor output shapes are shown for each of the network's layers. Convolutional kernels of size $(3, 3)$ were used on all convolutional layers. Max pooling with a kernel size 2 was applied to the GCC-PHAT dimension after all but the last convolutional layers. Parametric Rectified Linear Unit (PReLU) activation was used for all of the network's layers, apart from the RNN and DOA MLP output, which used a Hyperbolic Tangent (TANH) activation, and the activity output layer, which used sigmoidal activation. This architecture was chosen empirically. All the networks were implemented using the Pytorch library. The Adam optimizer was used for backpropagation.

A rectangular grid of size 64×32 was used for SRP, where the first dimension represents azimuth and the second elevation. The same configuration was used for generation of the input maps for the Cross3D baseline. The parameters used for the latter and the DOANet baseline were chosen similarly to those used in the respective original papers [17], [89].

5.3.1 Datasets

The first three experiments were performed using simulated datasets, which are referred to as SimSW, SimDirect and SimRandMic. All datasets contain samples of a source signal moving in a 3-dimensional sinusoidal trajectory inside a cuboid-shaped reverberant room containing a compact stationary microphone array. The trajectories were generated by

randomly selecting a start and end point inside the room, followed by randomly assigning a 3-dimensional vector referring to the frequency of oscillations within each direction. Finally, a second 3-dimensional vector is randomly generated representing the amplitude of each direction's oscillation. As in [89], the simulated datasets follow an "infinite-style" paradigm, meaning acoustic scenarios are randomly generated using the image source method [20] during training, i.e., no data is stored. The duration of each sample is 20 s. The ranges of the parameters are shown in Table 5.2. The sampling rate used for the simulations was equal to 16 kHz.

Both the first and second datasets, named SimSW and SimDirect use the pseudo-spherical array geometry of the NAO Robot as described in the LOCATA dataset [3]. SimSW and SimDirect differ in the type of noise used, respectively, spatially white (SW) sensor noise and directional noise. The goal of these datasets is to assess the robustness of the algorithms to different types of noise. For the third dataset, named SimRandMic, a random array geometry was generated for each dataset sample. The goal of this dataset was to assess the methods' generalizability to unseen microphone geometries.

Parameter	Min. value	Max. value
RT60 (ms)	0.2	1
SNR (dB)	5	30
Oscillations	0	2
Oscill. amp. (m)	0	1
# mics. (SimMicRand)	4	12
Array radius (SimMicRand, cm)	5	10

Table 5.2: Parameter ranges for simulated datasets.

The datasets were generated using the gpuRIR Python library [211], which can simulate audio recordings of cuboid-shaped, reveberant rooms including an arbitrary number of moving sources and microphone arrays. Simulating moving sources/microphones is a computationally expensive task, as high quality scenes are typically rendered by generating one Room Impulse Response (RIR) using the Image Source method between each source-microphone pair at every few milliseconds, and auralizing audio signals by convolving the source signals and RIRs using the Overlap-Add strategy [211]. gpuRIR significantly reduces the computational time in comparison to other libraries such as Pyroomacoustics

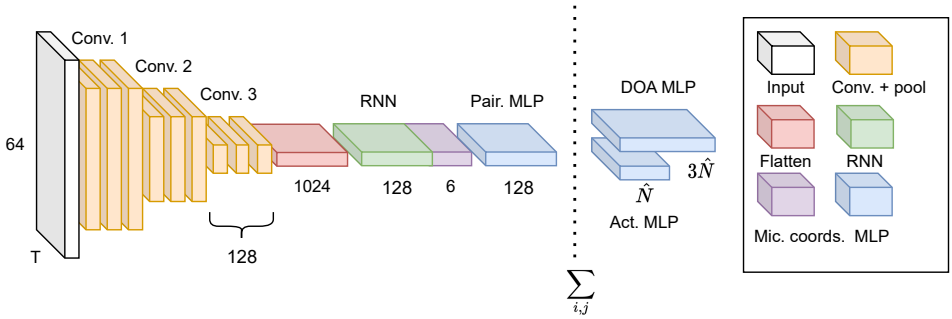

Figure 5.4: Detailed view of Neural-SRP architecture, where the numbers show the output dimension of each layer. The dotted line separates the pairwise network $\mathscr{P}$ from the global decoder $\mathscr{D}$, which receives the sum of pairwise features as its input. The input layer consists of T frames of 64 central GCC-PHAT bins each. The mic. coords. layer is of shape $(T, 6)$ where the three coordinates for each of the microphone in the pair are replicated for all frames.

[171] by generating the RIRs in a Graphics Processing Unit (GPU).

In the SimSW and SimRandMic datasets, random Gaussian white noise was added to the auralized (i.e. reverberant) signals at the desired SNR, computed using (5.9). In the SimDirect dataset, a second source emitting random Gaussian noise was randomly placed at a static position inside the room. The auralized noise signal was added to the source signal by scaling it to the desired SNR, computed using the mean energy of both auralized signals across all frames. In the SimRandMic dataset, a spherical microphone array is generated for every sample first by uniformly sampling its radius and number of microphones from the values ranges shown in Table 5.2, followed by randomly placing the microphones on the sphere's boundary. An utterrance from the LibriSpeech dataset [179] is randomly chosen as source signal for each dataset sample. Each epoch consists of a network pass through all of the Librispeech dataset, although different scenes are generated for each epoch.

The idealized noiseless signal frame received at microphone m is defined as $\mathbf{y}_m(t) = \mathbf{x}_m(t) - \epsilon_m(t)$. The average array-wide power of all signal frames p_y is defined as

$$p_y(t) = \frac{1}{LMT} \sum_{l=0}^{L-1} \sum_{m=1}^{M} \sum_{t=0}^{T-1} z(t) \mathbf{y}_m(t, l)^2, \tag{5.8}$$

where the ideal binary voice activity detector z is used to ignore silent frames. The array-wide power of each noise frame p_ϵ is defined analogously. The array-wide *spatially white*

$\mathrm{SNR}_{\mathrm{sw}}$ is computed as

$$\mathrm{SNR}_{\mathrm{sw}} = 10 \log_{10} \frac{p_y}{p_\epsilon}. \tag{5.9}$$

The LOCATA dataset [3] was released as part of the 2018 IEEE AASP Challenge on acoustic source LOCalization And TrAcking. It consists of 6 tasks of increasing complexity. In this work, only tasks 1, 3 and 5, namely, static, moving source and moving microphone localization, are selected. The dataset provides recordings from multiple microphone arrays. In this work, only recordings provenant from the NAO robot, which contains a pseudo-spherical 12-channel microphone array, were used. The goal of this dataset is to assess the performance of the algorithms in a real environment, as well as their ability to generalize to a real environment through training on simulated data. The sampling rate of the dataset is 48 kHz.

The TAU-NIGENS Spatial Sound Events dataset [212] was originally released for the Sound Event Localization and Detection task of the 2021 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge. It was generated by filtering source signals from the NIGENS Sound Events database [213] using time-varying RIRs recorded on 13 different rooms of Tampere University, Finland. These RIRs were recorded using a 32-channel Eigenmike spherical microphone array and a Genelec G Three [2] loudspeaker. Instead of providing the full 32-channel recordings, equivalent compressed 4-channel tetrahedral signals are provided. The dataset is subdivided into 400 training, 100 validation and 100 testing 1-minute recordings of up to two simultaneous moving sources. The samples may be corrupted by directional, moving interference emitting signals belonging to a noise class from the NIGENS database. The goal of this dataset is to assess the performance of the Neural-SRP method for tracking multiple sources. The sampling rate of the dataset is 24 kHz.

5.3.2 Experiment 1: spatially white noise

In this experiment, the performance of Cross3D, Neural-SRP and conventional SRP were evaluated in the presence of independent WGN added to each sensor. The neural models were trained for a duration of 80 epochs using a learning rate of 10^{-4}. As in [211], a frame size of 256 ms and a hop size of 192 ms were used. The noise signals were generated with

[2] https://www.genelec.com/g-three

unit variance, then scaled to the randomly selected SNR_{sw} by inverting (5.9), and then summed to the noiseless received signals. Both networks were trained using the range of SNRs defined in Table 5.2, and tested using a simulated dataset of unseen source signals from the Librispeech test set, as well as the unseen LOCATA dataset. The results are shown in Table 5.3. The dependence of the localization error to reverberation and noise on the test dataset are also shown in Fig. 5.5.

Model	SimSW.	LOCATA. (O)	SimDirect.	LOCATA. (D)
Cross3D	4.2	6.1	3.5	6.1
Neural-SRP	3.2	4.7	3.4	5.8
SRP	8.4	16.7	7.9	16.7

Table 5.3: Average localization error for experiment 1 (first and second columns) and 2 (third and fourth columns). All values are expressed in degrees. LOCATA (O) and (D) are the results following training using SimSW and SimDirect respectively. Both entries in the aforementioned columns show the same value for SRP, as it is not trained.

5.3.3 Experiment 2: directional noise

In this experiment, the Neural-SRP method was applied to the task of localizing a single speech source on a directional noise scenario, which is arguably more realistic than the diffuse case. For example, a directional noise source could be a fan, or a washing machine. The main difference from the experiment described in Sec. 5.3.2 is that, instead of adding independent noise to each microphone, the noise is itself modeled as a source in the room. In other words, for each training sample, the interferer is randomly placed within the room, with the restriction of being at least one metre away from the source and array. Then, a RIR between the microphones and interferer is computed, which is then convolved with a random unit variance Gaussian signal. Finally, the auralized result is scaled to the randomly assigned SNR in the same manner as (5.9). The results are shown in Table 5.3.

5.3.4 Experiment 3: testing on an unseen geometry

To assess the proposed model's ability to generalize to unseen microphone geometries, a dataset of multiple microphone array geometries is used, while testing it on the micro-

phones mounted on the NAO robot head of the LOCATA dataset, a geometry which is unseen in the training dataset. Although the Cross3D method can be theoretically trained using variable microphone geometries, training as unsuccessful using the SimRandMic dataset as the initialization of the SRP method was shown to be prohibitively costly, resulting in each epoch taking several hours on a GPU-enabled server. As a means of comparison baselines, conventional SRP, as well as Neural-SRP trained using the SimSW dataset, i.e., a matched array geometry are used. The results can be seen in Table 5.4.

Model	Trained on	SimSW (°)	LOCATA (°)
Neural-SRP	SimRandMic	6.7	6.0
Neural-SRP	SimSW	6.7	3.4
SRP	N/A	7.9	16.7

Table 5.4: Average localization error for experiment 3

5.3.5 Experiment 4: multi-source tracking

Model	Loc. err. (°)	Precision (%)	Recall (%)	F1 (%)
DOANet	9.4	88.6	81.9	85.1
Neural-SRP	8.2	92.0	79.9	85.4

Table 5.5: Average multi-source metrics and standard deviations of Neural-SRP and DOANet on the testing TAU-NIGENS dataset. Metrics and deviations were computed by averaging across the 3 training experiments.

In this experiment, the Neural-SRP method is applied to the task of multi-source tracking. The proposed method is compared to the state-of-the-art DOANet model with parameters described in [17] on the TAU-NIGENS dataset. The network was trained three times for a duration of 80 epochs using a learning rate of 10^{-4}. The average localization error was computed on the validation dataset at the end of each epoch, and the network weights that obtained the lowest validation localization error were used for evaluating the unseen test set. The results are shown in Table 5.5, where each value is the average metric obtained for each training round. The metrics used were the localization error in degrees, for true positive matches, as well as classical tracking metrics, namely, precision, recall, and the F1 score, defined as a geometric average of the two aforementioned scores. An example output of Neural-SRP successfully tracking two simultaneous sources is shown in Fig. 5.1. As in [17], a frame of size 20 ms with a hop size of 10 ms was used.

5.3.6 Complexity comparison

In this section, the complexity of the proposed Neural-SRP model and baselines is presented in terms of number of parameters, computational time and number of Floating-Point Operations (FLOPS) for microphone array sizes {4, 8, 12}. The number of FLOPS is obtained through the use of the THOP Python library [3]. This library is not compatible with the SRP implementation, so the complexity of the latter is computed theoretically, as in [70]. The inference time is measured as the time difference taken for the model to produce an output for an input stimulus of duration of one-second. These results were obtained using a 16 GB Macbook Pro with an M1 chip and are shown in Table 5.6.

Model	# params (10^6)	Inf. Times (ms)	FLOPS (10^9)
Cross3D	5.62	398, 423, 450	20
DOANet	1.57, 1.59, 1.64	11, 20, 35	0.1, 0.2, 0.3
Neural-SRP	0.92	13, 75, 149	0.45, 2.1, 4.9
SRP	0	8, 25, 54	0.39, 1.8, 4.3

Table 5.6: Complexity analysis of Neural-SRP and baselines. The cells containing three numbers refer to 4, 8 and 12 microphones respectively.

5.4 Experimentation for PSSL

5.4.1 Datasets

This section describes the different datasets used for training and evaluation of Neural-SRP and the baselines. To assess the dependence of Neural-SRP's performance on the number of microphones used, two testing variants of each dataset were produced, one containing 4 microphones and one containing 6 microphones. Furthermore, to train the baseline CRNN, training and validation datasets containing 4 and 6 microphones were used. No combination of microphone and source positions overlap between training, validation and testing datasets. The microphone signals used had a duration of 0.5 seconds.

The **AnechoicSim** dataset is a simulated dataset with no reverberation, used to train the first stage of the network. It consists of 10000 training, 2500 validation and 2500 test

[3] https://github.com/Lyken17/pytorch-OpCounter

samples. The source signal used are speech samples from the VCTK corpus [177]. For each dataset sample, the room's width, length and height are randomly uniformly chosen from the respective intervals of $[3, 10]$, $[3, 10]$ and $[2, 4]$ metres. The source and microphone positions were uniformly sampled within the room's dimensions, with the restriction of each device pair being at least 0.5 m apart.

The **ReverbSim** is a dataset generated using the Image Source method [20] and speech samples from the VCTK corpus, used to train the network at a second stage. Additionally to using the same random room dimension, microphone and source positions ranges of AnechoicSim, the reverberation of each room is set using the *reflectivity-biased* procedure [18], which assigns one absorption coefficient per surface, generating more realistic rooms than creating surfaces with a shared coefficient. It consists of 10000 training, 2500 validation and 2500 test samples.

The **Recorded** dataset [178] contains real recordings from a single room with a high reverberation time of 800 ms containing 40 microphones and 4 loudspeaker positions. The sound source is a single loudspeaker emitting speech samples from the LibriSpeech [179] corpus. This dataset was used for evaluating the proposed model and baselines, as well as for fine-tuning the neural models. It consists of 250 training samples and 2500 testing samples.

5.4.2 Methods and baselines

The proposed approach is evaluated against the classical SRP method using a similar approach to [214]. Furthermore, experiments are conducted training Neural-SRP using two or three stages. The two-stage approach, referrred to as *NeuralSRP*, consists of training the network using simulated data as defined in Equation 5.2.4, firstly by training the network on the AnechoicSim dataset, followed by training it on the ReverbSim dataset. The three-stage approach, denoted *NeuralSRP+*, is further training the *NeuralSRP* model using a small subset of the Recorded dataset. Using recorded data to complement synthetic training was used in other works such as [84], as simulated data may not completely match real scenarios.

To compare Neural-SRP against a trained baseline, the CRNN4 and CRNN6 models are proposed, which share a similar architecture to and number of parameters to Neural-

SRP, but differ in the input and output. CRNN4 and CRNN6 jointly process respectively 4 and 6 microphone signals, and therefore have to be trained specifically for each case. The architecture used is therefore the same as shown in Fig. 5.2, but the input to the Convolutional blocks is $(4 \times N \times F)$ for CRNN4, and $(6 \times N \times F)$ for CRNN6. Furthermore, a single output grid is produced instead of summing grids from all pairs. For that reason, the aforementioned models were trained using the Pseudo-Gaussian target as described in (5.5).

5.4.3 Experiment details

A public repository containing all methods is provided on Github, as well as a demonstration website for data access and reproduction [4]. Pytorch [172] was used as the main deep learning library, along with Pytorch Lightning [173] for abstracting common training routines. Pyroomacoustics [171] was used for generating the AnechoicSim and ReverbSim datasets.

The networks were trained for a maximum of 50 epochs with early stopping if the validation loss stops increasing after 3 epochs. A learning rate of 0.0005 using the Adam optimizer [176] was employed. Batch size of 16 samples was used. These parameters were chosen through multiple training runs using the validation dataset. All grids used are of dimensions 25×25. The neural networks contain four convolutional layers using 1×2 kernels and filter sizes and a MLP containing 3 layers, each of output size 625. The total number of parameters for the neural network methods is around one million. A ReLU activation function was used for all layers except for the output, which uses no activation. The source estimation procedure in the baselines and proposed methods consist of picking the location of the highest value, and GCC-PHAT is used to compute the cross-correlation ~~within the SRP baseline.~~

[4] https://github.com/egrinstein/gnn_ssl

5.5 Discussion and analysis

5.5.1 DOA estimation experiments

The single source experiments summarized in Table 5.3 show that Neural-SRP obtains favourable results in comparison to the Cross3D method both in the spatially white and directional noise scenarios, despite using a significantly smaller and more computationally efficient model. Like Cross3D, Neural-SRP can be trained using simulated data and tested using real recordings, as seen in the LOCATA results in Table 5.3. This is remarkable as, unlike Cross3D, Neural-SRP is required to learn its own spatial representation of sound. In other words, Neural-SRP is able to generalize despite having a less stringent inductive bias. Another relevant remark is that unlike SRP, Neural-SRP and Cross3D were able to eliminate the effect of a directional noise source, which is typically manifested as an additional peak in the GCC-PHAT (and therefore SRP) features.

The method's dependence of localization error to reverberation and SNR is shown in Fig. 5.5. The error of SRP increases significantly with high reverberation and low SNR, whereas Neural-SRP's error increases less significantly in those conditions. Fig. 5.5 also shows consistent incremental gains of Neural-SRP in comparison to the Cross3D baseline throughout all reverberation times and SNRs.

As shown in Table 5.4, Neural-SRP was able to be trained on a set of microphone geometries and tested on an unseen microphone geometry with only a small reduction in localization performance. This reduction is however expected, as the prolate spheroid (American football) geometry of the NAO array is not represented in the training dataset.

Turning to the multi-source experiment shown in Table 5.5, Neural-SRP achieves an improved localization performance in comparison to the DOANet method, as well as comparable tracking metrics. An important remark is that this increased performance is achieved despite the fact that the DOANet is able to obtain non-causal frame information, as a bidirectional RNN is employed by the latter, which also incurs in a greater number of parameters. A possible explanation for this increased performance is that the Neural-SRP pairwise architecture is more parameter-efficient than the DOANet's global architecture, which has to employ neurons to replicate information for each pair.

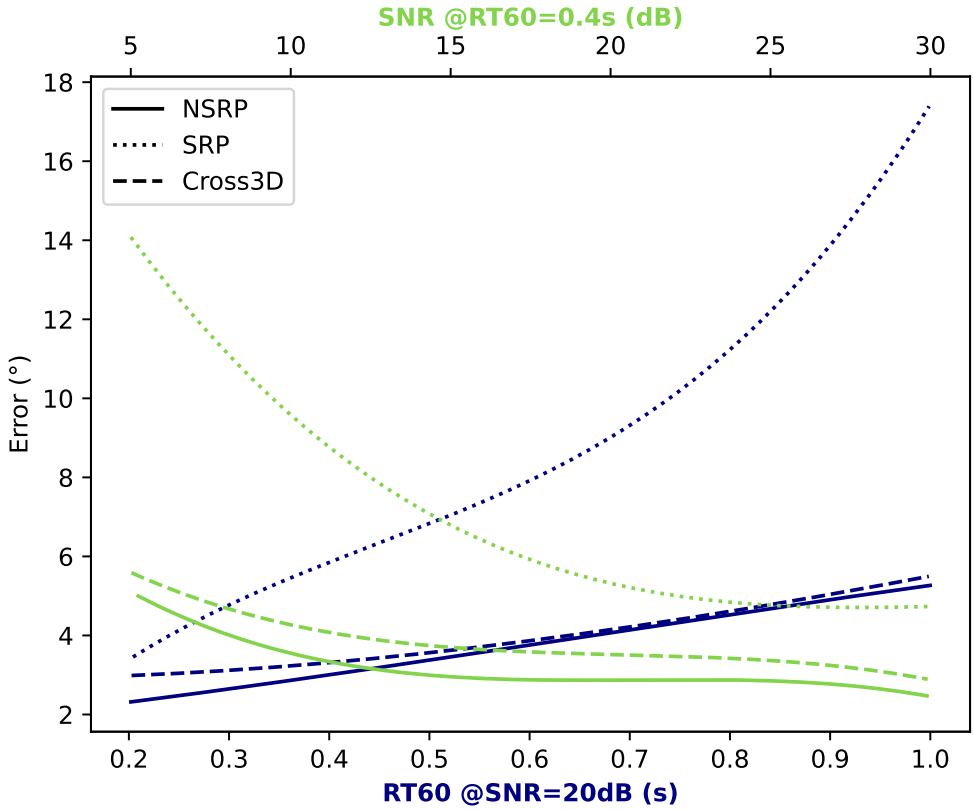

Figure 5.5: Localization error comparison between Neural-SRP, Cross3D and SRP for increasing levels of reverberation and SNR. The curves were smoothed using cubic interpolation.

Finally, as shown in Table 5.6, the Neural-SRP uses significantly fewer parameters than the other neural baselines, namely, over 6 times fewer parameters than Cross3D and a little over half as many as DOANet. In terms of computational complexity, Neural-SRP is positioned in-between Cross3D and DOANet, being at least 3 times faster than the former, and showing comparable performance with the latter in the case of a 4-microphone array. The proposed method's increase in computational cost is due to its pairwise formulation, which introduces a quadratic dependence with the number of microphone pairs ($M(M-1)/2$). However, this pairwise formulation also introduces flexibility, as microphone selection procedures such as [215] can be applied to reduce the number of pairs. The pairwise formulation also allows for distributed computing and only requires pairs to be synchronized, which is of particular relevance when using a distributed microphone network [188].

5.5.2 PSSL experiments

The error $\|\hat{\boldsymbol{u}} - \boldsymbol{u}\|$, i.e. the distance in metres between the predicted and actual source positions, was employed as the main metric of comparison. As a global comparison metric, the average error for all samples is computed. A table comparing the proposed methods and baselines is shown in Table 5.7.

The highlighted results in Table 5.7 show that NeuralSRP and NeuralSRP+ obtain the best results in terms of average localization error, overcoming the classical SRP method. The relative improvement between Neural and classical SRP was of 59, 54, 67, and 34% for the ReverbSim4, Recorded4, ReverbSim6, and Recorded6 datasets respectively. Interestingly, the CRNN baselines were unable to surpass the performance of SRP. An explanation for this performance may be that the Gaussian grid is less effective for training due to its sparse nature. The performance of NeuralSRP increased when using 6 microphones in comparison to 4, as expected.

In qualitative terms, an example comparison between grids generated by classical and Neural-SRP can be seen in Fig. 5.6, showing that while SRP produces a noisy grid with peaks not located at the true source position, Neural-SRP produces a smooth grid with the maximum located at the true source's position.

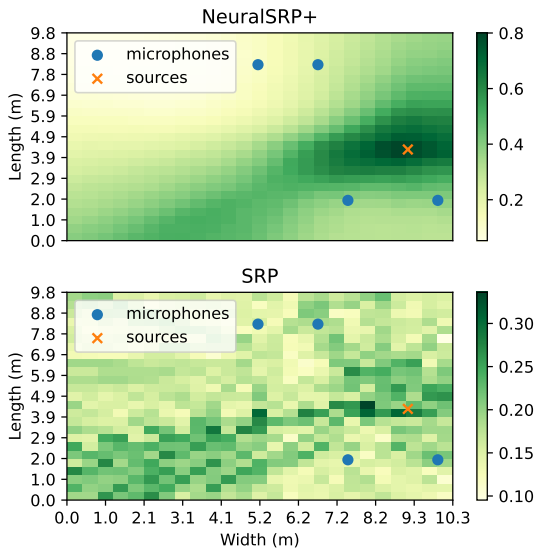

Figure 5.6: Neural-SRP and classical SRP output for real recorded signals.

Dataset	Model	Avg. err. (m)	Std. (m)
ReverbSim4	SRP	1.86	1.46
ReverbSim4	NeuralSRP	**1.17**	0.87
ReverbSim4	CRNN4	2.85	1.53
ReverbSim6	SRP	1.51	1.32
ReverbSim6	NeuralSRP	**0.90**	0.66
ReverbSim6	CRNN6	2.49	1.29
Recorded4	SRP	1.19	1.53
Recorded4	NeuralSRP+	**0.77**	0.66
Recorded4	CRNN4	3.78	1.63
Recorded6	SRP	0.75	1.04
Recorded6	NeuralSRP+	**0.56**	0.46
Recorded6	CRNN6	2.91	1.77

Table 5.7: Comparison of the mean and standard deviation of the localization error between multiple datasets and models.

5.6 Conclusions

This chapter presented The Neural-SRP, a state-of-the-art localization neural network which is able to overcome limitations of previous neural methods. Besides providing incremental gains in terms of localization performance, Neural-SRP is causal and shows a low computational complexity. Finally, Neural-SRP is the first method that was shown to work on unseen array geometries.

Chapter 6

A Generalized framework and review for the SRP method

This section provides a review of over 200 papers which either apply or modify the SRP method for source localization, followed by the creation of a modular framework and open-source software which allows for the reproduction of most SRP variants.

6.1 Introduction

Sound Source Localization (SSL) is the task of estimating the position of one or more active acoustic sources using one or more microphone arrays. Applications for SSL include event detection [42], [54], [204], camera steering [58], and sound source separation [32], [34], [205] among many others. In the last decades, many classical signal processing-based methods were developed for SSL, including MUSIC [216], ESPRIT [94], TDOA-based [62], [217], ML-based [63] and Steered Response Power (SRP) [68], [206], which is the focus of this review. Alternatively to signal processing-based methods, significant research interest has also been devoted to machine learning-based localization methods [79].

Choosing a localization method from all the available methods depends on the type of available acoustic and computational resources, assumptions about the localization scene, and knowledge of the method's mathematical formulation. SRP is known for its straightforward formulation and robust performance in many realistic environments [76].

A historical disadvantage of the method has been its significant computational complexity, although this is of diminishing importance due to the increased computational capacity of today's devices and to the many optimized modifications of SRP which have been developed. This has resulted in SRP becoming a standard SSL method in the literature.

Besides reducing its computational complexity, dozens of SRP variants have been developed to improve aspects of its performance, including increasing its robustness in adverse environments or in specific scenarios, and allowing multiple sources or moving sources to be localized. SRP can also be used as a feature extractor for neural-based localizers [89]. Therefore, one must not only choose SRP as a localizer, but must also decide which of the multiple SRP 'flavours' to use. A prominent flavour is the SRP-PHAT method, which uses the GCC-PHAT [64] method as its correlation function, which is shown to offer advantages to other correlation functions for processing speech signals. Unless stated otherwise, the term SRP refers to SRP-PHAT throughout this work.

The goal of this chapter is to provide a centralized resource for SRP research, to be used by both newcomers and experienced practitioners in the field of SSL. Over 200 papers are classified, described and compared, followed by the developement of a modular description of the algorithm, which can be used to develop implementations. A code library named X-SRP is also released as part of this work, with the goal of facilitating the usage of the algorithm. The remainder of the chapter comprises the following sections:

2. *Reducing SRP's complexity and computational time*, which discusses papers that focus on reducing SRP's computational cost at a minimal decrease in localization performance.

3. *Increasing robustness*, which focuses on improving SRP's performance on reverberant and noisy environments using, for example, neural network methods.

4. *X-SRP*, where a modular description of SRP is provided by decomposing the algorithm into functional building blocks. Each of the reviewed papers usually modify a single block in the proposed framework, allowing works to be combined and altered. The framework is applied through an open-source Python implementation of SRP denoted X-SRP, or eXtensible-SRP, with the goal of facilitating collaboration in the

field. The released code [1] includes implementations of many popular SRP variants.

5. *Conclusion*, where a discussion of future research directions is provided and the work is concluded.

6.2 Reducing SRP's complexity and computational time

6.2.1 Coarse grids and Volumetric-SRP

As mentioned above, reducing G is a straightforward strategy for reducing SRP's complexity. When applying equispaced grids such as those described in (2.12) and (2.13), this can be achieved by reducing the resolution parameters $R^{(1)}$, $R^{(2)}$ and $R^{(\psi)}$. However, this comes with the risk of not sampling the true source location, which may lead to the peak of the cross-correlation function not to be projected into the map, leading to a high localization error [218]. Nonetheless, many strategies can be applied to increase the localization performance of approaches using coarse grids.

As grids become coarser, each point is associated with an increasingly larger spatial region or volume. It is therefore reasonable to devise a way to modify SRP's operation to take into account the entire set of points around the candidate. Methods employing this strategy are referred to as *volumetric* SRP (V-SRP) [219]–[221]. An example comparison between conventional and volumetric SRP maps is shown in Fig. 6.1. The volume surrounding a candidate position is defined as

$$\begin{aligned}\mathcal{V}(\mathbf{u}) = \{ & \\ & [x\ y\ z]^T \mid \\ & |x - u^{(1)}| \leq r^{(1)}/2 \\ & |y - u^{(2)}| \leq r^{(2)}/2 \\ & |z - u^{(3)}| \leq r^{(3)}/2 \\ \}, & \end{aligned} \tag{6.1}$$

where $r^{(1)}$, $r^{(2)}$ and $r^{(3)}$ respectively represent the width, length and height of the volume. The Volumetric SRP (V-SRP) approach is typically defined by considering the SRP value

[1] https://github.com/egrinstein/xsrp

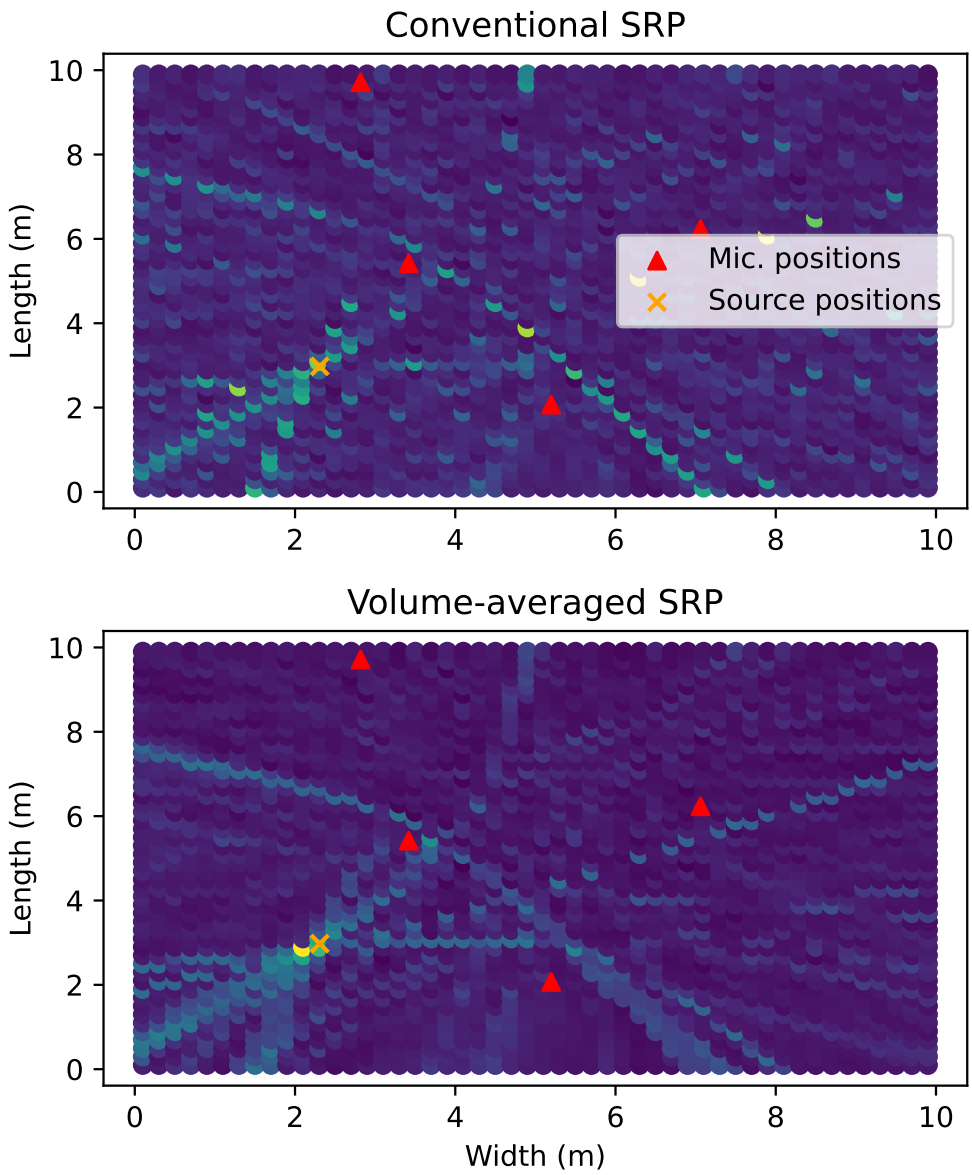

Figure 6.1: Comparison between SRP maps generated with (bottom) and without (top) volumetric techniques.

of all points within the volume, which are then combined using a pooling function such as summation. The pairwise V-SRP function can thus be defined as

$$\text{V-SRP}_{lm}(\mathcal{V};\mathcal{X}) = \sum_{\tau=\min(\tau_{lm}(\mathbf{u}\in\mathcal{V}))}^{\max(\tau_{lm}(\mathbf{u}))} \mathbf{g}[\tau \mid \mathbf{x}_l, \mathbf{x}_m]. \tag{6.2}$$

Different approaches and approximations can be used to find the summation interval in (6.2). A popular approach is the Modified SRP (M-SRP) [214], which approximates the minimum and maximum TDOA limits in the volume by first remarking that, due to the hyperboloidal nature of TDOAs, the extremes must be contained in the volume's boundary. These values are then approximated using the TDOA's gradient vector and the centre of the volume. Although summation is used as a pooling function in (6.2), it has been shown that average [222] or max [223] pooling may increase robustness to noise and reverberation.

The work of [219] proposes exacts bounds for the maximum and minimum and maximum TDOA limits used in the M-SRP algorithm [214] in anechoic conditions. In particular, the authors show that the minimum and maximum TDOAs of a cuboid volume can be always found by searching a set of only 26 points involving its vertices, edges and faces. Furthermore, this can be further appriximated by searching only the volume's 8 vertices, further simplifying finding the maximum and minimum TDOAS as these limits can be precomputed for any given cuboid and microphone array locations. The computational complexity of M-SRP (6.2) can be further reduced through an iterative subdivision of the maximal volume [219], [222], [224], [225].

6.2.2 Iterative grid refinement

A common strategy used in conjunction with coarse grids consists of iteratively modifying the initial search grid $\mathcal{G}(0)$ based on the candidate position's SRP values, allowing for the algorithm to 'focus' on promising regions. This procedure can be applied repeatedly until a stopping condition is reached, i.e.,

$$\mathcal{G}(i) = \text{ITERATE}(\mathcal{G}(i-1)), \tag{6.3}$$

where the ITERATE function usually involves evaluating the SRP function on the current grid candidate points, discarding points based on a criterion, and generating additional candidates based on some heuristic.

This iterative procedure may be performed using a quadtree [157], [226], a tree-based data structure commonly used for image processing. In [157], each cell of an initial azimuth-elevation square grid of size 16×16 is iteratively subdivided into four non-overlapping cells, where the SRP function is computed on each region's centre. To prevent the grid size from growing exponentially, only the cells with the highest SRP value are selected for further division.

When a coarse grid is used, the true source location $\mathbf{u}$ may lie on a grid point. A strategy to ensure $\mathbf{u}$'s neighbours exhibit a high SRP value in initial iterations was proposed in [157], which identify that the width of a peak on an SRP map is inversely proportional to the source's carrier frequency. Therefore, computing SRP using only low frequencies produces a smoother map. This is illustrated in Fig. 6.2, where only frequencies below 200 Hz are used for $\mathcal{F}$, which can be compared to Fig. 2.5 which shows a map generated using all frequencies up to the Nyquist rate.

The initial grid can also consist of points randomly sampled on the room's boundaries, as formulated in the Stochastic Region Contraction (SRC) method defined in [50]. The region contraction procedure is exemplified in Fig. 6.3. The subsequent grid can be chosen by resampling a set of points on the smaller boundary containing the previous candidates exhibiting the highest SRP values. This procedure may continue for a maximum number of iterations, or until a minimum search cuboid is obtained. Note that this contraction procedure can also be applied to deterministic grids. In this case, the SRP variant is referred to as Coarse-To-Fine Region Contraction (CFRC) [227].

Although the aforementioned methods significantly accelerate the computation of SRP, they provide no guarantees that the true source location will not be discarded, as they assume the SRP map to be a concave function with its maximum at the source location. The authors of [219]–[221] propose a procedure which theoretically guarantees not to discard the point maximizing the SRP function in anechoic conditions using the branch-and-bound iterative search method. The search starts by considering the entire search volume, typically the entire room, and subsequently divides it into smaller volumes

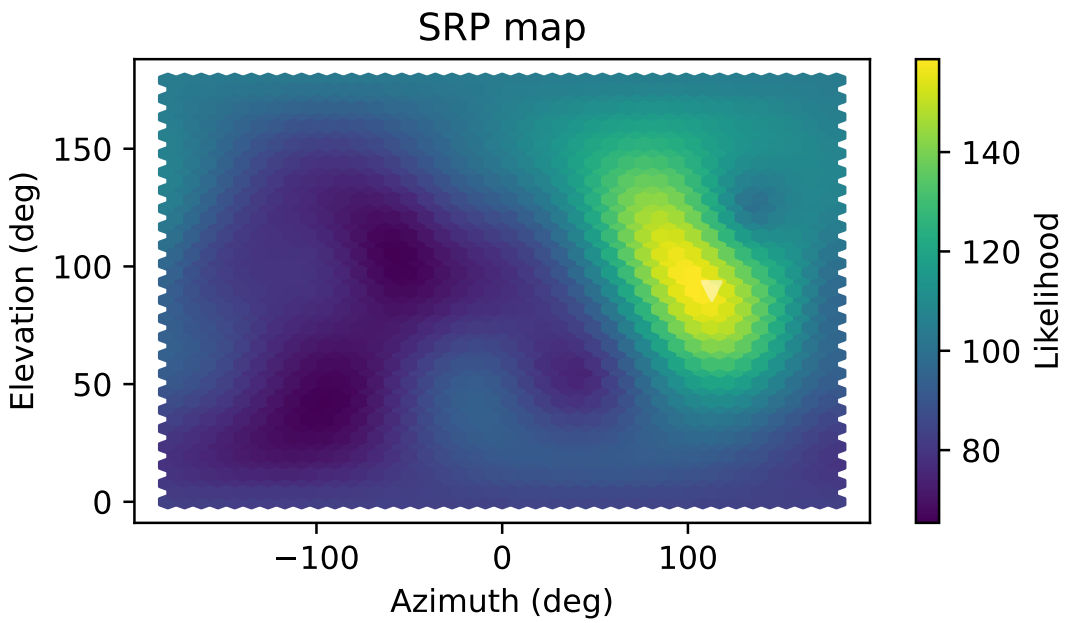

Figure 6.2: Low-pass version of the frequency-domain SRP, where only frequencies up to 200 Hz are considered.

using a branching function. Volumes are discarded through the aid of a bounding function similar to the bounds computed in (6.2).

Other iterative techniques used for SRP include the Artificial Bee Colony [228], Majorization-Minimization [229], [230] and Lagrange-Galerkin [231] search methods.

6.2.3 Grids based on prior location estimates

Alternatively, smaller grids can be built using lower-complexity, but less reliable source location estimators, such as those obtained using two-step methods. These candidates can then be more robustly selected and refined using SRP. In [158], the grid is initialized using the positions associated with the signals' highest GCC-PHAT peaks, which can be interpreted as estimates of the source's TDOA. These estimates are used to triangulate candidate source positions using a least squares approach, which are then evaluated using SRP. In [158], four peaks per pair were deemed to yield the best results.

As triangulation-based estimates are not robust to noise and reverberation, it is useful to include neighbouring points in the candidate grid, so as not to limit the performance

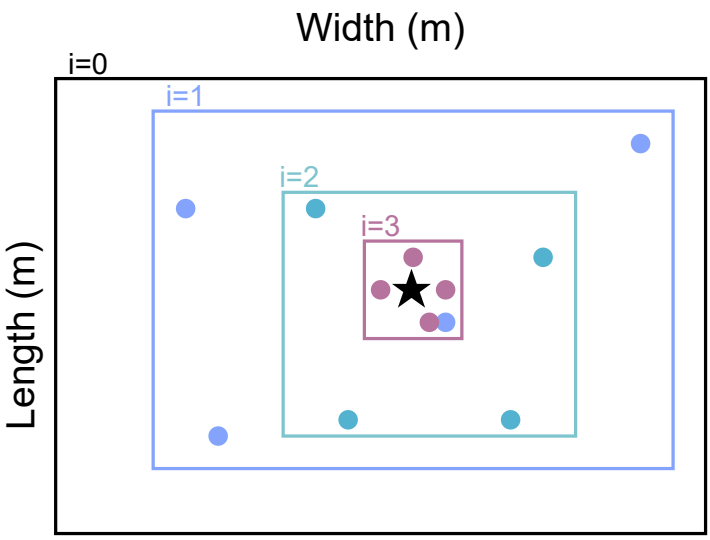

Figure 6.3: Iterative Region contraction procedure, where different colours represent search regions and grids of points related to iterations i. The true source location is represented by the black star.

of SRP. This can be achieved by sampling points in the cuboid region containing these candidates [232], [233]. Similar approaches are proposed by [234]–[237]. Grids based on prior location estimates were also explored on practical scenarios involving WASNs [232], [233].

6.2.4 Incorporation of prior scene information

Another strategy for reducing the grid size exploits the property that spatial regions exhibit different levels of sensitivity depending on their position in relation to the microphone array [238]–[242]. For instance, neighbouring candidate locations may have similar or identical sets of associated theoretical TDOAs, being therefore indistinguishable using SRP [238]–[240]. Those can therefore be replaced by their centroid without loss in performance [238]–[240].

A similar concept is proposed by [241], [242], where a non-uniform, *geometrically sampled grid*, is proposed. Based on their distance, each microphone pair within the system has a discrete set of integer TDOAs, in samples, each of which defines a hyperboloid in space. Candidate locations at the intersection of multiple hyperboloids have high definition, and can therefore be more reliably used for localization. Conversely, if the source is

located within a low-definition region, more grid points are used to improve its localization performance.

Alternatively, information about the environment can be included as prior information to build smaller grids. For example, for specific microphone array geometries such as the T-shaped orthogonal array used by [243], the 2D azimuth/elevation grid can be decomposed into two 1D grids, which can be independently maximized, significantly reducing the number of required SRP evaluations. In [244], a method combining SRP for both DOA estimation and PSSL is proposed, and tested with a large aperture, L-shaped microphone array. SRP is first used for estimating the source's DOA with respect to the array's branches. This direction is used to create the initial grid of candidate locations, from which the SRC variant of SRP is employed for 3D localization. A similar two-step approach is employed in [245], where distributed microphone arrays are used for DOA estimation. The intersection of these directions is then used to estimate the source location. The computational complexity of SRP can also be reduced, for linear arrays, by combining array interpolation and polynomial root solving [246]. Alternatively, if possible source locations are known, such as seat locations in a conference room, a database of possible source locations along with their respective microphone array responses can be precomputed, thereby significantly reducing the grid size [247].

Complexity can also be reduced by reducing the number of pairwise maps computed. For instance, centralized microphone arrays of symmetrical geometries such as spherical or rectangular exhibit multiple pairs of microphones with parallel directions. Computation can be reduced at a negligeable loss in performance by only using one pair for each of those directions [142]. Conversely, microphone pair selection can also be applied to distributed microphone networks, where data transmission is a secondary constraint which should be minimized. If each device contains at least two microphones, the SRP maps can be computed and transmitted independently for each device, an economic alternative to transmitting raw signals which was shown to incur only small losses in localization performance [248].

Finally, the computation of the SRP function can be avoided by only considering candidate positions with a high associated cross-correlation based on their theoretical TDOA and GCC-PHAT between microphone pairs [66], [249]. In practice, this can be

achieved by creating a hash table for each microphone pair where each key-value pair represents a TDOA and its set of possible candidate positions. The keys (TDOAs) with a low associated GCC-PHAT can then be filtered out. Finally, the table is traversed, where the SRP values for the remaining sets of TDOAs associated with a candidate position are summed to create a global SRP map.

6.2.5 Paralellization

When the device computing SRP supports parallel processing capabilities, such as multiple Central Processing Units (CPUs), multiple threads or one or more GPUs, the method can be sped up while using its original formulation, therefore guaranteeing its optimal performance. SRP is highly parallelizable, as the evaluation of the SRP function for each candidate location is independent.

A Compute Unified Device Architecture (CUDA) implementation of SRP was first proposed in [250], where the SRP function for each candidate location was computed independently on each GPU thread. In [251], a time-domain and a frequency-domain GPU implementation of SRP using CUDA were respectively compared with optimised CPU counterparts. Results show the GPU implementations resulted respectively in speed improvements of 70 and 275 times. In [252], the implementation provided by [251] is optimised by maximising usage of the GPU's internal memory in favour of the host's memory, resulting in significant speed-up in comparison to [251]. In [253], an implementation of SRP is proposed for three CUDA-enabled GPU types. In [254], [255], a GPU implementation of SRP using NVIDIA's Jetson chip, designed for low-power mobile computing, is evaluated for multiple grid resolutions. Conversely, in [254], a CUDA implementation of SRP using multiple GPUs is presented.

In [256], SRP's computation was vectorized using Intel's Integrated Performance Primitives (IPP) software library, reducing CPU load by a factor of two in comparison to a baseline scalar implementation. In [257], an implementation of SRP using OpenCL, an open-source parallel computing framework compatible with multiple processors including CPUs, GPUs and Field Programmable Gate Arrays (FPGAs), is presented. Experimental comparisons with device-specific implementations of SRP reveal that the proposed implementation achieves similar performance. An efficient hardware implementation of [70] is

presented in [258].

6.2.6 Other approaches

In [102], an SRP method based on the singular value decomposition (SVD) is proposed. Based on (2.11), a matrix is defined mapping all frequency-domain GCCs to all candidate locations, whereof a low-rank approximation is obtained using the SVD. This low-rank approximation allows to first project frequency-domain GCCs onto a subspace with reduced dimensions and subsequently employing a k-d tree search scheme [259], resulting in a lower computational cost at a similar localization performance to that obtained with the conventional SRP-PHAT. The performance of this method is increased in [96], where a spectral subraction procedure is applied to the correlation matrix.

It was shown in [70] that an SRP map can be efficiently approximated through interpolation while critically sampling the GCCs, based on Nyquist-Shannon sampling. Such approach is formulated while accounting for the physical bound over the range of possible TDOAs for a given microphone array, as well as the assumed GCC bandlimit. Simulation results indicate that the computational cost of the proposed interpolation-based approach for obtaining the approximated SRP map can be several orders of magnitude lower than the cost of computing the conventional SRP map, while the localization performance is maintained.

6.3 Increasing robustness

Although SRP has been shown to provide satisfactory performance in realistic scenarios [76], its performance is reduced in challenging scenarios including high reverberation and/or noise. Localization performance is often inversely related to the strategies presented in Sec. 6.2, as fine grids provide better resolution. However, other techniques are required to remove artifacts caused by noise and reverberation from the SRP maps.

6.3.1 Modified GCC-PHAT functions

The quality of SRP is dependent on the quality of the cross-correlation between microphone pairs. Most approaches employ GCC-PHAT to obtain the correlation information, as it

was shown to outperform temporal CC [68], [206]. Nonetheless, modifications can be employed to improve GCC-PHAT in challenging scenarios. One of such modification is GCC-PHAT$_\beta$, a parameterized version of GCC-PHAT which was shown to improve localization performance, defined as [97], [260]–[262]

$$\text{GCC-PHAT}_\beta(f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m) = \frac{\bar{\mathbf{x}}_l(f)\bar{\mathbf{x}}_m^*(f)}{|\bar{\mathbf{x}}_l(f)\bar{\mathbf{x}}_m^*(f)|^\beta + \gamma}, \tag{6.4}$$

where γ provides numerical stability, and β controls the relevance attributed to the signals' magnitudes. Note that conventional GCC-PHAT is achieved when $\beta = 1$, whereas conventional CC is obtained using $\beta = 0$. The experiments in [97] show that intermediary values of β (e.g., $\beta = 0.8$) improve localization of narrowband signals under the interference of directional noise sources at low SNRs. Although γ is often set to a small value to prevent a null denominator, Shen et al. [261] propose setting γ to the minimum *coherence* between the signal pair over all frequency bins. Coherence is here defined as the ratio between the signals' cross- and auto-spectral densities. In [263], the authors perform an experimental analysis of the SRP-PHAT$_\beta$ method and they verify the simulation study in [97] which shows the acceptable range of values for the partial whitening parameter β for a general signal to be between 0.65 and 0.7. They also point out that the experiments exhibit more significant performance fluctuations for especially $\beta = 1$ corresponding to the conventional PHAT method. This outcome supports the use of the partial whitening over the conventional PHAT.

An alternative to PHAT filtering consists of using the kurtosis of the signal pair, motivated by the assumption that noise is frequently modelled as a Gaussian random process, which is theoretically eliminated in the kurtosis computation [264]. The GCCs can also be replaced by a sum of Gaussians centered at the former's most prominent peaks, thus producing a smoother SRP map [265]. The effects of the phase transform can also be replaced by a linear predictor incorporating sparsity constraints [266]. The Multichannel Cross-Correlation (MCCC) function [267] can also be employed [268]. Instead of providing a single correlation value for two signals and a delay τ, MCCC provides a correlation value for a vector of M signals and a vector of delays $\boldsymbol{\tau}$. The MCCCs can therefore be used to construct a beamformer which is applied as a preprocessing step before SRP [268]. Finally, the CC between microphone pair signals can be computed using an eigenvalue decomposition of the cross-correlation matrix of the microphone signals. Instead of computing

the CC between microphone signals, the correlation between corresponding eigenvectors can be used, ignoring directions related to noise and reverberation [269] and therefore improving the quality of the SRP map.

The GCC-PHAT function of a broadband signal in an ideal, anechoic scenario is an impulse with its main peak occurring at the microphone pair's TDOA. However, as the source signal becomes narrowband, the pair's GCC-PHAT becomes a sinc function ($\text{sinc}(x) = \sin(x)/x$), i.e. a function exhibiting multiple ripples which translate into low-quality SRP maps. In this case, the envelope of the Generalized Cross-Correlation (GCC) function, obtained by extracting the magnitude of its analytic signal, can be applied instead to remove the aforementioned ripples.

In other broadband cases, some frequency bands may be more affected by noise than others. In those cases, it is advantageous to analyse the CCs in different frequency bands. This is done, for example in [270], which proposes the creation of a GCC matrix, where columns represent frequency bands and rows represent time delays. The conventional GCC-PHAT can be obtained from this matrix as long as the Constant Overlap-Add principle is satisfied when selecting the frequency band centres and widths. The authors show that degradations from noisy frequency bands can be reduced by applying SVD to obtain a low-order approximation of the GCC matrix, improving the robustness over the conventional GCC-PHAT.

Many challenges also arise when applying SRP in large outdoor environments. Firstly, these environments suffer from intense low-frequency environmental noise, often requiring the signals to filtered before processing, thus creating a band-passed input signal which introduces challenges for the SRP method as described in [271]. Secondly, the size of the search area may require very large grids, significantly increasing the method's computational cost. Finally, factors such as changes in temperature, terrain, wind and position of the sensors make the propagation time model defined in (2.1) unreliable. The authors of [272] propose a modified GCC function based on Wavelet theory which takes the three aforementioned factors into account to improve the performance of SRP in outdoor environments.

Finally, the GCC-PHAT function can be substituted by a neural network [273], as will be discussed in Sec. 6.3.4.

6.3.2 Improving combination

The formulation defined in (2.11) combines pairwise and frequency-wise SRP values through unweighted summation. A more general formulation of SRP, which is denoted Weighted SRP (W-SRP) can be written as

$$\text{W-SRP}(\mathbf{u} \mid \bar{\mathbf{X}}) = \bigcup_{(l,m)\in\binom{M}{2}} \bigcap_{f\in\mathcal{F}} \frac{\text{SRP}_{lm}(\mathbf{u}, f \mid \bar{\mathbf{x}}_l, \bar{\mathbf{x}}_m)}{k_f k_{lm}}, \tag{6.5}$$

where $\bigcap$ represents the operation combining frequency information, $\bigcup$ represents the combination of pairwise information, and weighting factors k_f and k_{lm} respectively weight frequency and pairwise information. Besides classical summation, choices for the pairwise combinator $\bigcap$ are the product $\prod$ and the Hamacher t-norm, among others [274]. Conventional SRP combines pairs through summation, meaning that pairwise SRP maps combined in such manner will exhibit high values if any pair does so. Conversely, if multiplication is used, all pairwise maps must exhibit high values for the global SRP to do so. In an extreme case, if any pairwise map is null, so will be the global SRP map. The simulated experiments in [274] show that combining pairwise SRPs through their product results in a significant increase in localisation performance over their sum, reducing the localization Root Mean Squares (RMSs) error by 45%.

The weights k_{lm} can be computed on pairwise SRP maps, for example, from a fractal theory standpoint, giving less importance to noisier, pairwise SRPs [275], or by measuring the noise of the GCC-PHAT vector by computing the ratio between the GCC-PHAT's peak and its average [276]. Note that microphone pair selection is also included in (6.5) for the special case $k_{lm} = \infty$.

Conversely, the frequency weight k_f can be set as the maximum SRP value across all pairs, therefore equalizing the contribution of each frequency bin to the global SRP. This is shown to offer a similar effect to the PHAT weighting [277]. Another approach estimates k_f using neural networks [278]–[281], as will be discussed in Sec. 6.3.4.

6.3.3 Pre/Post-processing

Applying pre- or post-processing to the microphone signals in search of anomalies may improve SRP maps. For example, a VAD can be used to detect the presence of speech in a

noisy environment, in order to prevent SRP from unintentionally localizing noise sources [282], or to improve the localization of impulsive sources [283]. A VAD can also be used to discard directional noise sources [284]. SRP maps can also be improved through the application of a Wiener filter [276].

6.3.4 Neural approaches

As in many other tasks in acoustic signal processing, neural networks have also been applied for the task of SSL, frequently obtaining state-of-the-art results in comparison to classical methods such as SRP [79]. However, SRP still presents several advantages over classical neural network methods, which usually require matched training/testing microphone geometries. Furthermore, SRP maps serve as an excellent input feature for neural networks. Finally, SRP's building blocks can be advantageously replaced by neural blocks, bridging the gap with neural methods' performance in challenging environments. The approaches below are related to the strategies mentioned in the above subsections.

One of such blocks which can be improved is GCC-PHAT. A deep neural block can be used to estimate an idealized GCC-PHAT vector which removes peaks associated with reverberation and noise. A Deep-GCC function can be formulated in the time [273], [285] or frequency [138] domain. In the time domain, the Deep-GCC vector should exhibit a single peak at the source's true TDOA τ_{lm}, modeled as a Gaussian with standard deviation σ_d as [273], [285]

$$\text{Deep-GCC}(\tau) = \exp\left(\frac{-|\tau - \tau_{lm}|^2}{2\sigma_d^2}\right). \tag{6.6}$$

In practice, (6.6) serves as the target loss function for the network being trained. The choice of input feature and architecture for a Deep-GCC function may vary. In [273], [285], GCC-PHAT itself is chosen as the networks's input and a 1-D Convolutional autoencoder is selected as architecture. In [138], the magnitude and phase spectrograms of both microphone signals are chosen as the input features, and a CRNN is chosen as the neural architecture.

Many approaches focus on using neural networks to estimate a weighting function, similarly to the signal processing based procedures described in Sec. 6.3.2. Most approaches focus on the frequency weights k_f, inspired by the task of speech enhancement, where neural time-frequency masks have attained significant success [278]. For instance,

a CNN can be trained to estimate a time-frequency mask to reduce the interference of directional sources, using the output of a Wiener filter as its target function [278]. Other targets can be used, such as the distance between the true and SRP-estimated locations for a single frequency band [279], [281]. Alternatively, the SNR on each microphone can be used as a weight for each frequency band [280]. Similar approaches been also employed other machine learning methods, namely, Support Vector Machines (SVMs) and Radial Basis Function Networks (RBFNs) [286]–[288].

Another prominent manner of improving localization performance using SRP uses the SRP maps as the input feature of a deep neural network. In this case, the neural network may have two goals: to enhance the maps produced by SRP [90], [91], and/or to extract the source locations using the map [9], [14], [89], [139], [258], [289], i.e., to improve the grid search/peak-picking function defined in (2.14). The networks differ in the architecture used, such as the MLP [9], [14], 3D [89], [258], [289], spherical [139] and icosahedral [90], [91] convolutions.

Finally, other neural approaches simulate the pairwise processing used by SRP for the task of source localization. In chapter 4, it is remarked that the SRP algorithm shares architectural similarities with the Relation Network, a type of GNN. In the context of SRP, a *relation* between two microphones consists of the pairwise SRP maps shared between them. All pairwise relations are then summed, creating a global relationship between all microphones, which can be used to estimate the source locations. Neural-SRP approaches chapter 5 therefore replace SRP's function with a neural network, reducing the detrimental effects of noise and reverberation by including challenging scenarios during network training. An example of a map produced using a Neural-SRP method is shown in Fig. 6.4.

6.3.5 Other approaches

SRP maps can also be analysed by decomposing them using a set of idealized pairwise maps, computed using the theoretical TDOA between the microphone pairs and the candidate locations. Instead of estimating the source location through peak-picking, the search can be done by matching the pairwise SRP maps with a subset of idealized maps according to a similarity metric [115], [290].

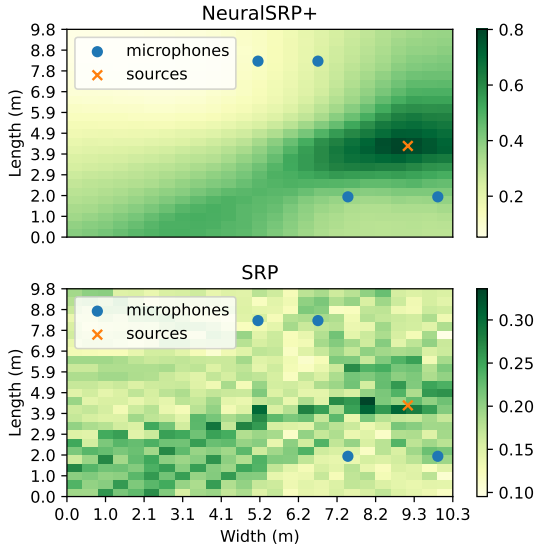

Figure 6.4: Neural-SRP+ (see chapter 5) conventional SRP map in a highly reverberant room. The source position is shown with a cross and the microphone positions with circles.

When the distance between microphones in a centralized array is small, so is the range of possible TDOAs between the pair as expressed in (2.4). It is therefore desirable to perform interpolation in the CC function to obtain sub-sample TDOA resolution when using the temporal SRP formulation. The work of [72] evaluates the performance of SRP for DOA estimation from concert hall recordings using three different interpolation methods, namely, parabolic, exponential and Fourier. The study reports best performance using exponential peak interpolation.

The work in [291] presents a system where, before performing localization using SRP, a speaker verification step to remove unwanted speakers and noise is applied.

In [292], the authors exploit spatial diversity in order to improve SRP's performance in reverberant environments. Their simulation results show that large arrays are affected by the reverberation more than smaller ones and that having a smaller distance between microphone arrays results in more accurate localization. When the number of microphones in an array is increased the localization results are more robust as expected, but separating it into two array makes it even more favourable compared to merely increasing the number of microphones in a single array.

In [293], a mel-frequency extraction technique is employed with Steered Response

Power with Phase Transform (SRP-PHAT) in order to obtain an enhancement of human speech and process it more robustly in a noisy environment. As a performance metric, peak SNR (PSNR) is used. The results show that utilizing Mel-frequency Cepstral Coefficientss (MFCCs) in conjunction with SRP-PHAT yields higher PSNR values compared to using only the SRP-PHAT, which results in a more accurate localization.

In [294], the authors compare the SRP-PHAT localization performances using a Uniform Linear Array (ULA) and a Coprime Microphone Array (CPMA) interleaving two linear arrays with coprime dimensions. They show that a CPMA offers better localization results than a ULA with the same number of microphones. In another study [295] by the same authors, a performance analysis of Semi-Coprime Microphone Arrays (SCPMAs) for localization using the SRP-PHAT algorithm is conducted. They evaluate the performance in terms of beam pattern, array gain and DOA estimation. The results on beam pattern an array gain suggest that the SCPMA outperforms the CPMA in reducing the peak side lobe level and minimising the total side lobe area. Moreover, it shows an enhanced ability to amplify the target signal while suppressing the noise. The results of DOA estimation in anechoic and low reverberant environments show that the SCPMA delivers accurate estimates which are on par with the estimates obtained from the full ULA. However, in highly reverberant conditions such as a 400 ms reverberation time, side lobes in the beam pattern of the SCPMA result in less accurate estimates.

As discussed in Sec. 1.4, the range ρ can only be accurately estimated when the source is located in the near-field with respect to the microphone array. The field type can be estimated by comparing the SRP of two circular candidate grids at different distances, one in the far-field, the other in the near-field. The grid exhibiting the highest SRP value dictates the field regime. If near-field conditions are found, a second SRP grid search can be applied for range estimation [296].

6.4 X-SRP

In this section, the SRP method is described from an algorithmic perspective, with the goal of unifying the previously described extensions of the method within a common framework. The main functionality described in Sec. 2.1.2 is revisited by substituting specific functions

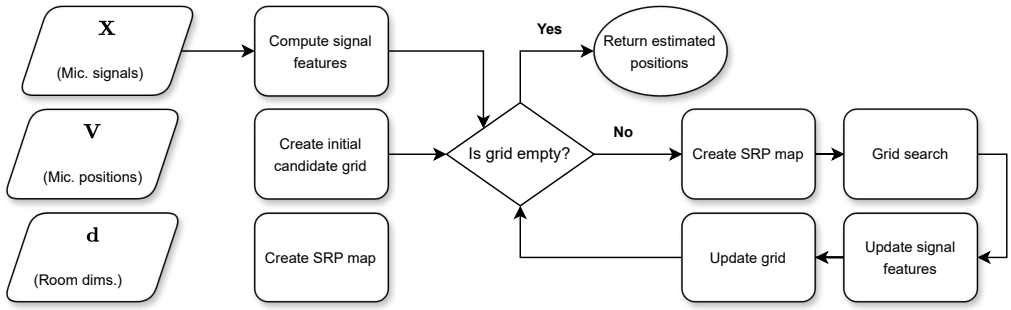

Figure 6.5: Flowchart of the generalized SRP algorithm. Parallelograms represent input data, rectangles represent functions, diamonds represent decisions and ellipses represent terminal states.

with generic ones, which are referred to as *modules*. For example, the CC function used in the classical SRP is substituted by a module called compute_signal_features, which can be instantiated as the temporal CC, GCC-PHAT, or a neural-based feature as in [138], [278].

This modular perspective allows for SRP papers to be grouped in an alternative way to the task-oriented manner used in the previous sections. Conversely, the categorization presented here groups the works by their implementation details, facilitating their combination and comparison.

To facilitate the reproduction of SRP variants and explore novel variants, the eXtensible-SRP, or X-SRP Python library is released, which provides a modular implementation of SRP following Alg. 2, which is also shown as a flow diagram in Fig. 6.5. multiple modules are included within eXtensible SRP (XSRP), which allow for selected variants to be implemented. The refer is referred to the project's repository [2] for further documentation on the library. The library includes several unit tests to ensure correct functionality of the SRP variants. Tests were carried using simulated data generated using Pyroomacoustics [171].

Algorithm 2 accepts three input parameters: A matrix $\mathbf{X}$ of microphone signal frames, a matrix of microphone positions $\mathbf{V}$ and a vector $\mathbf{d}$ containing the room dimensions. It is made optional as it is only necessary for SSL, not for DOA estimation. Note that configuration parameters such as the sampling rate f_s are omitted for the sake of conciseness.

[2] https://github.com/egrinstein/xsrp

Algorithm 2 X-SRP

1: **function** X-SRP($\mathbf{X}, \mathbf{V}, \mathbf{d} = \text{null}$)
2: $\quad \hat{\mathcal{U}} \leftarrow \emptyset$
3: $\quad \mathbf{G} \leftarrow$ create_initial_candidate_grid($\mathbf{d}$)
4: $\quad \mathbf{C} \leftarrow$ compute_signal_features($\mathbf{X}$)
5: $\quad$ **while** $\mathbf{G} \neq \emptyset$ **do**
6: $\quad\quad \mathbf{S} \leftarrow$ create_srp_map($\mathbf{V}, \mathbf{G}, \mathbf{C}$)
7: $\quad\quad \hat{\mathcal{U}} =$ grid_search($\mathbf{G}, \mathbf{S}, \hat{\mathcal{U}}$)
8: $\quad\quad \mathcal{C} =$ update_signal_features($\mathcal{C}, \hat{\mathcal{U}}, \mathbf{V}$)
9: $\quad\quad \mathbf{G} =$ update_grid($\hat{\mathcal{U}}, \mathbf{d}$)
10: $\quad$ **return** $\hat{\mathcal{U}}$

The first line initializes the estimated source coordinates $\hat{\mathcal{U}}$ as an empty set. $\hat{\mathcal{U}}$ is a set of points and not a single point to accommodate multi-source localization approaches.

Then, an initial grid of candidate positions $\mathbf{G}$ is created using the **create_initial_candidate_grid** module. In most SRP variants, this will be the only grid created. However, in iterative approaches such as [50], [219] as well as multi-source approaches [99], this function only provides an initial grid $\mathbf{G}$, which is further updated as part of their grid search or refinement procedure. Typically, the grid created is a 2D or 3D Cartesian grid for SSL, or a polar or spherical grid for DOA estimation. In the latter case, the room dimensions $\mathbf{d}$ are not used, as the grid is produced with respect to the microphone array's centre. This grid can typically be computed as a pre-processing step if the microphone and room geometries are known beforehand.

The **compute_signal_features** module computes $\mathbf{C}$, which can be the CC function between microphone pairs, their GCC-PHAT, or neural-based features [138], [278].

Line (5) begins a loop, which represents the grid search procedure. For most approaches, this loop will only execute once, and will only execute multiple times in approaches such as [50], [219]. The loop's stopping criterion is the candidate grid $\mathbf{G}$ being empty, symbolizing that end of the grid search.

The module **create_srp_map** computes the SRP map $\mathbf{S}$, which assigns a likelihood value to each grid point in $\mathbf{G}$, using the microphone positions $\mathbf{V}$ and the temporal features $\mathbf{C}$.

Then, the **grid_search** module searches for the grid points in $\mathbf{G}$ that maximize the SRP map $\mathbf{S}$ to estimate the source coordinates $\hat{\mathcal{U}}$, as well as a new grid of candidate

locations $\mathbf{G}$. When localizing a single source, grid_search returns $\hat{\mathscr{P}} = \{\arg\max_{\mathbf{G}} \mathbf{S}\}$, and an empty grid, i.e., $\mathbf{G} = \emptyset$.

The **update_signal_features** module is used to alter the signal features $\mathbf{C}$. This is mainly used in iterative and multi-source approaches such as the source de-emphasis procedure [99] (cf. Fig. 2.11). Finally, the **update_grid** module may be used to generate a new grid based on the current source estimates $\hat{\mathcal{U}}$. An example of variant using this grid is the SRC approach [50].

6.5 Conclusion

This chapter showed that the SRP method remains an important localization method, and is still under continuous improvement. The detailed description of the conventional SRP method, followed by a presentation of the combination of the literature into multiple categories allows for analyses and extensions on SRP to be further conducted.

Future research directions on SRP include further improvement of neural methods, by allowing an arbitrary number of sources to be concurrently detected, inclusion of prior information such as noise statistics as a secondary network input, or architectural modifications, for example. Signal processing-based SRP modifications can also be improved by exploring other types of multi-source and tracking strategies, as well as alternative strategies for combining pairwise and frequency-wise information.

Chapter 7

On the use of complex-valued neural networks for SSL

This final chapter presents exploratory results in applying Complex-Valued Neural Networks (CVNNs) for the task of SSL. This type of network was explored in the beggining of the author's PhD. Although promising results were obtained, they were not further explored due to their lack of appropriate software tooling in comparison to real-valued networks.

7.1 Introduction

In recent years, deep learning techniques have been extensively explored for the task of Sound Source Localization (SSL) using microphone arrays. Out of the multiple proposed neural network architectures, the Convolutional Recurrent Neural Network (CRNN), which combines advantages of CNNs and RNNs, has achieved state of the art results on many scenarios. Multiple input features have been explored for this network, many of them based on the STFT. Although the STFT is a complex-valued feature, most approaches represent it as a real quantity either by discarding its magnitude [81], phase [82], or by treating the magnitude and phase independently [83]. Few approaches have explored the usage of the STFT in its original complex-valued form.

An explanation to why using the complex-valued STFT has been historically avoided

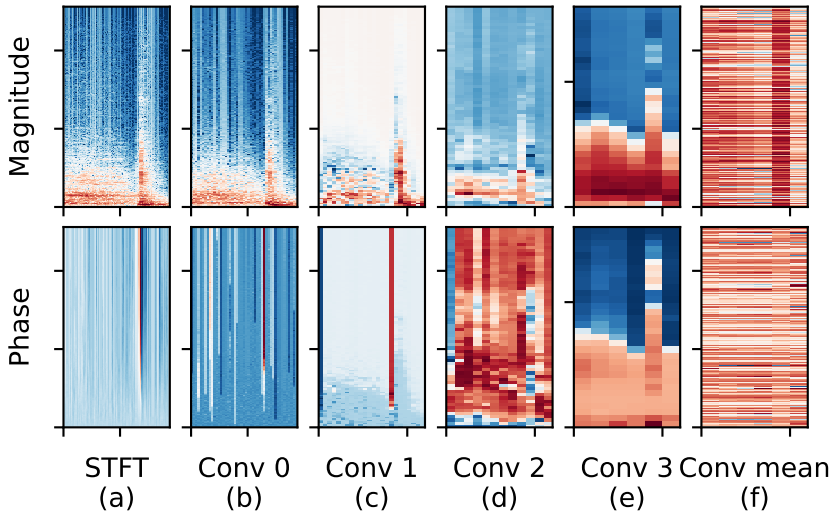

Figure 7.1: Example of features generated at the output of the neural network's convolutional layers.

is that it requires CVNNs to be used, which have been less explored than real-valued networks. However, this combination of input features and network presents multiple advantages. Firstly, unlike the aforementioned features, the complex-valued STFT does not lead to loss of information. Secondly, since the network operates directly on the STFT, intuitive visualizations on the intermediate representations learned of the network can be obtained. This allows the model to be more easily explainable and debuggable. Finally, CVNNs have been shown to be more stable than real-valued networks, being less prone to overfitting and to train faster [297].

In this chapter, using the complex-valued STFT alongside a complex-valued CRNN for estimating the azimuthal DOA of a single source is proposed. The proposed method is shown to compare favourably to a real-valued network with a similar number of parameters. For the experiments, networks are trained and evaluated using data from the DCASE 2019 dataset [298], which is composed of signals generated using real-life impulse responses on five different rooms.

Most approaches mentioned in Sec. 2.3 use real-valued neural networks. To the author's knowledge, the only dedicated study of CVNNs for SSL was conducted in [299], where a complex-valued MLP was employed for SSL using two directive microphones. In [83], many neural architectures for SSL are compared, including a CVNN. Outside the

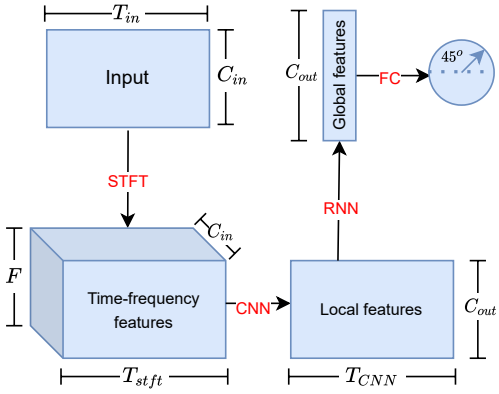

Figure 7.2: Architecture of the proposed CRNN.

SSL domain, CVNNs have recently attained state of the art performance on the task of speech enhancement [300], as well as music transcription and speech spectrum prediction [301]. The work of [301] also introduced complex-valued adaptations for important deep learning training techniques such as batch normalization [175]. The reader is referred to [302] and [297] for dedicated books on CVNNs.

This chapter continues as follows. The proposed neural network model is described in Sec. 7.2. Sec. 7.3 describes the experimentation procedure. Sec. 7.4 presents the results, and the last section provides a conclusion.

7.2 Complex-valued CRNN

This section describes the adopted architecture, as well as motivating the complex-valued operations performed by the network. A CRNN is divided into three sequential sub-networks: a CNN block, responsible for extracting local patterns from the input data, a RNN, responsible for combining these pattens into global, time-independent features, and a fully connected layer, which maps the global features into a single complex number representing the estimated source direction. A diagram representing the components of the network can be viewed in Fig. 7.2. The functioning of each block is further detailed below.

The convolutional block receives a tensor of shape (C_{in}, T_{stft}, F) representing a multi-

channel complex STFT. The first dimension C_{in} represents the number of audio channels, T_{stft} represents the number of time frames generated by the STFT, and F represents the number of frequency bins used. The role of this block is two-fold: firstly, to combine local information across all microphone channels, and secondly to reduce the dimensionality of the data to make it more tractable for the RNN layer.

The convolutional block consists of four sequential layers, where each performs three sequential operations. The output channel k at layer l is described as:

$$\text{output}(k,l) = \mathscr{P}(\sigma(\sum_{c=1}^{C_{in}^{l}} \text{f}(k,c) \star \text{input}(c,l))) \tag{7.1}$$

where $\text{f}(k,c) \star \text{input}(c,l)$ represents the layer l operation of cross-correlation between kernel f and the input input. This operation is carried out independently across each input channel c, after which the results are summed. This operation may be viewed as a sliding dot product across the time-frequency dimensions of the input with a kernel with the same number of channels. Different kernel shapes are shown in Fig. 7.3. After this operation, an activation function σ is applied, followed by a pooling function $\mathscr{P}$. As in [300], PReLU activation is used for all of the network's layers. An advantage of the PReLU over the more common ReLU is that it allows phase information to propagate across all complex quadrants instead of only the positive one.

Depending on the kernel shape used, the cross-correlation operation defined in (7.1) may have different meanings. Specifically, the operation performed by the kernel with shape $(1, 1, C)$, represented by the yellow block in Fig. 7.3 is a linear combination of the input channels. This operation may be interpreted as a narrowband weighted delay-and-sum beamformer [69], as each channel is scaled by its corresponding weight's magnitude and delayed by their phase. This is interpretation of complex-convolutional layers is relevant for the task of SSL, as classical approaches such as the SRP method [153] also perform beamforming as part of their location estimation procedure. However, for the experiments, $(2, 2, C)$ kernels are used, which are able to combine information from neighbouring frequencies and time frames.

After passing the input through the four convolutional layers, global average pooling operation is performed across all frequencies, generating a two-dimensional output matrix.

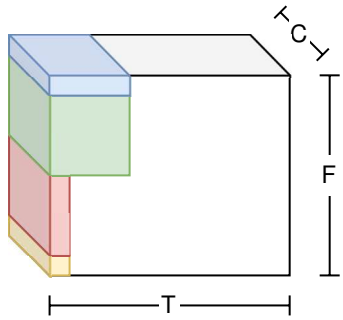

Figure 7.3: Representation of different types of convolutional kernels used on a multichannel spectrogram containing T time frames, F frequency bins and C channels.

After the convolutional block, the resulting matrix is fed to a bidirectional, gated recurrent unit neural network (GRU-RNN) [170]. As sound may not be present thoughout the whole duration of the audio signal, the RNN is important for propagating location information to silent time-steps. After this network, the rank of the features is once again reduced by performing average pooling on the time dimension, resulting in a vector of time-independent features.

The last block of the network is a fully connected layer which maps the global features to a single complex number, which is interpreted as a vector in the two-dimensional plane. The direction of this vector represents the azimuthal direction of the active source. Such representation is preferred to representing the azimuth as a scalar within the $[0, 2\pi]$ interval, as the circular nature of problem is lost and an ambiguity is created since the limits of the interval are the same [88].

The loss function is defined as the error between the network output $\hat{\boldsymbol{v}}$ and the target $\boldsymbol{v}$ a function of the cosine similarity as

$$\mathcal{L}(\boldsymbol{v}, \hat{\boldsymbol{v}}) = 1 - \frac{\boldsymbol{v} \cdot \hat{\boldsymbol{v}}}{|\boldsymbol{v}||\hat{\boldsymbol{v}}|}. \tag{7.2}$$

$\mathcal{L}$ ranges over $[0, 2]$, where 2 represents the maximum angular distance between the prediction and target of π radians.

Similarly to the case of real-valued networks, CVNNs require the computation of derivatives for their learning procedure. In this work, these derivatives are computed with respect to their conjugate values, according to the theory of CR calculus [303]. CR

calculus provides advantages over classic complex calculus by allowing non-holomorphic functions to be differentiable. This is important as the function defined by (7.2) is real, and therefore non-holomorphic. The weight update equation using gradient descent is:

$$\boldsymbol{W}_{n+1}^{l} = \boldsymbol{W}_{n}^{l} - \lambda \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{n}^{l*}}, \tag{7.3}$$

where $\boldsymbol{W}_{n}^{l}$ corresponds to layer l's weight matrix at time-step n, λ is the learning rate used and $*$ is the complex conjugate operator.

7.3 Experimentation

7.3.1 Dataset

To evaluate the proposed network, samples from the DCASE 2019 dataset [298] are used. The dataset was originally developed for the development of Sound Event Localization and Detection (SELD) algorithms using a tetrahedral microphone. In this work, the event types information is ignored The i-th element of the dataset $s[i]$ is generated by randomly selecting a room impulse response $\mathrm{rir}(k)$, convolving it with a randomly selected event signal $\mathrm{event}(l)$ and adding randomly selected ambient noise $\epsilon(j)$ afterwards. This procedure is summarized as

$$\mathrm{d}(i) = \mathrm{mask}(s(i), \tau, d) + \mathrm{noise}(j) \tag{7.4}$$

where

$$s(i) = \mathrm{rir}(k) * \mathrm{event}(l) + \epsilon(j). \tag{7.5}$$

There are 324 recorded angles beteween the source and microphone array for each of the five rooms available, as well as 11 event classes such as coughing, keyboard tapping and phone ringing, each of which contain 20 examples. Note that some of the beginning and end samples of each event signal may be silent, which makes the localization task more realistic. The ambient noise is scaled to produce a signal-to-noise ratio of 30 dB. The resulting dataset samples have each a duration of one second at a sampling rate of 24 kHz. The dataset consists of 31924 samples, out of which 50% are used for training,

25% for validation and 25% for testing.

7.3.2 Neural network training and evaluation

In this subsection, the neural network architecture described in the previous section is compared with an equivalent real-valued CRNN. To produce a fair comparison in terms of trainable parameters, the output size of all convolutional, recurrent and fully connected layers of the real-valued network is set to twice the size of the complex one.

For the STFT, the DFT of 1024 samples and hop length of 512 samples is used. During training, the Adam optimizer [176] was used with a fixed learning rate of 10^{-4}, along with a batch size of 32 signals for a duration of 20 epochs. The network is evaluated against the validation set at the end of each epoch. The final model used for testing is the one which obtains the lowest validation error.

The code was developed under the Pytorch [172] ecosystem, using Pytorch Lightning [173] for training. The code is made available on GitHub [1], and experiments may be reproduced on a provided Kaggle notebook [2]

7.4 Results and discussion

Fig. 7.4 shows a comparison between the training errors for the real and complex-valued networks during training and validation, as well as an error histogram for the test dataset. Results show that during training, the complex-valued network is able to converge more quickly than its real counterpart. The difference is particularly pronounced by the end of the first epoch, where the error of the real network considerably higher than the complex one. Turning to the validation dataset, it is shown that while the real network is overfitting the data during the first epochs, the complex network is able to generalize from the first epochs. Finally, the error histograms for the test set show that the end result for both networks is very similar. While the real network produced an average test error and standard deviation of respectively 9.3° and 18.1°, the complex version had a slightly better performance of 9.0° and 17.5°.

[1] https://github.com/SOUNDS-RESEARCH/complex_neural_source_localization
[2] https://www.kaggle.com/code/egrinstein/neural-doa-training-notebook/notebook

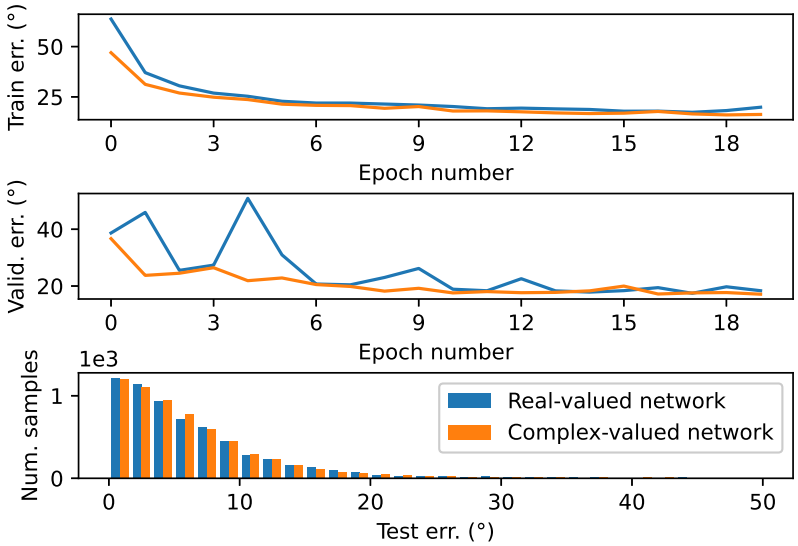

Figure 7.4: Top and middle: training and validation errors at the end of every epoch for the proposed complex architecture as well as for the real baseline. Bottom: error histogram for the test set.

To further analyze the behaviour of the proposed network, visualization of the outputs of its convolutional layers are produced. Fig. 7.1 shows the output layers for an example audio in the test set. (a) shows the magnitude and phase plots of the STFT for the first microphone channel. (b-e) show the first output channel of the convolutional layers of the network. On these columns, the horizontal axis represents time while the vertical axis represents frequency. Finally, (f) shows the output feature generated after averaging the frequency channels on the last convolutional layer. In this plot, the horizontal axis represents time while the vertical bins represent the channel number. It is shown that the original magnitude structural information is preserved across all layers. In the last layer, the magnitude is considerably stronger at the time frames which seem to be of most interest within the signal. It is therefore reasonable to consider that the network is able to focus on regions of interest within the signal to perform localization.

7.5 Conclusion

This chapter presented an experiment of using complex-valued neural networks for estimating the DOA of a single source on a plane. Results indicate that complex-valued CRNNs are able to converge to lower errors faster than their real counterparts, as well as being less prone to overfitting. This latter characteristic indicates they could be particularly useful when little training data is available. A second advantage of these networks is being able to visualize their learned features, which provides intuitive explanations to their functioning. A future research direction is to extend this work for localization and detection of multiple sources.

Chapter 8

Conclusion

SSL is a fundamental task in acoustic signal processing, with applications in speech enhancement and acoustic scene understanding. There are multiple challenges to this task, including detecting multiple sources, moving sources, real-time causal constraints, resource constraints, reverberation, directional and diffuse noise, types of microphones, among others.

Neural network-based methods have received increased attention in the past years due to their ability to overcome most of the aforementioned problems for specific scenarios. However, neural network models are known to offer poor generalization to unseen scenarios, limiting their application in practice. Conversely, this issue is not shared with classical signal processing algorithms, which, however, come with their own performance limitations.

This thesis focused on improving the generalization of neural methods for SSL. Specifically, it focused on the development of neural methods capable of working on multiple microphone array geometries without the need of retraining. This is advantageous for companies with many arrays on their product lines, or for companies or individuals without the resources to train and maintain their own SSL modules.

The challenge of developing universal methods for SSL is divided into two subproblems: different number of microphones and different relative microphone positions. The problem of different positions is discussed on chapter 3, where strategies for incorporating microphone positions into networks is discussed, culminating in the Dual-input Neural

Network architecture, which is capable of conditioning its output on the microphone positions. The problem of different microphone numbers is studied in chapter 4, where a GNN is proposed to estimate the position of a source for arrays of any size. Finally, chapter 5 builds on the previous chapters to develop Neural-SRP, a robust algorithm to track multiple sound sources in challenging environments.

A second contribution of this thesis is the review and generalization of over 200 variations of the SRP method, a classical signal processing technique from which the proposed Neural-SRP takes inspiration from. More than 200 papers were reviewed, showing that the original SRP formulation proposed more than 30 years ago can be modified in multiple ways to improve its robustness to multiple sources, noise types, reverberation, among other conditions. The proposed generalization of the methods allows for faster and more structured innovation in this ever-growing field.

Several research directions are interesting in this active field. One is reducing model sizes and complexities, so they can be embedded into smart devices such as watches and glasses. This can be achieves using state-of-the-art techniques such as model pruning, transfer learning, and Neural Architecture Search (NAS). A second direction consists of combining SSL models to other tasks, such as event detection, source separation and beamforming. Increasing controllability of SSL models can also be useful: for instance, a model for selective localization of a given input sound type could be devised. Finally, multi-source tracking models are still typically limited to the maximum number of detectable sources, a constraint which can be removed by iteratively decoding sources locations and their probabilities instead of using a fixed output.

Bibliography

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[3] C. Evers, H. W. Löllmann, H. Mellmann, *et al.*, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, 2020.

[4] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 1997, pp. 187–190.

[5] P. Cabañas-Molero, M. Lucena, J. M. Fuertes, P. Vera-Candeas, and N. Ruiz-Reyes, "Multimodal speaker diarization for meetings using volume-evaluated SRP-PHAT and video analysis," *Multimedia Tools and Applications*, vol. 77, no. 20, pp. 27 685–27 707, 2018.

[6] K. Wiggers, *Amazon's AI uses a microphone array to localize multiple speakers in a room*, https://venturebeat.com/ai/amazons-ai-uses-a-microphone-array-to-localize-multiple-speakers-in-a-room/, Apr. 2020.

[7] E. Grinstein, C. M. Hicks, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP Method for universal robust multi-source tracking," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 19–28, 2024.

[8] E. Grinstein, V. W. Neo, and P. A. Naylor, "Dual input neural networks for positional sound source localization," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2023, no. 1, p. 32, Aug. 2023.

[9] E. Grinstein, M. Brookes, and P. A. Naylor, "Graph neural networks for sound source localization on distributed microphone networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023.

[10] E. Grinstein, T. van Waterschoot, M. Brookes, and P. A. Naylor, "The Neural-SRP method for positional sound source localization," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2023.

[11] E. Grinstein and P. A. Naylor, "Deep complex-valued convolutional-recurrent networks for single source DOA estimation," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sep. 2022, pp. 1–5.

[12] E. Grinstein, E. Tengan, B. Çakmak, *et al.*, "Steered response power for sound source localization: A tutorial review," Imperial College London, Tech. Rep. arXiv:2405.02991, 2024.

[13] A. Chinaev, P. Thüne, and G. Enzner, "A double-cross-correlation processor for blind sampling rate offset estimation in acoustic sensor networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 641–645.

[14] D. Diaz-Guerra and J. R. Beltran, "Direction of arrival estimation with microphone arrays using SRP-PHAT and neural networks," in *Proc. IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, 2018, pp. 617–621.

[15] K. E. Iverson, *A programming language*. Wiley, 1962.

[16] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[17] S. Adavanne, A. Politis, and T. Virtanen, "Differentiable tracking-based training of deep learning sound source localizers," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2021, pp. 211–215.

[18] C. Foy, A. Deleforge, and D. Di Carlo, "Mean absorption estimation from room impulse responses using virtually supervised learning," *J. Acoust. Soc. Am.*, vol. 150, no. 2, pp. 1286–1299, Aug. 2021.

[19] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. (AES) Conv.*, Audio Engineering Society, 2000.

[20] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[21] "Simulation of sound in rooms," in *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 175–226.

[22] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *J. Audio Eng. Soc. (AES)*, vol. 62, no. 11, pp. 748–766, 2014.

[23] M. R. Bai and P. Ho, "Using a steered-response power-phase transform to optimize speech pickup in reverberant environments," *J. Audio Eng. Soc. (AES)*, vol. 56, no. 4, pp. 280–291, 2008.

[24] A. Levi and H. F. Silverman, "An alternate approach to adaptive beamforming using SRP-PHAT," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2010, pp. 2726–2729.

[25] J. Traa, D. Wingate, N. D. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 493–503, 2016.

[26] D. A. Blauth, V. P. Minotto, C. R. Jung, B. Lee, and T. Kalker, "Voice activity detection and speaker localization using audiovisual cues," *Pattern Recognition Lett.*, Intelligent Multimedia Interactivity, vol. 33, no. 4, pp. 373–380, 2012.

[27] O. Schwartz, A. David, O. Shahen-Tov, and S. Gannot, "Multi-microphone voice activity and single-talk detectors based on steered-response power output entropy," in *Proc. IEEE Int. Conf. on the Science of Elec. Eng. in Israel (ICSEE)*, 2018, pp. 1–4.

[28] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for joint source localization and separation," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2015, pp. 1–5.

[29] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 281–295, 2018.

[30] W. Kang, B. C. Roy, and W. Chow, "Multimodal speaker diarization of real-world meetings using d-vectors with spatial features," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 6509–6513.

[31] T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, and R. Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," Paderborn University, Tech. Rep. arXiv:2205.00944, 2022.

[32] H. Q. H. Dam, H. Ho, and M. H. L. Ngo, "Blind speech separation using SRP-PHAT localization and optimal beamformer in two-speaker environments," *Int. J. of Computer and Information Eng.*, vol. 10, no. 8, pp. 1529–1533, 2016.

[33] H. Q. H. Dam and S. Nordholm, "Source separation employing beamforming and SRP-PHAT localization in three-speaker room environments," *Vietnam J. of Computer Science*, vol. 4, no. 3, pp. 161–170, 2017.

[34] C. Wu, L. Zhou, X. Chen, and L. Chen, "Microphone array speech separation algorithm based on DNN," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2021, pp. 1305–1310.

[35] M. Hennecke, T. Plotz, G. A. Fink, J. Schmalenstroer, and R. Haeb-Umbach, "A hierarchical approach to unsupervised shape calibration of microphone array networks," in *Proc. IEEE/SP Workshop on Statistical Signal Process.*, 2009, pp. 257–260.

[36] A. Sedunov, H. Salloum, A. Sutin, N. Sedunov, and S. Tsyuryupa, "UAV passive acoustic detection," in *Proc. IEEE Int. Symp. on Technologies for Homeland Security (HST)*, 2018, pp. 1–6.

[37] B. Harvey and S. O'Young, "A harmonic spectral beamformer for the enhanced localization of propeller-driven aircraft," *J. of Unmanned Vehicle Systems*, vol. 7, no. 2, pp. 156–174, 2019.

[38] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: dataset and methods for UAV-embedded sound source localization," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS)*, 2018, p. 5735.

[39] E. Tengan, "Spatial audio analysis with constrained microphone setups in adverse acoustic conditions," in KU Leuven: PhD thesis, 2024, ch. DOA-informed speech enhancement with a UAV, pp. 111–123.

[40] C. Zieger, A. Brutti, and P. Svaizer, "Acoustic based surveillance system for intrusion detection," in *Proc. IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, 2009, pp. 314–319.

[41] K. Kim, S. Wang, H. Ryu, and S. Q. Lee, "Acoustic-based position estimation of an object and a person using active localization and sound field analysis," *Applied Sciences*, vol. 10, no. 24, p. 9090, 2020.

[42] J. Lopez-Morillas, F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, R. Mata-Campos, and V. Montiel-Zafra, "Gunshot detection and localization based on non-negative matrix factorization and SRP-PHAT," in *Proc. IEEE Sensor Array and Multichannel Signal Process. Workshop (SAM)*, 2016, pp. 1–5.

[43] J. H. Park, W. Cho, and S.-C. Kim, "Improving acoustic localization accuracy by applying interaural level difference and support vector machine for AoA outlier removal," in *Proc. Int. Conf. on Elec., Inf., and Commun. (ICEIC)*, 2021, pp. 1–4.

[44] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization: Reviews, methodology and applications," *Mech. Syst. and Signal Process.*, vol. 120, pp. 422–448, Apr. 2019.

[45] M. Royvaran, K. D. Donohue, and B. Davis, "Localization of stationary source of floor vibration using steered response power method," in *Dynamics of Civil Structures, Volume 2*, 2021, pp. 141–149.

[46] J. Tiete, F. Domínguez, B. D. Silva, L. Segers, K. Steenhaut, and A. Touhafi, "SoundCompass: A distributed mems microphone array-based sensor for sound source localization," *Sensors*, vol. 14, no. 2, pp. 1918–1949, Feb. 2014.

[47] P. Nie, B. Liu, P. Chen, K. Li, and Y. Han, "SRP-PHAR combined velocity scanning for locating the shallow underground acoustic source," *IEEE Access*, vol. 7, pp. 161 350–161 362, 2019.

[48] P. DeVille, "Localization of soniferous fish using a sparse hydrophone array and conventional steered response power method," M.S. thesis, East Carolina University, 2019.

[49] J. Chen, X. Shen, M. Lu, J. Wu, N. Zhou, and L. Luo, "Equipment fault acoustic source direction of arrival estimation with microphone arrays using SRP-PHAT method," in *Proc. Asia Conf. on Power and Elec. Eng. (ACPEE)*, 2020, pp. 1388–1392.

[50] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, 2007, pp. I–121–I–124.

[51] S. Shon, E. Kim, J. Yoon, and H. Ko, "Sudden noise source localization system for intelligent automobile application with acoustic sensors," in *Proc. IEEE Int. Conf. on Consumer Elec. (ICCE)*, 2012, pp. 233–234.

[52] A Swerdlow, T Machmer, and K Kroschel, "Speaker position estimation in vehicles by means of acoustic analysis," in *Proc. Int. Conf. on Acoust. (DAGA)*, 2008.

[53] B. Van Den Broeck, L. Vuegen, H. Van hamme, M. Moonen, P. Karsmakers, and B. Vanrumste, "Footstep localization based on in-home microphone-array signals," in *Assistive Technology: From Research to Practice*, P. Encarnação *et al.*, Eds., IOS Press, 2013, pp. 90–94.

[54] Y. Li, K. C. Ho, and M. Popescu, "A microphone array system for automatic fall detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1291–1301, 2012.

[55] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, Robust Speech Proc. Vol. 5, no. 2, pp. 131–140, 2004.

[56] R. Lebarbenchon, E. Camberlein, D. di Carlo, C. Gaultier, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge," *arXiv*, Dec. 2018.

[57] J. J. Gamboa-Montero, M. Basiri, J. C. Castillo, S. Marques-Villarroya, and M. A. Salichs, "Real-time acoustic touch localization in human-robot interaction based on steered response power," in *Proc. IEEE Int. Conf. on Development and Learning (ICDL)*, 2022, pp. 101–106.

[58] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2011, pp. 2592–2595.

[59] A. Johansson, N. Grbic, and S. Nordholm, "Speaker localisation using the far-field SRP-PHAT in conference telephony," in *Proc. 2002 Int. Symp. on Intelligent Signal Process. and Commun. Syst. (ISPACS)*, 2002.

[60] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2011, pp. 1317–1321.

[61] Y. Zhang and S. Meng, "Sound source localization algorithm based on a helmet-mounted microphone array," in *Intern. Symp. on Parallel Architectures, Algorithms and Programming*, Jul. 2014, pp. 183–186.

[62] F. Gustafsson and F. Gunnarsson, "Positioning using time-difference of arrival measurements," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2003.

[63] H. C. So, "Source localization: Algorithms and analysis," in *Handbook of Position Location: Theory, Practice, and Advances*, Z. Seyed A. and R. M. Buehrer, Eds., John Wiley & Sons, Ltd, 2011, pp. 25–66.

[64] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, pp. 320–327, Aug. 1976.

[65] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, 1997.

[66] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Language Process.*, pp. 2510–2526, Nov. 2007.

[67] S. Birchfield and D. Gillmor, "Acoustic source direction by hemisphere sampling," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 5, 2001, 3053–3056 vol.5.

[68] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays.* USA, 2000.

[69] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds., Berlin, Heidelberg: Springer, 2001, pp. 157–180.

[70] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on nyquist-shannon sampling," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2021, pp. 206–210.

[71] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, 2009.

[72] S. Tervo and T. Lokki, "Interpolation methods for the SRP-PHAT algorithm," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2008.

[73] L. O. Nunes, W. A. Martins, M. V. S. Lima, *et al.*, "Discriminability measure for microphone array source localization," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[74] W. Cochran, J. Cooley, D. Favin, *et al.*, "What is the Fast Fourier Transform?" *Proc. of the IEEE*, vol. 55, no. 10, pp. 1664–1674, Oct. 1967.

[75] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Audio, Speech, Language Process.*, pp. 943–956, Nov. 2001.

[76] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2008, pp. 2565–2568.

[77] X. Sheng and Y.-h. Hu, "Energy based acoustic source localization," in *Information Processing in Sensor Networks*, 2003, pp. 285–300.

[78] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[79] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, 2021.

[80] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signal to source position coordinates," *Sensors*, vol. 18, no. 10, Oct. 2018.

[81] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2017, pp. 136–140.

[82] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017, pp. 6125–6129.

[83] D. Krause, A. Politis, and K. Kowalczyk, "Comparison of convolution types in CNN-based feature extraction for sound source localization," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 820–824.

[84] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. Int. Conf. Robotics and Automation*, May 2018, pp. 74–79.

[85] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *J. of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[86] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1462–1466.

[87] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "CRNN-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 22–33, Mar. 2019.

[88] L. Perotin, A. Défossez, E. Vincent, R. Serizel, and A. Guérin, "Regression versus classification for neural network based audio source localization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2019, pp. 343–347.

[89] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 300–311, 2021.

[90] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Direction of arrival estimation of sound sources using icosahedral CNNs," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 313–321, 2023.

[91] D. Diaz-Guerra, A. Politis, and T. Virtanen, "Position tracking of a varying number of sound sources with sliding permutation invariant training," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2023, pp. 251–255.

[92] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[93] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, pp. 276–280, Mar. 1986.

[94] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[95] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS)*, Oct. 2007, pp. 2009–2014.

[96] F. Grondin and J. Glass, "Fast and robust 3-D sound source localization with DSVD-PHAT," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS)*, Nov. 2019, pp. 5352–5357.

[97] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Process.*, vol. 87, no. 7, pp. 1677–1691, 2007.

[98] T. Padois, O. Doutres, F. Sgard, and A. Berry, "On the use of geometric and harmonic means with the generalized cross-correlation in the time domain to improve noise source maps," *J. Acoust. Soc. Am.*, vol. 140, no. 1, EL56–EL61, 2016.

[99] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2010, no. 1, pp. 1–17, 2010.

[100] D. Diaz-Guerra and J. R. Beltran, "Source cancellation in cross-correlation functions for broadband multisource DOA estimation," *Signal Process.*, vol. 170, p. 107 442, 2020.

[101] F. Grondin and J. Glass, "Multiple sound source localization with SVD-PHAT," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 2698–2702.

[102] F. Grondin and J. Glass, "SVD-PHAT: A fast sound source localization method," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 4140–4144.

[103] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. Joint Workshop on Hands-free Speech Commun. and Microphone Arrays (HSCMA)*, 2011, pp. 92–97.

[104] Y. Oualil, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012.

[105] Y. Oualil, F. Faubel, and D. Klakow, "A fast cumulative steered response power for multiple speaker detection and localization," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2013, pp. 1–5.

[106] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.

[107] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Multispeaker localization and tracking in intelligent environments," in *Multimodal Technologies for Perception of Humans*, 2008, pp. 82–90.

[108] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2008.

[109] H. Do and H. F. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 301–304.

[110] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2010, pp. 125–128.

[111] W. Cai, X. Zhao, and Z. Wu, "Localization of multiple speech sources based on sub-band steered response power," in *Proc. Int. Conf. on Elec. and Control Eng.*, 2010, pp. 1246–1249.

[112] S. Khanal and H. F. Silverman, "Multi-stage rejection sampling (MSRS): A robust SRP-PHAT peak detection algorithm for localization of cocktail-party talkers," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2015, pp. 1–5.

[113] R. Boora and S. K. Dhull, "A TDOA-based multiple source localization using delay density maps," *Sādhanā*, vol. 45, no. 1, p. 204, 2020.

[114] A. Dehghan Firoozabadi, P. Irarrázaval, P. Adasme, *et al.*, "Multi-speaker localization by central and lateral microphone arrays based on the combination of 2D-SRP and subband GEVD algorithms," in *Proc. Int. Conf. on Signal Processing and Communication (ICSC)*, 2022, pp. 433–438.

[115] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, no. 10, pp. 13 781–13 812, 10 2012.

[116] E. Tengan, T. Dietzen, F. Elvander, and Toon van Waterschoot, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered response power maps," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, USA, 2023, pp. 1–5.

[117] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, 2011.

[118] H. Park and J. Li, "A frequency-domain SPICE approach to high-resolution time delay estimation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 360–363, 2018.

[119] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[120] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, 2002, pp. I–529–I–532.

[121] E. Hadad and S. Gannot, "Multi-speaker direction of arrival estimation using SRP-PHAT algorithm with a weighted histogram," in *IEEE Int. Conf. on the Science of Elec. Eng. in Israel (ICSEE)*, 2018, pp. 1–5.

[122] T. Padois, O. Doutres, F. Sgard, and A. Berry, "Time domain localization technique with sparsity constraint for imaging acoustic sources," *Mechanical Systems and Signal Proc.*, vol. 94, pp. 85–93, 2017.

[123] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, Nov. 1993, pp. 40–44.

[124] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[125] Z. Chu, J. Weng, and Y. Yang, "Determination of propagation model matrix in generalized cross-correlation based inverse model for broadband acoustic source localization," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2098–2109, 2020.

[126] S. Thakallapalli, S. V. Gangashetty, and N. Madhu, "NMF-weighted SRP for multi-speaker direction of arrival estimation: Robustness to spatial aliasing while exploiting sparsity in the atom-time domain," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2021, no. 1, p. 13, 2021.

[127] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[128] A. Abad, C. Segura, D. Macho, J. Hernando, and C. Nadeu, "Audio person tracking in a smart-room environment," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2006.

[129] S. Astapov, J. Berdnikova, and J.-S. Preden, "Predictive acoustic localization and speaker tracking for distributed sensor networks," in *Proc. Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 833–838.

[130] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics and Autonomous Syst.*, vol. 113, pp. 63–80, 2019.

[131] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[132] J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, 2006, pp. IV–IV.

[133] T. Habib and H. Romsdorfer, "Comparison of SRP-PHAT and Multiband-PoPi algorithms for speaker localization using particle filters," in *Proc. Conf. on Digital Audio Effects*, 2010.

[134] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[135] K. Wu and A. W. H. Khong, "Acoustic source tracking in reverberant environment using regional steered response power measurement," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2013, pp. 1–6.

[136] K. Wu and A. W. H. Khong, "Sound source localization and tracking," in *Context Aware Human-Robot and Human-Agent Interaction*, N. Magnenat-Thalmann, J. Yuan, D. Thalmann, and B.-J. You, Eds., Springer, 2016, pp. 55–78.

[137] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Tracking a moving sound source from a multi-rotor drone," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS)*, 2018, pp. 2511–2516.

[138] B. Yang, H. Liu, and X. Li, "SRP-DNN: learning direct-path phase difference for multiple moving sound source localization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 721–725.

[139] T. Zhong, I. M. Velázquez, Y. Ren, H. M. P. Meana, and Y. Haneda, "Spherical convolutional recurrent neural network for real-time sound source tracking," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 5063–5067.

[140] M. F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1228–1242, Aug. 2010.

[141] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 3, pp. 1526–1540, 2004.

[142] F. Grondin, M.-A. Maheux, J.-S. Lauzon, J. Vincent, and F. Michaud, "SMP-PHAT: Lightweight DOA estimation by merging microphone pairs," Université de Sherbrooke, Tech. Rep. arXiv:2203.14409, 2022.

[143] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2005.

[144] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2006.

[145] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura, and H. Tsujino, "Sound source tracking with directivity pattern estimation using a 64 ch microphone array," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS)*, 2005, pp. 1690–1696.

[146] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2007.

[147] M. Togami and Y. Kawaguchi, in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2010, pp. 133–136.

[148] C. Segura and F. J. Hernando Pericás, "GCC-PHAT based head orientation estimation," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 1–4.

[149] H. Silverman, Y. Yu, J. Sachar, and W. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 593–606, 2005.

[150] U. T. K. Nguyen and T. V. Pham, "Performance assessment of generalized cross-correlation based algorithms for multisource point-based localization and detection," in *Int. Conf. on Advanced Technologies for Commun. (ATC)*, 2011, pp. 303–306.

[151] G. Lathoud, "Further applications of sector-based detection and short-term clustering," IDIAP, Martigny, Switzerland, Tech. Rep. Idiap-RR-26-2006, 2006.

[152] A. Johansson, G. Cook, and S. Nordholm, "Acoustic direction of arrival estimation, a comparison between Root-MUSIC and SRP-PHAT," in *IEEE Region 10 Conf. (TENCON)*, vol. B, 2004, pp. 629–632.

[153] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[154] J. P. Dmochowski and J. Benesty, "Steered beamforming approaches for acoustic source localization," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds., Springer, 2010, pp. 307–337.

[155] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 602–606.

[156] J. Peterson and C. Kyriakakis, "Analysis of fast localization algorithms for acoustical environments," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2005, pp. 1385–1389.

[157] D. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 499–508, 2004.

[158] J. Peterson and C. Kyriakakis, "Hybrid algorithm for robust, real-time source localization in reverberant environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, 2005, pp. iv/1053–iv/1056.

[159] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP J. on Advances in Signal Process.*, vol. 2003, no. 4, pp. 1–10, 2003.

[160] J. Velasco, C. J. Martín-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios," *Signal Process.*, vol. 119, pp. 209–228, 2016.

[161] M. Swartling and N. Grbić, "Calibration errors of uniform linear sensor arrays for DOA estimation: An analysis with SRP-PHAT," *Signal Process.*, vol. 91, no. 4, pp. 1071–1075, 2011.

[162] P. Nie, B. Liu, P. Chen, and Y. Han, "Coherence-weighted steered response power for acoustic source localization," *Acoustics Australia*, vol. 50, no. 3, pp. 365–371, 2022.

[163] D. B. Haddad, M. V. S. Lima, W. A. Martins, L. W. P. Biscainho, L. O. Nunes, and B. Lee, "Acoustic sensor self-localization: Models and recent results," *Wireless Communications and Mobile Computing*, vol. 2017, e7972146, Oct. 2017.

[164] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, "Parameterized neural networks for high-energy physics," *The European Physical Journal C*, May 2016.

[165] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, pp. 345–379, Nov. 2010.

[166] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 423–443, Feb. 2019.

[167] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[168] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 4642–4646.

[169] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," *Proc. Detect. and Classific. of Acoust. Scenes and Events (DCASE)*, pp. 30–34, 2019.

[170] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Neural Inform. Process. Conf*, vol. abs/1412.3555, 2014.

[171] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2018, pp. 351–355.

[172] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Neural Inform. Process. Conf*, vol. 32, Curran Associates, Inc., 2019.

[173] W. Falcon and The PyTorch Lightning team, *PyTorch lightning*, https://www.pytorchlightning.ai, Mar. 2019.

[174] O. Yadan, *Hydra - A framework for elegantly configuring complex applications*, https://www.hydra.cc, 2019.

[175] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learning (ICML)*, Jun. 2015, pp. 448–456.

[176] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Arxiv*, Jan. 2017. arXiv: 1412.6980 [cs].

[177] J. Yamagishi, C. Veaux, and K. MacDonald, *Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, Nov. 2019.

[178] S. Guan, S. Liu, J. Chen, W. Zhu, S. Li, *et al.*, "Libri-Adhoc40: A dataset collected from synchronized ad-hoc microphone arrays," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2021.

[179] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[180] A. M. Aurand, J. S. Dufour, and W. S. Marras, "Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume," *Journal of Biomechanics*, vol. 58, pp. 237–240, Jun. 2017.

[181] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2013, pp. 106–110.

[182] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Passive self-localization of microphones using ambient sounds," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 1314–1318.

[183] H. Gamper and I. J. Tashev, "Blind Reverberation Time Estimation Using a Convolutional Neural Network," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sep. 2018, pp. 136–140.

[184] P. S. López, P. Callens, and M. Cernak, "A universal deep room acoustics estimator," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2021, pp. 356–360. arXiv: 2109.14436 [cs, eess].

[185] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, "Turning enemies into friends: Using reflections to improve sound source localization," in *IEEE International Conference on Multimedia and Expo*, Jul. 2010, pp. 731–736.

[186] W. Xue, Y. Tong, C. Zhang, G. Ding, X. He, and B. Zhou, "Sound event localization and detection based on multiple doa beamforming and multi-task learning," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Oct. 2020, pp. 5091–5095.

[187] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 241–245.

[188] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symp. on Commun. and Veh. Technol. in the Benelux (SCVT)*, Nov. 2011, pp. 1–6.

[189] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Feb. 2021.

[190] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, 2017.

[191] J. Zhou, G. Cui, S. Hu, *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Jan. 2020.

[192] P. W. Battaglia, J. B. Hamrick, V. Bapst, *et al.*, "Relational inductive biases, deep learning, and graph networks," Google Inc., Mountain View, CA, USA, Tech. Rep., 2018.

[193] A. Santoro, D. Raposo, D. G. Barrett, *et al.*, "A simple neural network module for relational reasoning," in *Proc. Neural Inform. Process. Conf*, vol. 30, 2017.

[194] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2008, Jul. 2008.

[195] D. Li and Y. H. Hu, "Energy based collaborative source localization using acoustic micro-sensor array," *EURASIP J. on Applied Signal Process.*, 2003.

[196] N. Furnon, R. Serizel, S. Essid, and I. Illina, "Attention-based distributed speech enhancement for unconstrained microphone arrays with varying number of nodes," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021.

[197] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Apr. 2020.

[198] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020, pp. 6394–6398.

[199] Y.-K. Luo, S.-X. Chen, L. Zhou, and Y.-Q. Ni, "Evaluating railway noise sources using distributed microphone array and graph neural networks," *Transportation Research Part D: Transport and Environment*, vol. 107, Jun. 2022.

[200] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. on Learning Representations*, Feb. 2017.

[201] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, Jun. 1977.

[202] R. Neubauer and B. Kostek, "Prediction of the reverberation time in rectangular rooms with non-uniformly distributed sound absorption," *Archives of Acoustics*, vol. 26, no. 3, 2001.

[203] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)*, Nov. 2019.

[204] S. Li, X. Chang, C. Yang, *et al.*, "A fast vehicle horn sound location method with improved SRP-PHAT," in *IEEE Int. Conf. on Progress in Informatics and Comp. (PIC)*, 2018, pp. 435–439.

[205] H. Do and H. F. Silverman, "A robust sound-source separation algorithm for an adverse environment that combines MVDR-PHAT with the CASA framework," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2011, pp. 273–276.

[206] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. ii, 1994, II/273–II/276 vol.2.

[207] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2010, pp. 85–88.

[208] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 241–245.

[209] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[210] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, Cham: Springer International Publishing, 2018, pp. 270–279.

[211] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[212] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events Workshop*, 2020.

[213] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS general sound events database," *arXiv*, no. 1902.08314, 2020. eprint: 1902.08314 (cs, eess).

[214] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, 2011.

[215] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011, pp. 1–6.

[216] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[217] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

[218] G. García-Barrios, J. M. Gutiérrez-Arriola, N. Sáenz-Lechón, V. J. Osma-Ruiz, and R. Fraile, "Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps," *IEEE Access*, vol. 9, pp. 121 549–121 560, 2021.

[219] L. O. Nunes, W. A. Martins, M. V. S. Lima, *et al.*, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.

[220] M. V. S. Lima, W. A. Martins, L. O. Nunes, *et al.*, "A volumetric SRP with refinement step for sound source localization," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1098–1102, 2015.

[221] M. V. S. Lima, W. A. Martins, L. O. Nunes, *et al.*, "Efficient steered-response power methods for sound source localization using microphone arrays," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1098–1102, 2015.

[222] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 2627–2630, 2013.

[223] D. Salvati, C. Drioli, and G. L. Foresti, "Acoustic source localization using a geometrically sampled grid SRP-PHAT algorithm with max-pooling operation," *IEEE Signal Process. Lett.*, vol. 29, pp. 1828–1832, 2022.

[224] R. Boora and S. K. Dhull, "Iterative volumetric reduction (IVR) steered response power method for acoustic source localization," *Int. J. of Sensors Wireless Commun. and Control*, vol. 11, no. 4, pp. 428–436, 2021.

[225] R. Boora and S. K. Dhull, "Iterative modified SRP-PHAT with adaptive search space for acoustic source localization," *IETE Tech. Rev.*, vol. 39, no. 1, pp. 28–36, 2022.

[226] R. Duraiswami, D. Zotkin, and L. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 5, 2001, pp. 3309–3312.

[227] H. Do and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2007, pp. 295–298.

[228] Y. Guo, J. Wu, and S. Zhu, "SRP-PHAT source location algorithm based on chaos artificial bee colony algorithm," in *Proc. Int. Conf. on Information Eng. for Mechanics and Materials*, 2015, pp. 153–158.

[229] R. Scheibler and M. Togami, "Refinement of direction of arrival estimators by majorization-minimization optimization on the array manifold," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2021, pp. 436–440.

[230] M. Togami and R. Scheibler, "Sound source localization with majorization minimization," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 2122–2126.

[231] D. Liu, X. Cai, D. Yu, Z. Qiao, H. Dong, and M. Wu, "Sound source localization methods based on Lagrange-Galerkin spherical grid," in *Proc. IEEE Int. Conf. on Electrical Eng. and Mechatronics Technol. (ICEEMT)*, Jul. 2021, pp. 665–670.

[232] S. Astapov, J. Berdnikova, and J.-S. Preden, "Optimized acoustic localization with SRP-PHAT for monitoring in distributed sensor networks," *Int. J. of Elec. and Telecommun.*, vol. 59, no. 4, 2013.

[233] S. Astapov, J. Berdnikova, and J.-S. Preden, "A method of initial search region reduction for acoustic localization in distributed systems," in *Proc. Int. Conf. Mixed Design of Integrated Circuits and Systems (MIXDES)*, 2013, pp. 451–456.

[234] Y. Zhao, X. Chen, and B. Wang, "Real-time sound source localization using hybrid framework," *Applied Acoust.*, vol. 74, no. 12, pp. 1367–1373, 2013.

[235] M. Seifipour and S. Seyedtabaii, "Computation saving in a SRP-PHAT sound source locator variant," in *Proc. Iranian Conf. on Elec. Eng. (ICEE)*, 2013, pp. 1–5.

[236] M. Ranjkesh Eskolaki and R. Hasanzadeh, "A fast and accurate sound source localization method using optimal combination of SRP and TDOA methodologies," *Journal of Information Systems and Telecommunication (JIST)*, vol. 2, no. 10, p. 1, Jun. 2015.

[237] M. A. Awad-Alla, A. Hamdy, F. A. Tolbah, M. A. Shahin, and M. A. Abdelaziz, "A two-stage approach for passive sound source localization based on the SRP-PHAT algorithm," *APSIPA Trans. on Signal and Information Processing*, vol. 9, no. e8, 2020.

[238] Y. Cho, D. Yook, S. Chang, and H. Kim, "Sound source localization for robot auditory systems," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp. 1663–1668, 2009.

[239] X. Yuan, D. Cai, J. Deng, P. Li, and P. Gong, "Performance enhancement of SSC sound source localization for indoor environment," in *Proc. Int. Conf. on Signal Process.*, vol. 1, 2012, pp. 79–83.

[240] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Trans. on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.

[241] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the SRP-PHAT for localization improvement and power response sensitivity analysis," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 586–601, 2017.

[242] D. Salvati, C. Drioli, and G. L. Foresti, "Sensitivity-based region selection in the steered response power algorithm," *Signal Process.*, vol. 153, pp. 1–10, 2018.

[243] W. Cai, S. Wang, and Z. Wu, "Accelerated steered response power method for sound source localization using orthogonal linear array," *Applied Acoust.*, vol. 71, no. 2, pp. 134–139, 2010.

[244] A. Dehghan Firoozabadi and H. R. Abutalebi, "A new region search method based on DOA estimation for speech source localization by SRP-PHAT method," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2010, pp. 656–660.

[245] H. R. Zarghi, M. Sharifkhani, and I. Gholampour, "Implementation of a cost efficient SSL based on an angular beamformer SRP-PHAT," in *Proc. IEEE Int. Conf. on Elec., Circuits, and Systems*, 2011, pp. 49–52.

[246] A. Johansson and S. Nordholm, "Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a realtime implementation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, Mar. 2005, pp. iv/933–iv/936.

[247] D.-B. Zhuo and H. Cao, "Fast sound source localization based on SRP-PHAT using density peaks clustering," *Applied Sciences*, vol. 11, no. 1, p. 445, 1 2021.

[248] B. Çakmak, T. Dietzen, R. Ali, P. Naylor, and T. van Waterschoot, "A distributed steered response power approach to source localization in wireless acoustic sensor networks," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2022.

[249] J. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 289–292.

[250] L. Gonzaga da Silveira Jr, V. Peruffo Minotto, C. Jung, and B. Lee, "A GPU implementation of the SRP-PHAT sound source localization algorithm," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2010.

[251] V. P. Minotto, C. R. Jung, L. Gonzaga da Silveira, and B. Lee, "GPU-based approaches for real-time sound source localization using the SRP-PHAT algorithm," *Int. J. of High Performance Computing Applications*, vol. 27, no. 3, pp. 291–306, 2013.

[252] T. Lee, S. Chang, and D. Yook, "Parallel SRP-PHAT for GPUs," *Computer Speech & Language*, vol. 35, pp. 1–13, 2016.

[253] J. A. Belloch, A. Gonzalez, A. M. Vidal, and M. Cobos, "Real-time sound source localization on graphics processing units," *Procedia Computer Science*, Int. Conf. on Computational Science, vol. 18, pp. 2549–2552, 2013.

[254] J. A. Belloch, A. Gonzalez, A. M. Vidal, and M. Cobos, "On the performance of multi-GPU-based expert systems for acoustic localization involving massive microphone arrays," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5607–5620, 2015.

[255] J. A. Belloch, M. Cobos, A. Gonzalez, and E. S. Quintana-Ortí, "Real-time sound source localization on an embedded GPU using a spherical microphone array," *Procedia Computer Science*, Int. Conf. On Computational Science, ICCS 2015, vol. 51, pp. 201–210, 2015.

[256] B. Lee and T. Kalker, "A vectorized method for computationally efficient SRP-PHAT sound source localization," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2010.

[257] J. M. Badía, J. A. Belloch, M. Cobos, F. D. Igual, and E. S. Quintana-Ortí, "Accelerating the SRP-PHAT algorithm on multi- and many-core platforms using OpenCL," *J. of Supercomputing*, vol. 75, no. 3, pp. 1284–1297, 2019.

[258] J. Yin and M. Verhelst, "CNN-based robust sound source localization with SRP-PHAT for the extreme edge," *ACM Trans. on Embedded Computing Systems*, vol. 22, no. 3, pp. 1–27, 2023.

[259] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.

[260] D. V. Rabinkin, R. J. Renomeron, A. J. Dahl, J. C. French, J. L. Flanagan, and M. Bianchi, "DSP implementation of source location using microphone arrays," in *Proc. SPIE Adv. Algorithms Architectures Signal Process.*, vol. 2846, 1996, pp. 88–99.

[261] M. Shen and H. Liu, "A modified cross power-spectrum phase method based on microphone array for acoustic source localization," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, Oct. 2009, pp. 1286–1291.

[262] T. Padois, O. Doutres, and F. Sgard, "On the use of modified phase transform weighting functions for acoustic imaging with the generalized cross correlation," *J. Acoust. Soc. Am.*, vol. 145, no. 3, pp. 1546–1555, Mar. 2019.

[263] A. Ramamurthy, H. Unnikrishnan, and K. D. Donohue, "Experimental performance analysis of sound source detection with SRP PHAT-β," in *Proc. IEEE Southeastcon*, 2009, pp. 422–427.

[264] M. Swartling, B. Sallberg, and N. Grbic, "Direction of arrival estimation for speech sources using fourth order cross cumulants," in *Proc. Int. Symp. on Circuits and Syst.*, 2008, pp. 1696–1699.

[265] A. Cirillo, R. Parisi, and A. Uncini, "Sound mapping in reverberant rooms by a robust direct method," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 285–288.

[266] H. He, X. Wang, Y. Zhou, and T. Yang, "A steered response power approach with trade-off prewhitening for acoustic source localization," *J. Acoust. Soc. Am.*, vol. 143, no. 2, pp. 1003–1007, 2018.

[267] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.

[268] M. Liu, J. Hu, Q. Zeng, Z. Jian, and L. Nie, "Sound source localization based on multi-channel cross-correlation weighted beamforming," *Micromachines*, vol. 13, no. 7, p. 1010, 7 2022.

[269] X. Wan and Z. Wu, "Improved steered response power method for sound source localization based on principal eigenvector," *Applied Acoust.*, vol. 71, no. 12, pp. 1126–1131, 2010.

[270] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, "Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1270–1281, 2020.

[271] M. Cobos, M. García-Pineda, and M. Arevalillo-Herráez, "Steered response power localization of acoustic passband signals," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 717–721, 2017.

[272] Y. Huang, J. Tong, X. Hu, and M. Bao, "A robust steered response power localization method for wireless acoustic sensor networks in an outdoor environment," *Sensors*, vol. 21, no. 5, p. 1591, 2021.

[273] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Acoustic source localization with deep generalized cross correlations," *Signal Process.*, vol. 187, p. 108 169, 2021.

[274] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2008, Jul. 2008.

[275] X. Wan and Z. Wu, "Improved speech source localization in reverberant environments based on correlation dimension," in *Proc. Int. Conf. on Wireless Commun. & Signal Proc.*, 2009, pp. 1–4.

[276] F. Hummes, J. Qi, and T. Fingscheidt, "Robust acoustic speaker localization with distributed microphones," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2011, pp. 240–244.

[277] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 581–585, 2014.

[278] P. Pertila and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, New Orleans, LA, 2017, pp. 6125–6129.

[279] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Trans. on Emerging Topics in Comput. Intel.*, vol. 2, no. 2, pp. 103–116, 2018.

[280] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 178–188, 2019.

[281] J. Wechsler, W. Mack, and E. A. P. Habets, "End-to-end signal-aware direction-of-arrival estimation using weighted steered-response power," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2022, pp. 41–45.

[282] J. Moragues, L. Vergara, J. Gosalbez, T. Machmer, A. Swerdlow, and K. Kroschel, "Background noise suppression for acoustic localization by means of an adaptive energy detection approach," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 2421–2424.

[283] T. Machmer, A. Swerdlow, K. Kroschel, J. Moragues, L. Vergara, and J. Gosálbez, "Robust impulsive sound source localization by means of an energy detector for temporal alignment and pre-classification," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2009, pp. 1409–1412.

[284] H. Lim, I.-C. Yoo, Y. Cho, and D. Yook, "Speaker localization in noisy environments using steered response voice power," *IEEE Trans. Consum. Electron.*, vol. 61, no. 1, pp. 112–118, 2015.

[285] X. Qian, Q. Zhang, G. Guan, and W. Xue, "Deep audio-visual beamforming for speaker localization," *IEEE Signal Process. Lett.*, vol. 29, pp. 1132–1136, 2022.

[286] D. Salvati, C. Drioli, and G. L. Foresti, "Frequency map selection using a RBFN-based classifier in the MVDR beamformer for speaker localization in reverberant rooms," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 3298–3301.

[287] D. Salvati, C. Drioli, and G. L. Foresti, "On the use of machine learning in microphone array beamforming for far-field sound source localization," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, 2016, pp. 1–6.

[288] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," en, *Pattern Recognition Lett.*, vol. 84, pp. 15–21, Dec. 2016.

[289] X. Zhao, L. Zhou, Y. Tong, Y. Qi, and J. Shi, "Robust sound source localization using convolutional neural network based on microphone array.," *Intelligent Automation & Soft Computing*, vol. 30, no. 1, 2021.

[290] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, 2007, pp. IV–493–IV–496.

[291] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari, "Verified speaker localization utilizing voicing level in split-bands," *Signal Process.*, vol. 89, no. 6, pp. 1038–1049, 2009.

[292] G. García-Barrios, E. Latorre Iglesias, J. M. Gutiérrez-Arriola, R. Fraile, N. Sáenz-Lechón, and V. J. Osma-Ruiz, "Exploiting spatial diversity for increasing the robustness of sound source localization systems against reverberation," *Applied Acoust.*, vol. 202, p. 109 138, 2023.

[293] A. Das H., L. Gopalakrishnan Pillai, and M. Chellappa, "Human voice localization in noisy environment by SRP-PHAT and MFCC," *Int. Research J. of Advanced Eng. and Science*, vol. 1, no. 3, pp. 33–37, 2016.

[294] J. Zhao and C. Ritz, "Investigating co-prime microphone arrays for speech direction of arrival estimation," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2018, pp. 1658–1664.

[295] J. Zhao and C. Ritz, "Semi-coprime microphone arrays for estimating direction of arrival of speech sources," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, 2019, pp. 308–313.

[296] X. Zhao, J. Tang, L. Zhou, and Z. Wu, "A fast search method of steered response power with small-aperture microphone array for sound source localization," *J. of Electronics (China)*, vol. 30, no. 5, pp. 483–490, 2013.

[297] A. Hirose, *Complex-Valued Neural Networks*. Springer Science & Business Media, 2012, vol. 400.

[298] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. Detect. and Classific. of Acoust. Scenes and Events (DCASE)*, 2019, pp. 10–14.

[299] H. Tsuzuki, M. Kugler, S. Kuroyanagi, and A. Iwata, "An approach for sound source localization by complex-valued neural network," *IEICE Tech. Rep.*, no. 10, pp. 2257–2265, 2013.

[300] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2020.

[301] C. Trabelsi, O. Bilaniuk, Y. Zhang, *et al.*, "Deep complex networks," in *Proc. Int. Conf. on Learning Representations*, 2018.

[302] D. P. Mandic and V. S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models.* John Wiley & Sons, 2009, vol. 59.

[303] K. Kreutz-Delgado, "The complex gradient operator and the cr-calculus," *Arxiv*, Jun. 2009.

Appendix A

Energy-based SSL: proof of the locus of source candidates

This proof was a joint work with Mike Brookes. In the aforementioned energy-based approaches for source localization [195], a the ratio between the measured energy levels at two microphones is used to restrict the available positions a source.

Although many papers claim such ratio restricts the candidate source positions to a hypersphere, no proof of this statement was found by the authors within the literature. We provide a proof for such statement below.

In (7) of [195], the so-called "energy ratio" (actually a distance ratio) for a pair of microphones located at $\mathbf{v}_l$ and $\mathbf{v}_m$ is defined as

$$k_{i,j} = \frac{|\mathbf{u} - \mathbf{v}_l|}{|\mathbf{u} - \mathbf{v}_m|}$$

where u is the source position and all vectors, in bold, are column vectors in d dimensions.

In order to remove the modulus operator, we square both sides of the equation. By using the relationship $|\mathbf{x}|^2 = \mathbf{x}^T\mathbf{x}$, we obtain:

$$k_{i,j}^2 = \frac{(\mathbf{u} - \mathbf{v}_l)^T (\mathbf{u} - \mathbf{v}_l)}{(\mathbf{u} - \mathbf{v}_m)^T (\mathbf{u} - \mathbf{v}_m)} \tag{A.1}$$

By multiplying both sides by the denominator and expanding the terms, we obtain:

$$\left(1-k_{i,j}^2\right)\mathbf{u}^T\mathbf{u}-2\left(\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m\right)\mathbf{u}+\left(\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m\right)=0 \tag{A.2}$$

This equation is the vectorial equivalent of the second-order polynomial $ar^2+br+c=0$. To obtain an equation of the form $\mathbf{u}^2+b\mathbf{u}+c=0$, we divide both sides by $\left(1-k_{i,j}^2\right)$:

$$\mathbf{u}^T\mathbf{u}-2\frac{\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m}{1-k_{i,j}^2}\mathbf{u}+\frac{\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m}{1-k_{i,j}^2}=0 \tag{A.3}$$

By defining $\mathbf{c}_{i,j}=\frac{\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m}{1-k_{i,j}^2}$, we obtain:

$$\mathbf{u}^T\mathbf{u}-2\mathbf{c}_{i,j}\mathbf{u}+\frac{\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m}{1-k_{i,j}^2}=0 \tag{A.4}$$

We recall the equation of the circle is defined as $(x-c)^T(x-c)=\rho^2$, where c represents the central point of the circle while ρ represents its radius. We then add $\mathbf{c}_{i,j}^2$ to both sides of the equation and see that the term $\mathbf{u}^T\mathbf{u}-2\mathbf{c}_{i,j}+\mathbf{c}_{i,j}^2$ appears. We are able to factor this polynomial to obtain the first side of the circle equation. This technique is informally known as "completing the square".

$$\left(r-\mathbf{c}_{i,j}\right)^T\left(r-\mathbf{c}_{i,j}\right)=\mathbf{c}_{i,j}^T\mathbf{c}_{i,j}-\frac{\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m}{1-k_{i,j}^2} \tag{A.5}$$

It then remains to prove that

$$\rho^2=\mathbf{c}_{i,j}^T\mathbf{c}_{i,j}-\frac{\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m}{1-k_{i,j}^2} \tag{A.6}$$

By replacing $\mathbf{c}_{i,j}$ with its original value, we obtain:

$$\rho^2=\frac{\left(\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m\right)^T\left(\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m\right)-\left(1-k_{i,j}^2\right)\left(\mathbf{v}_l^T\mathbf{v}_l-k_{i,j}^2\mathbf{v}_m^T\mathbf{v}_m\right)}{\left(1-k_{i,j}^2\right)^2} \tag{A.7}$$

By expanding the terms in the numerator, many elements cancel out, yielding the following equation:

$$\rho^2 = \frac{k_{i,j}^2 \mathbf{v}_l^T \mathbf{v}_l - 2k_{i,j}^2 \mathbf{v}_l^T \mathbf{v}_m + k_{i,j}^2 \mathbf{v}_m^T \mathbf{v}_m}{\left(1 - k_{i,j}^2\right)^2} \tag{A.8}$$

Finally, using the relationship $|\mathbf{x}|^2 = \mathbf{x}^T \mathbf{x}$, we are able to group the square term:

$$\rho^2 = \left(\frac{k_{i,j} \left| \mathbf{v}_l - \mathbf{v}_m \right|}{1 - k_{i,j}^2} \right)^2 \tag{A.9}$$

Which concludes the proof.