Acoustic Reflector Localisation for Blind 
Source Separation and Spatial Audio

by

Luca Remaggi

A thesis submitted in partial fulfillment for the degree of 
Doctor of Philosophy

UNIVERSITY OF 
SURREY

Centre for Vision, Speech and Signal Processing 
Faculty of Engineering and Physical Sciences 
University of Surrey 
Guildford, Surrey, GU27XH, UK

August 2017
Abstract

From a physical point of view, sound is classically defined by wave functions. Like every other physical model based on waves, during its propagation, it interacts with the obstacles it encounters. These interactions result in reflections of the main signal that can be defined as either being supportive or interfering. In the signal processing research field, it is, therefore, important to identify these reflections, in order to either exploit or avoid them, respectively.

The main contribution of this thesis focuses on the acoustic reflector localisation. Four novel methods are proposed: a method localising the image source before finding the reflector position; two variants of this method, which utilise information from multiple loudspeakers; a method directly localising the reflector without any pre-processing. Finally, utilising both simulated and measured data, a comparative evaluation is conducted among different acoustic reflector localisation methods. The results show the last proposed method outperforming the state-of-the-art. The second contribution of this thesis is given by applying the acoustic reflector localisation solution into spatial audio, with the main objective of enabling the listeners with the sensation of being in the recorded environment. A novel way of encoding and decoding the room acoustic information is proposed, by parametrising sounds, and defining them as reverberant spatial audio objects (RSAOs). A set of subjective assessments are performed. The results prove both the high quality of the sound produced by the proposed parametrisation, and the reliability on manually modifying the acoustic of recorded environments.

The third contribution is proposed in the field of speech source separation. A modified version of a state-of-the-art method is presented, where the direct sound and first reflection information is utilised to model binaural cues. Experiments were performed to separate speech sources in different environments. The results show the new method to outperform the state-of-the-art, where one interferer is present in the recordings.

The simulation and experimental results presented in this thesis represent a significant addition to the literature and will influence the future choices of acoustic reflector localisation systems, 3D rendering, and source separation techniques. Future work may focus on the fusion of acoustic and visual cues to enhance the acoustic scene analysis.
Declaration of Authorship

I confirm that the submitted work is my own work and that I have clearly identified and fully acknowledged all material that is entitled to be attributed to others (whether published or unpublished) using the referencing system set out in the programme handbook. I agree that the University may submit my work to means of checking this, such as the plagiarism detection service Turnitin® UK. I confirm that I understand that assessed work that has been shown to have been plagiarised will be penalised.

Signed: ____________________________

Date: ________________________________
“The most exciting phrase to hear in science, the one that heralds the most discoveries, is not “Eureka!” but “That’s funny...”

- Isaac Asimov
Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Philip Jackson, for the continuous support during my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I would also like to extend my gratitude to the rest of my supervision team: Dr. Wenwu Wang and Dr. Jean-Yves Guillemaut. Thank you for all the invaluable effort and guidance.

I would like to thank all the officemates I had during these years, and my research group components: Phil, Teo, Sam Zargham, Sam Fowler, Mark, Quinju, Hanne, Jon, Miguel, and Ben, for the great discussions and hilarious moments.

I must thank the other two components of the “trio”, Simone and Estephan. It has been an honour to meet you guys. I will never forget all the time spent together.

Thank you to all my family: nonna Alberta, nonno Luciano, nonna Bruna, nonno Rodolfo, zio Stefano, zia Annalisa, zio Valerio, zia Suzy, Anna, Dominic, Viola, Valentina, maman Vida, baba Mansour, and Armin. I would like to thank, in particular, my parents Grazia e Maurizio, my sister Roberta and my hairy brother Brian. You are amazing.

Finally, to Nazli for always being there, during both the cheerful and sad periods of this journey. I would have never done it without you.
## Contents

Abstract                                                i  
Declaration of Authorship                                ii 
Acknowledgements                                        iv  
List of Figures                                         ix  
List of Tables                                          xi  
Abbreviations                                           xiii 
Symbols                                                 xvii 

### 1 Introduction

1.1 Motivation                                           2  
1.2 Problem Statements and Applications                  3  
   1.2.1 Acoustic Reflector Localisation                   3  
   1.2.2 Spatial Audio                                    4  
   1.2.3 Source Separation                                5  
1.3 Contributions and Thesis Outline                     7  
   1.3.1 Chapter 2: Conceptual Foundations                7  
   1.3.2 Chapter 3: Reflector Localisation and its Applica-
   tions: a Literature Review                           7  
   1.3.3 Chapter 4: Acoustic Reflector Localisation Meth-
   ods                                                   8  
   1.3.4 Chapter 5: Parametrisation of Reverberant Spatial
   Audio Objects                                         9  
   1.3.5 Chapter 6: Source Separation through Multipath
   Propagation Analysis                                  11 
   1.3.6 Chapter 7: Conclusion                            11 
   1.3.7 Appendixes                                       11 

### 2 Conceptual Foundations

2.1 Sound Wave Propagation                              13  
   2.1.1 Acoustical Wave Equation                        14  
   2.1.2 Solution to the Wave Equation                   16
2.2 Acoustic Reflection .............................................. 19
  2.2.1 Specular Reflection ......................................... 19
  2.2.2 Scattering and Diffraction ................................ 21
2.3 Room Acoustic Modelling ...................................... 23
  2.3.1 The Green’s Function ....................................... 23
  2.3.2 The Image Source Model ................................... 25
  2.3.3 The Room Impulse Response ............................... 27
2.4 The Sound Perception in Indoor Environments .............. 29
  2.4.1 The Perception of the First Reflection .................... 29
  2.4.2 The Mixing Time .......................................... 30
  2.4.3 The Reverberation Time ................................. 31
3 Reflector Localisation and its Applications: a Literature Review 33
  3.1 On Acoustic Reflector Localisation ......................... 33
    3.1.1 Reflector Estimation from RIRs .......................... 34
    3.1.2 Literature Limitations ................................ 39
    3.1.3 Reflector Estimation From Continuous Signals .......... 39
  3.2 On Spatial Audio .............................................. 40
    3.2.1 Spatial Audio Scene Representation ..................... 40
    3.2.2 Object based Audio Literature Limitations .......... 44
    3.2.3 Spatial Audio Object Coding ............................ 44
  3.3 On Blind Source Separation ................................ 45
    3.3.1 Speech Source Separation ............................... 46
    3.3.2 Source Separation including Early Reflection Information 48
    3.3.3 Literature Limitations ................................ 49
  3.4 Fundamental Tools ........................................... 50
    3.4.1 Epoch Detectors ........................................ 50
    3.4.2 Beamformers ............................................ 51
    3.4.3 Linear Predictive Coding ............................... 52
    3.4.4 Homogeneous Coordinates ............................... 53
    3.4.5 Linear Transformations ................................ 54
4 Acoustic Reflector Localisation given Room Impulse Responses 57
  4.1 Introduction .................................................. 57
  4.2 Background and Preliminaries .............................. 60
    4.2.1 The Room Impulse Response (RIR) ....................... 60
    4.2.2 The Image Source Model ............................... 60
    4.2.3 Method Classification and Overviews .................... 61
    4.2.4 Method Assumptions .................................. 62
    4.2.5 Common Pre-Processing ................................ 63
  4.3 State-of-the-Art Methods .................................. 65
    4.3.1 Maximum Likelihood (ML) ............................... 66
    4.3.2 Multilateration ........................................ 66
    4.3.3 The Loudspeaker-Image Bisection (LIB) Algorithm .... 67
  4.4 Proposed Methods ........................................... 68
    4.4.1 Image-Source Reversion Methods ....................... 68
4.5 Experimental Evaluation and Discussion ........................................ 74
  4.5.1 Datasets .................................................................................... 75
  4.5.2 Evaluation Metrics ...................................................................... 79
  4.5.3 C-DYPSA Evaluation .................................................................. 81
  4.5.4 Simulated Experiments ................................................................. 82
  4.5.5 Comparative Evaluation with Recorded Data ................................. 84
  4.5.6 Computational Complexity .......................................................... 88

4.6 Conclusion ...................................................................................... 88

5 Parametrisation of Reverberant Spatial Audio Objects ....................... 91
  5.1 Background .................................................................................... 91
    5.1.1 Auralisation .............................................................................. 92
    5.1.2 Spatial Audio Object Coding .................................................... 92
    5.1.3 Spatial Audio Object Parametrisation ........................................ 93
    5.1.4 The Reverberant Spatial Audio Object Novelties ....................... 95
    5.1.5 RIR Definition ......................................................................... 96
  5.2 Reverberant Spatial Audio Object Encoding ...................................... 98
    5.2.1 TOA Estimation and RIR Segmentation .................................... 98
    5.2.2 Direction of Arrival Estimation ................................................ 100
    5.2.3 Colour Estimation ...................................................................... 101
    5.2.4 Late Reverberation Parametrisation ......................................... 102
  5.3 Reverberant Spatial Audio Object Decoding and Rendering ............... 102
    5.3.1 Direct and Early Spatial Audio Objects .................................... 103
    5.3.2 Late Diffuse Spatial Audio Object ............................................. 104
    5.3.3 Mixer ....................................................................................... 105
  5.4 Experimental Evaluation ................................................................. 105
    5.4.1 Subjective Assessments ............................................................ 106
  5.5 Conclusion ...................................................................................... 112

6 Source Separation through Multipath Propagation Analysis ............... 115
  6.1 Introduction .................................................................................... 115
  6.2 Background Definitions ................................................................. 118
    6.2.1 Proposed Method Assumptions ................................................. 118
    6.2.2 Binaural Room Impulse Response ........................................... 119
    6.2.3 Interaural Spectrogram ............................................................. 121
  6.3 Proposed Interaural Cue Model ....................................................... 121
    6.3.1 Interaural Level and Phase Differences .................................... 121
    6.3.2 Probability Distributions .......................................................... 123
  6.4 Source Separation Model ............................................................... 123
    6.4.1 Parameter Estimation from Mixtures ....................................... 123
    6.4.2 Expectation-Maximisation (EM) .............................................. 124
    6.4.3 Initialisation Method ................................................................. 126
  6.5 Experiments and Results ............................................................... 127
    6.5.1 Datasets .................................................................................. 127
    6.5.2 Evaluation Metrics ................................................................. 131
    6.5.3 Control Masks .......................................................................... 132
    6.5.4 Source Separation Experiments .............................................. 133
6.5.5 Implementation Issues ........................................ 138
6.6 Conclusion ...................................................... 138

7 Conclusion ....................................................... 141

A Multichannel Room Impulse Response Visualisation 145
A.1 Introduction .................................................. 145
A.2 Room Visualisation Techniques .............................. 146
A.3 Recorded Dataset Visualisation ................................. 147
  A.3.1 Recorded Multichannel RIR Datasets .................. 147
  A.3.2 Multichannel RIRs Visualisation ....................... 148
A.4 Conclusion ....................................................... 150

B The CISPE Algorithm 151
B.1 Introduction .................................................. 151
B.2 Recording Setup Estimation .................................. 153
  B.2.1 Sensor Positions - The CISPE Algorithm ............ 153
  B.2.2 Source and Reflector Positions ......................... 155
B.3 Experiments ................................................... 157
  B.3.1 Recording Setup ........................................ 158
  B.3.2 Reflector Estimation .................................... 158
B.4 Conclusion ....................................................... 160

Bibliography 161
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Space-time domain visualisation of the sound wave propagation.</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Visualisation of the plane wave assumption.</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Representation of wavefronts and related harmonic wave.</td>
<td>18</td>
</tr>
<tr>
<td>2.4</td>
<td>Graphical representation of a specular reflections.</td>
<td>20</td>
</tr>
<tr>
<td>2.5</td>
<td>Representation of different type of reflections.</td>
<td>22</td>
</tr>
<tr>
<td>2.6</td>
<td>Representation of different diffraction causes.</td>
<td>23</td>
</tr>
<tr>
<td>2.7</td>
<td>Representation of the image source position calculation.</td>
<td>25</td>
</tr>
<tr>
<td>2.8</td>
<td>Simulated and recorded room impulse responses (RIRs).</td>
<td>27</td>
</tr>
<tr>
<td>2.9</td>
<td>Representation of the three RIR components.</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>Ellipse and parabola properties.</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>Loudspeaker image bisection (LIB) algorithm schematisation.</td>
<td>38</td>
</tr>
<tr>
<td>3.3</td>
<td>Loudspeaker setups for spatial audio.</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Beamforming, independent component analysis (ICA) and time-frequency (TF) masking source separation methods.</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Linear geometric transformations.</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview of the tested acoustic reflector localisation methods.</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>Overview of the common pre-processing stage.</td>
<td>63</td>
</tr>
<tr>
<td>4.3</td>
<td>The DYPSA phase-slope function.</td>
<td>64</td>
</tr>
<tr>
<td>4.4</td>
<td>Clustering the DYPSA outputs.</td>
<td>65</td>
</tr>
<tr>
<td>4.5</td>
<td>The LIB algorithm.</td>
<td>67</td>
</tr>
<tr>
<td>4.6</td>
<td>Different loudspeaker midpoints.</td>
<td>70</td>
</tr>
<tr>
<td>4.7</td>
<td>Output of the ellipsoid tangent sample consensus (ETSAC) algorithm.</td>
<td>71</td>
</tr>
<tr>
<td>4.8</td>
<td>Floor plan view of the four measured rooms.</td>
<td>76</td>
</tr>
<tr>
<td>4.9</td>
<td>Pictures of the four recorded rooms.</td>
<td>77</td>
</tr>
<tr>
<td>4.10</td>
<td>The average absorption coefficient of the recorded rooms.</td>
<td>77</td>
</tr>
<tr>
<td>4.11</td>
<td>Double circular microphone array.</td>
<td>78</td>
</tr>
<tr>
<td>4.12</td>
<td>Analysis of the ETSAC RANSAC sample number.</td>
<td>85</td>
</tr>
<tr>
<td>4.13</td>
<td>Overview of the reflection positions estimated using ETSAC.</td>
<td>87</td>
</tr>
<tr>
<td>5.1</td>
<td>General spatial audio object coding structure overview.</td>
<td>93</td>
</tr>
<tr>
<td>5.2</td>
<td>System diagram overview of the reverberant spatial audio object (RSAO) encoder, decoder and renderer.</td>
<td>97</td>
</tr>
<tr>
<td>5.3</td>
<td>The clustered dynamic dynamic programming projected phase-slope algorithm (C-DYPSA) output.</td>
<td>99</td>
</tr>
<tr>
<td>5.4</td>
<td>Power in output of the delay-and-sum beamformer.</td>
<td>100</td>
</tr>
<tr>
<td>5.5</td>
<td>The linear predictive coding (LPC) output.</td>
<td>101</td>
</tr>
<tr>
<td>5.6</td>
<td>Spectrogram of three late reverberation octave bands.</td>
<td>104</td>
</tr>
</tbody>
</table>
5.7  The uniform circular array close-up. ................................. 106  
5.8  Representation of the parametric RIRs. .............................. 108  
5.9  Perceptual scores for the subjective assessment tests. ............... 110  

6.1  Representation of an ideal BRIR direct sound and first reflection. ....... 120  
6.2  Interaural phase difference for a source convolved with an ideal BRIR. .... 122  
6.3  Pictures of the four room setups. .................................. 127  
6.4  Plan view of the four datasets employed. ............................ 128  
6.5  RT60s and averaged absorption coefficients for the recorded datasets. .... 129  
6.6  MESSL, ER-MESSL and Oracle TF masks. ............................ 133  
6.7  Signal to distortion ratios for different target source positions. .......... 135  
6.8  Polar plot of the signal to distortion ratio for two datasets. .............. 136  
6.9  Signal to distortion ratios for different interferer positions. .............. 136  

A.1  Raw RIR data visualisation and direction of arrival (DOA)-time energy  
analysis. ........................................................................ 149  
A.2  Visualisation of the reflection and reflector estimation. ................. 149  

B.1  Initialised and estimated microphone positions. .......................... 155  
B.2  The “Surrey sound sphere” and measurement setup. ...................... 157  
B.3  Root mean square errors (RMSEs) and standard errors related to the  
reflector position estimation with and without CISPE. ...................... 159
List of Tables

4.1 3D method categorisation. .............................................. 61
4.2 Parameter values used for the experiments. ....................... 74
4.3 Room dimensions and volumes for the 10 simulated rooms. ........ 75
4.4 RIR dataset room properties. ........................................... 77
4.5 Gross and fine errors for the first reflection time of arrival estimation. 81
4.6 Gross and fine errors for the reflector position estimation with simulated rooms. ........................................ 82
4.7 Gross and fine errors for the reflector position estimation with simulated rooms, varying the direct to noise ratio. .................... 83
4.8 Gross and fine errors for the image source localisation with recorded rooms. ........................................ 85
4.9 Fine errors for the image source localisation with recorded rooms. ........................................ 86

5.1 Overview of the proposed reverberant spatial audio object and the state-of-the-art. ............................................. 94
5.2 Parameter values used for the listening tests. ....................... 105
5.3 Volume, RT60 and source-listener distance, for every room. ........ 107
5.4 p-values for the apparent source distance results. ................... 109
5.5 p-values for the apparent room size results. ......................... 111
5.6 p-values for the listener envelopment results. ....................... 112

6.1 Recorded room reverberation time, direct to reverberant ratio, and target-interferer separation angle. ............................. 127
6.2 Signal to distortion ratios for the one interferer source case. ........ 134
6.3 Perceptual evaluation of speech quality for the one interferer source case. ........................................ 134
6.4 P-values of the paired t-test comparing the MESSL and ER-MESSL results, with one interferer. ................ ...................... 134
6.5 Signal to distortion ratio for the two interferer sources case. ........ 137
6.6 Perceptual evaluation of speech quality for the two interferer sources case. ........................................ 137
6.7 Allowed ranges around the initialised parameter values. .............. 138

A.1 Properties of the three presented rooms. .............................. 147
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>1 Dimensional</td>
</tr>
<tr>
<td>2D</td>
<td>2 Dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>3 Dimensional</td>
</tr>
<tr>
<td>ADM</td>
<td>Audio Definition Model</td>
</tr>
<tr>
<td>AR</td>
<td>Augmented Reality</td>
</tr>
<tr>
<td>AVG</td>
<td>AVeraGe</td>
</tr>
<tr>
<td>AVG-TISA</td>
<td>AVeraGe Target-Interferer Separation Angle</td>
</tr>
<tr>
<td>BCMDS</td>
<td>Basis-point Classical MDS</td>
</tr>
<tr>
<td>BIFS</td>
<td>BIinary Format Scenes</td>
</tr>
<tr>
<td>BRIR</td>
<td>Binaural Room Impulse Response</td>
</tr>
<tr>
<td>CASA</td>
<td>Computational Auditory Scene Analysis</td>
</tr>
<tr>
<td>C-DYPSA</td>
<td>Clustered-DYPSA</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CISPE</td>
<td>Cross-correlation based Iterative Sensor Position Estimation</td>
</tr>
<tr>
<td>COTA</td>
<td>COmmon TAngent</td>
</tr>
<tr>
<td>DirAC</td>
<td>Directional Audio Coding</td>
</tr>
<tr>
<td>DNR</td>
<td>Direct to Noise Ratio</td>
</tr>
<tr>
<td>DOA</td>
<td>Direction Of Arrival</td>
</tr>
<tr>
<td>DRR</td>
<td>Direct to Reverberant Ratio</td>
</tr>
<tr>
<td>DSB</td>
<td>Delay-and-Sum-Beamformer</td>
</tr>
<tr>
<td>DYPSA</td>
<td>DYnamic programming projected Phase Slope Algorithm</td>
</tr>
<tr>
<td>Ebrief</td>
<td>Engineering brief</td>
</tr>
<tr>
<td>EDM</td>
<td>Euclidean Distance Matrix</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximisation</td>
</tr>
</tbody>
</table>
Abbreviations

**ER-MESSL** Early Reflection Model-based Expectation-maximisation Source Separation and Localisation

**ESPRIT** Estimation of Signal Parameters via Rotational Invariance Techniques

**ETSAC** Ellipsoid Tangent SAmple Consensus

**FDN** Feedback Delay Network

**FIR** Finite Impulse Response

**GMM** Gaussian Mixture Model

**IC** Independent Component

**ICA** Independent Component Analysis

**ILD** Interaural Level Difference

**IPD** Interaural Phase Difference

**ISDAR** Image Source Direction And Ranging

**ITD** Interaural Time Difference

**ITU** International Telecommunication Union

**JND** Just Noticeable Difference

**JSON** JavaScript Object Notation

**LFE** Low Frequency Effect

**LIB** Loudspeaker Image Bisection

**LOLO** Leave One Loudspeaker Out

**LR-TOA** Late Reverberation TOA

**LS** Least-Squares

**LPC** Linear Predictive Coding

**MDS** MultiDimensional Scaling

**MESSL** Model-based Expectation-maximisation Source Separation and Localisation

**ML** Maximum Likelihood

**MLS** Maximum Length Sequence

**MOS** Mean Opinion Score

**MPEG** Moving Picture Experts Group

**MUSIC** MUltiple SIgnal Classification

**MV** Mixing Vector

**MVDR** Minimum Variance Distortionless Response
<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PESQ</td>
<td>Perceptual Evaluation of Speech Quality</td>
</tr>
<tr>
<td>PWD</td>
<td>Plane Wave Decomposition</td>
</tr>
<tr>
<td>RANSAC</td>
<td>RANdom SAmple Consensus</td>
</tr>
<tr>
<td>RIR</td>
<td>Room Impulse Response</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>RSAO</td>
<td>Reverberant Spatial Audio Object</td>
</tr>
<tr>
<td>RT60</td>
<td>Reverberation Time (60 dB)</td>
</tr>
<tr>
<td>RTF</td>
<td>Room Transfer Function</td>
</tr>
<tr>
<td>SAC</td>
<td>Spatial Audio Coding</td>
</tr>
<tr>
<td>SAO</td>
<td>Spatial Audio Object</td>
</tr>
<tr>
<td>SAOC</td>
<td>Spatial Audio Object Coding</td>
</tr>
<tr>
<td>SDA</td>
<td>SuperDiRective Array</td>
</tr>
<tr>
<td>SDM</td>
<td>Spatial Decomposition Method</td>
</tr>
<tr>
<td>SDR</td>
<td>Signal to Distortion Ratio</td>
</tr>
<tr>
<td>SIRR</td>
<td>Spatial Impulse Response Rendering</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SOFA</td>
<td>Spatially Oriented Format for Acoustics</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound Pressure Level</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time Difference Of Arrival</td>
</tr>
<tr>
<td>TF</td>
<td>Time-Frequency</td>
</tr>
<tr>
<td>TISA</td>
<td>Target-Interferer Separation Angle</td>
</tr>
<tr>
<td>TOA</td>
<td>Time Of Arrival</td>
</tr>
<tr>
<td>UCA</td>
<td>Uniform Circular Array</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>ULA</td>
<td>Uniform Linear Array</td>
</tr>
<tr>
<td>URA</td>
<td>Uniform Rectangular Array</td>
</tr>
<tr>
<td>USAC</td>
<td>Unified Speech and Audio Coding stage</td>
</tr>
<tr>
<td>VBAP</td>
<td>Vector-Based Amplitude Panning</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual Reality</td>
</tr>
<tr>
<td>W-AVG</td>
<td>Weighted-AVerAge</td>
</tr>
<tr>
<td>WFS</td>
<td>Wave Field Synthesis</td>
</tr>
</tbody>
</table>
Symbols

\( a \) Acceleration
\( a_{\text{ILD}} \) Interaural level difference model
\( a(\cdot) \) All-pole filter
\( A(\cdot) \) All-pole filter power spectrum
\( A \) Vector containing the sensor position coordinates
\( A^{\text{part}} \) Vector containing the partially estimated sensor coordinates
\( A^{\text{init}} \) Vector containing the initialised sensor coordinates
\( A_x \) \( x \) coordinate of the sensor position
\( A_y \) \( y \) coordinate of the sensor position
\( A_z \) \( z \) coordinate of the sensor position
\( b \) All-pole filter coefficient index
\( b_{x,0} \) \( x \) coordinate of the source position
\( b_{y,0} \) \( y \) coordinate of the source position
\( b_{z,0} \) \( z \) coordinate of the source position
\( b_{x,e} \) \( x \) coordinate of the \( e \)-th image source position
\( b_{y,e} \) \( y \) coordinate of the \( e \)-th image source position
\( b_{z,e} \) \( z \) coordinate of the \( e \)-th image source position
\( B \) Number of time-frequency bins
\( B_0 \) Vector containing the source position coordinates
\( B_e \) Vector containing the \( e \)-th image source position coordinates
\( B_G \) Vector with the image source coordinate groundtruth
\( \Delta B_r \) Vector with the \( r \)-th permutated image source coordinates
\( \Delta B_m \) Vector with the reflector length information
\( c \) Source combination index
\( c_0 \) Speed of sound
Symbols

\( C \)  RANSAC samples
\( \mathbf{C} \)  Vector containing the interaural phased difference parameters
\( \text{CI} \)  Confidence interval for the image source localisation error \( \epsilon \)
\( \text{CI}_{\text{RMSE}} \)  Confidence interval for the reflector localisation root mean square error
\( \text{CI}_{\text{TOA}} \)  Confidence interval for the time of arrival error
\( d \)  Reflector irregularity width
\( d(\cdot) \)  Arbitrary signal
\( d^{+}(\cdot) \)  Arbitrary signal positively propagating in space
\( d^{-}(\cdot) \)  Arbitrary signal negatively propagating in space
\( d_{\text{cm}} \)  Distance between a microphone and the microphone array centre
\( d_{\text{ff}} \)  Far field limit distance
\( d_{\text{mm}} \)  Microphone array aperture
\( D \)  Number of datasets available for the experiment
\( D_{\text{PS}}(\cdot) \)  Power spectrum of a signal \( d(\cdot) \)
\( e \)  Reflection index
\( e^{D} \)  Energy decay
\( e^{\overline{D}} \)  Mean of the energy decays
\( \mathbf{E} \)  Quadratic surface matrix
\( \mathbf{E}_{I} \)  Sphere matrix
\( \mathbf{E}_{g} \)  Reduced version of an ellipsoid matrix
\( E(\cdot) \)  Error function
\( E^{\text{dist}}(\cdot) \)  Distortion function in a mixture of sources
\( E_{\text{ML}}(\cdot) \)  Maximum-likelihood algorithm error function
\( E_{\text{artif}} \)  Artefact error term
\( E_{\text{interf}} \)  Interference error term
\( E_{\text{noise}} \)  Noise error term
\( E_{\text{tot}} \)  Total error
\( E^{C} \)  Error in the microphone positioning
\( f \)  Frequency
\( f_{c} \)  Cutoff frequency
\( f_{s} \)  Sampling frequency
\( F \)  Force
\( g \)  Sensor combination index
Symbols

\( g(\cdot) \) Green’s function
\( G_a \) Gain factor of the all-pole filter \( a(\cdot) \)
\( G \) Gross error for the image source localisation
\( G_{\text{RMSE}} \) Gross error for the reflector localisation
\( G_{\text{TOA}} \) Gross error for the time of arrival estimation
\( G(\cdot) \) Laplace transform of the Green’s function
\( G_{\text{gd}}(\cdot) \) Group delay function
\( G(\cdot) \) Fourier transform of the Green’s function
\( h \) Frequency subband index
\( h_{11} \)
\( h_{12} \)

... The sixteen quadratic surface matrix coefficients
\( h_{34} \)
\( h_{44} \)
\( h(\cdot) \) General room impulse response component
\( \hat{h}(\cdot) \) Estimated general room impulse response component
\( h^D(\cdot) \) Direct sound of a room impulse response
\( h^E(\cdot) \) Early reflections of a room impulse response
\( h^L(\cdot) \) Late reverberation of a room impulse response
\( H \) General transformation matrix
\( H^{\text{gd}}(\cdot) \) Hann window (used for the group delay function)
\( H^{\text{seg}}(\cdot) \) Hamming window (used for segmenting the room impulse response)
\( i \) Sensor index
\( I(\cdot) \) Room impulse response
\( \hat{I}(\cdot) \) Approximated room impulse response
\( \hat{I}^{\text{ang}}(\cdot) \) Approximated room impulse response phase
\( I^B(\cdot) \) Beamformed room impulse response
\( I^{\text{BD}}(\cdot) \) Delayed beamformed room impulse response
\( I^D(\cdot) \) Delayed room impulse response
\( I^S(\cdot) \) Segmented room impulse response
\( I^{\text{SF}}(\cdot) \) Segmented and filtered Room impulse response
\( j \) Imaginary unit \((j = \sqrt{-1})\)
\( J(\cdot) \) Fisher information function
Symbols

\( k \)  
Angular wavenumber

\( K \)  
All-pole filter order

\( \mathbf{K}^H \)  
Translated plane normal vector

\( l \)  
Source index

\( l' \)  
Index for a source different from \( l \)

\( l \)  
Bisecting line

\( L \)  
Number of sources selected from the dataset for the experiments

\( L_I \)  
Number of sources localising image sources with fine errors

\( L_R \)  
Number of sources localising reflectors with fine errors

\( L_x \)  
Room \( x \) dimension length

\( L_y \)  
Room \( y \) dimension length

\( L_z \)  
Room \( z \) dimension length

\( L_{TOT} \)  
Number of total loudspeakers in the recorded dataset

\( L(\cdot) \)  
Log-likelihood function

\( m \)  
Discrete time bin index

\( m_a \)  
Mass

\( m_1 \)

\( m_2 \)  
Three integer numbers

\( m_3 \)

\( \mathbf{m}_x \)  
Vector of integer numbers

\( M \)  
Number of sensors available

\( M_1 \)  
Midpoint first component

\( M_2 \)  
Midpoint second component

\( M_3 \)  
Midpoint third component

\( \mathbf{M} \)  
Source-image source midpoint

\( \overline{\mathbf{M}} \)  
Mean of the source-image source midpoint

\( \widetilde{\mathbf{M}} \)  
Median of the source-image source midpoint

\( M(\cdot) \)  
Time-frequency mask

\( M^{ora}(\cdot) \)  
Time-frequency Oracle mask

\( n \)  
Discrete time variable

\( \hat{n} \)  
Estimated times of arrival

\( \bar{n} \)  
Mean of the estimated times of arrival

\( \tilde{n} \)  
Median of the estimated times of arrival
Symbols

\( n_e \) Time of arrival (in samples) related to the \( e \)-th reflection

\( n_{T_{m+1}} \) Mixing time

\( n_{T_{m+1}}^P \) Perceptual mixing time

\( n^C \) Time delay applied to a signal

\( n^D \) Direct sound time of arrival for the first binaural sensor

\( n^{DF} \) Delay between direct sound and first reflection (first binaural sensor)

\( n^{DS} \) Interaural time difference of arrival for the direct sound

\( n^{ML} \) Time of arrival for a randomly generated point in space

\( n^{ST} \) Interaural time difference of arrival for the first reflection

\( \hat{n} \) Vector containing the estimated times of arrival

\( n^{ML} \) Vector with the times of arrival for randomly generated points

\( n^{DSB} \) Vector with the delays used by the delay-and-sum beamformer

\( N \) Number of source-sensor combination

\( o \) Dataset index

\( O \) Weight coefficient

\( p \) Tested plane index

\( p_1 \) First plane coefficient

\( p_2 \) Second plane coefficient

\( p_3 \) Third plane coefficient

\( p_4 \) Fourth plane coefficient

\( p(\cdot) \) Probability distribution

\( p^{ML}(\cdot) \) Probability distribution for the maximum-likelihood based reflector localiser

\( P \) Air pressure

\( \hat{P} \) Air pressure amplitude

\( \bar{P} \) Air pressure amplitude of the reflection

\( \bar{P}^\prime \) Altered air pressure amplitude of the reflection

\( P^{inc} \) Air pressure of the incident sound

\( P^{ref} \) Air pressure of the reflected sound

\( P_{B_0}(\cdot) \) Air pressure wave function of the direct sound

\( P_e(\cdot) \) Air pressure wave function including the \( e \)-th image source

\( P_{sp}(\cdot) \) Total air pressure wave function for specular reflections

\( p \) Vector containing a plane coefficients
\( \mathbf{p}_l \)  Vector of the plane coefficients obtained by ISDAR-LIB for the \( l \)-th source

\( \mathbf{p}_{\text{ETSAC}} \)  Vector containing the plane coefficients obtained by ETSAC

\( \mathbf{p}_G \)  Vector containing the groundtruth plane coefficients

\( \mathbf{p}_M \)  Vector with the plane coefficients obtained by Mean-ISDAR-LIB

\( \tilde{\mathbf{p}}_{\text{MED}} \)  Vector with the plane coefficients obtained by Median-ISDAR-LIB

\( q \)  Index of the points used for the reflector localisation evaluation

\( \mathbf{q}_b \)  All-pole filter coefficients

\( \mathbf{q} \)  Vector containing ellipsoid coefficients (i.e. \( h_{14} \), \( h_{24} \) and \( h_{34} \))

\( Q \)  Temperature

\( Q_{\text{scal}}^{x} \)  Scaling factor along the \( x \)-axis

\( Q_{\text{scal}}^{y} \)  Scaling factor along the \( y \)-axis

\( Q_{\text{scal}}^{z} \)  Scaling factor along the \( z \)-axis

\( Q(\cdot) \)  Autocorrelation function

\( Q_a(\cdot) \)  All-pole filter autocorrelation function

\( Q_{\text{CC}}(\cdot) \)  Crosscorrelation function

\( r \)  Permutation index

\( r_{\text{head}} \)  Radius of the human head

\( R \)  Pressure reflector factor

\( R \)  Total rotation matrix

\( R_x \)  Rotation matrix around the \( x \)-axis

\( R_y \)  Rotation matrix around the \( y \)-axis

\( R_z \)  Rotation matrix around the \( z \)-axis

\( \text{RMSE}_{\text{TOA}} \)  RMS error for the TOA estimation

\( \text{RT60} \)  Reverberation time

\( s \)  Laplace domain variable

\( S \)  Volumetric surface

\( S_{\text{TOT}} \)  Total reflective surface area

\( S \)  Scaling matrix

\( S_{\text{gd}}(\cdot) \)  Phase-slope function

\( t \)  Continuous time variable

\( t_0 \)  Time of arrival (in seconds) of the direct sound

\( \delta t \)  Continuous time increment
$T$  Temporal period
$T_{gd}$  Length of the group delay Hann window
$T_m$  Last reflection before the mixing time
$T_L$  Last reflection recorded in the room impulse response
$T_{seg}$  Length of the room impulse response segmentation Hamming window
$T$  Translation matrix
$u$  Autocorrelation variable
$u_1$  First component of the normal vector
$u_2$  Second component of the normal vector
$u_3$  Third component of the normal vector
$u$  Normal vector to a plane
$\bar{u}$  Mean of a normal vector
$\tilde{u}$  Median of a normal vector
$u$  Normal vector to a plane
$u^E$  Normal vector to an ellipsoid
$u^s$  Normal vector to a surface
$U$  Number of utterances used
$U_{en}$  Energy of a signal
$U(\cdot)$  Unitary step function
$v_x$  $x$ component of the total sound wave particle velocity
$v^\text{inc}_x$  $x$ component of the incident sound wave particle velocity
$v^\text{ref}_x$  $x$ component of the reflected sound wave particle velocity
$v_y$  $y$ component of the total sound wave particle velocity
$v_z$  $z$ component of the total sound wave particle velocity
$v$  Particle velocity vector
$v^\text{inc}$  $x$ Incident sound wave particle velocity vector
$v^\text{ref}$  $x$ Reflected sound wave particle velocity vector
$V$  Volume
$V_{TOT}$  Total volume of a room
$w_1$  First constant to calculate the fourth plane coefficient in ETSAC
$w_2$  Second constant to calculate the fourth plane coefficient in ETSAC
$w_3$  Third constant to calculate the fourth plane coefficient in ETSAC
$w(\cdot)$  White Gaussian noise
$w^I(\cdot)$  Interaural noise

$w^R(\cdot)$  Exponential decaying Gaussian noise

$W$  Window defining the $W$ disjoint orthogonality

$W^B$  Number of frequency subbands

$x$  $x$-axis spatial variable

$x'$  $x$-axis spatial variable for the rotated axis

$x^{ev}$  Number of points used to evaluate one reflector position

$x^{ML}$  Points used by the maximum-likelihood based reflector locator

$x^{mul}$  Number of image sources estimated by the multilateration algorithm

$\delta x$  $x$ coordinate increase

$\Delta x$  Translation component for the $x$-axis

$x(\cdot)$  Signal produced by a source

$X$  Vector containing the three Cartesian coordinates of a point

$X^H$  Set of points used by the Hough transform

$X_{hom}$  Vector containing the Homogeneous coordinates of a point

$X^{ML}$  Set of points used by the maximum-likelihood based reflector locator

$X_T$  Vector containing the parabola tangent point Cartesian coordinates

$\Delta X$  Vector containing the three translation components

$y$  $y$-axis spatial variable

$y(\cdot)$  Signal received at the sensor position

$y^{IS}(\cdot)$  Interaural spectrogram

$y^{tar}(\cdot)$  Target signal for the separation

$\hat{y}(\cdot)$  Estimated direct sound early reflections received by a decoder

$\hat{y}^P(\cdot)$  Signal panned by a renderer/decoder

$\hat{y}^{P}(\cdot)$  Signal estimated for the reverberation received by a decoder

$\Delta y$  Translation component for the $y$-axis

$Y$  Admittance coefficient of the reflector

$z$  $z$-axis spatial variable

$z_2$  Fourth coordinate in the homogeneous coordinate system

$\Delta z$  Translation component for the $z$-axis

$Z$  Impedance coefficient of the reflector

$Z_0$  Characteristic impedance of the medium
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\alpha}$</td>
<td>Averaged absorption coefficient of the room</td>
</tr>
<tr>
<td>$\alpha_m$</td>
<td>Attenuation coefficient of the medium</td>
</tr>
<tr>
<td>$\alpha_p$</td>
<td>Absorption coefficient of the reflector</td>
</tr>
<tr>
<td>$\alpha_{\text{path}}$</td>
<td>Path dependent attenuation</td>
</tr>
<tr>
<td>$\alpha_{\text{rot}}$</td>
<td>Angle of rotation related to the $x$-$y$ plane</td>
</tr>
<tr>
<td>$\alpha_{\text{ILD}}(\cdot)$</td>
<td>Interaural level difference observation</td>
</tr>
<tr>
<td>$\beta_{\text{rot}}$</td>
<td>Angle of rotation related to the $x$-$z$ plane</td>
</tr>
<tr>
<td>$\gamma_{\text{rot}}$</td>
<td>Angle of rotation related to the $y$-$z$ plane</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>Vector containing both azimuth and elevation angles</td>
</tr>
<tr>
<td>$\delta(\cdot)$</td>
<td>Dirac function</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Image source localisation error</td>
</tr>
<tr>
<td>$\epsilon_{\text{ref}}$</td>
<td>Reflector localisation error calculated for a single point</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>Tangency coefficient</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Total air density</td>
</tr>
<tr>
<td>$\delta \eta$</td>
<td>Variable part of the air density</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>Static part of the air density</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>Azimuth direction of arrival</td>
</tr>
<tr>
<td>$\iota$</td>
<td>Marginal class membership</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Wavelength</td>
</tr>
<tr>
<td>$\mu_{\text{ILD}}$</td>
<td>Mean value of the interaural level difference model</td>
</tr>
<tr>
<td>$\mu_{\text{IPD}}$</td>
<td>Mean value of the interaural phase difference model</td>
</tr>
<tr>
<td>$\mu_{\text{RMSE}}$</td>
<td>Mean value of the root mean square errors for the reflector localisers</td>
</tr>
<tr>
<td>$\mu^\epsilon$</td>
<td>Mean value of the image source localisation errors</td>
</tr>
<tr>
<td>$\nu$</td>
<td>Auxiliary function for the expectation-maximisation algorithm</td>
</tr>
<tr>
<td>$\xi$</td>
<td>Incidence angle</td>
</tr>
<tr>
<td>$\Xi$</td>
<td>Vector containing the normal Gaussian mixture parameters</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Pi number</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Radial distance</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
<td>Estimated radial distance</td>
</tr>
<tr>
<td>$\bar{\rho}$</td>
<td>Averaged radial distance</td>
</tr>
<tr>
<td>$\rho'$</td>
<td>Altered radial distance</td>
</tr>
<tr>
<td>$\sigma_{\text{ILD}^2}$</td>
<td>Variance of the interaural level difference model</td>
</tr>
<tr>
<td>$\sigma_{\text{IPD}^2}$</td>
<td>Variance of the interaural phase difference model</td>
</tr>
</tbody>
</table>
Symbols

$\Sigma$  Diagonal matrix containing the Fisher information

$\tau$  Time difference of arrival

$\tau_A$  Threshold for the room impulse response time domain amplitude

$\tau_s$  Threshold for the slope function $S_{gd}(\cdot)$

$\tau_t$  Threshold for the tangency coefficient $\zeta$

$\Upsilon$  Number of tested mixtures

$\Phi$  Elevation direction of arrival

$\Phi^{\text{IPD}}(\cdot)$  Interaural phase difference of the observation

$\hat{\Phi}^{\text{IPD}}(\cdot)$  Predicted interaural phase difference

$\psi$  Critical value for the confidence interval

$\omega$  Angular frequency

$\Omega$  Vector containing the log-likelihood parameters

$\Omega_{\text{est}}$  Vector containing the estimated log-likelihood parameters

$\ast$  Convolution operator

$[\cdot]^*$  Adjoint of a matrix operator

$[\cdot]^T$  Transpose of a matrix operator

$\|\cdot\|$  Euclidean norm operator

$|\cdot|$  Absolute value operator

$(\cdot)!$  Factorial operator

$\partial(\cdot)$  Partial differential operator

$\nabla$  Gradient operator

$\nabla \cdot$  Divergence operator

$\nabla^2$  Laplacian operator

$\Sigma(\cdot)$  Summation operator

$\mathbb{N}$  Natural number set

$\mathcal{N}(\cdot)$  Normal distribution

$\mathbb{P}^3$  3D projective space

$\mathbb{R}^3$  3D Euclidean space

$\Re(\cdot)$  Real part operator
Alla mia famiglia, per l’enorme supporto durante questi anni.

To my family, for the tremendous support during these years.
Chapter 1

Introduction

Sound is classically defined as the auditory perception caused by rapid variations of the medium pressure. From a physical point of view, these oscillations are usually approximated by defining wave equations. Like every other physical model based on waves, the sound interacts with the natural and artificial obstacles that encounters during its propagation. These interactions generate additional waves, i.e. reflections, that can either support or modify the perception of the main sound. Therefore, reflections can be defined as either supportive or interfering, respectively. Considering static listener and sound source within an enclosed environment, these reflections produce a unique perception of the sound.

The human auditory system is able to understand and classify the differences among sound perceptions. Converting the sound pressure to an electric signal first, the central nervous system is then able to process it. The output of this process allows humans to accomplish different tasks, for instance: to localise target sources; to separate them from background noise; to understand the size of the environment; to detect the proximity to acoustic reflectors. Features that are taken into account by the human system must be considered, to develop signal processing methods that are able to exploit information related to supportive reflections and suppress the contribution given by interfering ones. This thesis contributes to the literature by presenting four novel acoustic reflector localisation techniques. Moreover, it also contributes to two of the major audio signal processing fields, i.e. spatial audio and source separation. Their respective state-of-the-art is improved, by proposing novel methods that incorporate information about the acoustical reflections. This information is provided by the novel acoustic reflector localisation solutions.

This chapter is composed by different sections, having different aims. Firstly, the motivation that lead to study the topic of this thesis will be underlined. Then, the three
main problems that have been tackled will be stated. Finally, the actual contributions, together with the outline of the thesis, will be presented.

1.1 Motivation

Supportive and interfering reflection information can be utilised to improve several areas in the audio signal processing research field. Strong reflections that arrive early in time can be identified as supportive reflections, since they are usually high correlated to the produced sound. On the other hand, later reflections may loose this correlation, hence, they can be considered as interfering reflections. The opportunity of defining a relationship between a sound and the reflections produced during its propagation mainly motivated this thesis.

The main process of defining a relationship between an audio signal and the environment within it is reproduced falls into the research area named as virtual acoustic environment modelling. It is composed of three main tasks [Savioja et al., 1999]: the source modelling [Coleman et al., 2014a] (including research areas such as natural audio, synthetic audio, source directivity), the room modelling [Lee et al., 2012] (e.g. modelling of acoustic spaces or artificial reverberation) and the listener modelling [Masterson et al., 2012] (e.g. study of head-related transfer functions or microphone directivity).

The focus of this thesis is centred on the room modelling part. During the last decade, several research groups focused their attention on this topic. Room geometry estimation, i.e. the localisation of acoustic reflectors in a room, was mainly investigated by exploiting information extracted from recorded room impulse responses (RIRs). A RIR is a particular acoustic signal which relates the listener position to the source position, within a specific environment. Acoustic reflector localisation, given RIRs, is also the main focus of this thesis. This choice was motivated by the fact that, by localising the main reflector positions, the related acoustic reflection can be interpreted. This is demonstrated, later in the thesis, by employing the acoustic reflector localisation solutions to create novel spatial audio and source separation models.

In general, creating an accurate model to identify acoustic reflector positions from RIRs is important for several research areas. For instance, as already mentioned, such a model can provide information about the acoustic of the listening environment with respect to a listener position. This information can be exploited in audio forensics, e.g. to verify the authenticity of digital evidence [Malik, 2013]; in simultaneous localisation and mapping, e.g. to let a robot understanding its own position [Dokmanić et al., 2016]; in spatial audio, e.g. to estimate the acoustical parameters that are necessary to reproduce
1. Introduction

a sound, maintaining the acoustical characteristics of the recording environment [Herre et al., 2015]. In addition, knowledge of the reflector positions can be utilised to enhance target signals, improving probabilistic models, applied in fields such as automatic speech recognition [Yoshioka et al., 2012] and music transcription [Benetos et al., 2013]. It can also aid source separation techniques, for instance, by recovering the acoustic channels [Asaei et al., 2014]; audio tracking models by enhancing the estimation of the binaural cues [Liu et al., 2015a]; dereverberation, e.g. by improving applied beamforming techniques [Habets and Benesty, 2013]; personal sound zones reproduction by improving the loudspeaker array directivity [Coleman et al., 2014b]. Beyond all of these, an acoustic reflection localiser can be combined with image processing to construct a robust hybrid room geometry model [Hussain et al., 2014, Ye et al., 2015]: acoustic information can aid detection of mirrors and windows, which cannot be correctly analysed by only visual sensors.

1.2 Problem Statements and Applications

1.2.1 Acoustic Reflector Localisation

As it has been already mentioned upon, the main focus of this thesis is the creation of mathematical models, able to localise acoustic reflectors. This kind of research area falls into the room modelling task, which is a component of the more general virtual acoustic environment modelling.

The main assumption made by this thesis is to have availability of RIRs, recorded in the environment under investigation. Several definitions can be found in literature, describing a RIR. In general, a RIR is an acoustic signal that, convolved with the sound generated at the source, produces the signal perceived at the listener’s position. Typically, researchers agree on the fact that a RIR is composed by three main components: the direct sound, the early reflections, and the late reverberation [Kuttruff, 2009]. The late reverberation (i.e. the characteristic decaying “tail”) provides the listener with the perception of many remote reflections from distant surfaces, whereas early reflections are the sonic manifestation of nearby objects (e.g. walls) [Blesser, 2001]. In other words, while the direct sound reveals the direction of the audio source, the early reflections convey a sense of the geometry, whereas the late reverberation is indicative of the size of the environment [Välimäki et al., 2012]. Therefore, direct sound and early reflections can be defined as providers of the main information related to the recording setup.

In the literature, several issues related to the room modelling are usually investigated, and they can be summarised through four main research questions:
1.2.2 Spatial Audio

- Is it possible to extract the reflector information from recorded audio signals?
- How to detect reflections given recorded audio signals?
- How to identify reflections recorded by different microphones as coming from the same acoustic reflector?
- What is the best approach to localise reflectors once the related information has been extracted?

The assumption of having recorded RIRs allows to answer the first question. In fact, it opens to the possibility of developing algorithms which can directly extract information related to the room acoustics. To answer the second question, a novel algorithm to detect direct sound and early reflections given RIRs is proposed, later in this thesis (see Chapter 4). It is named as Clustered Dynamic Programming Projected Phase-Slope Algorithm (C-DYPSA), and it is an evolution of the state-of-the-art DYPSA algorithm [Naylor et al., 2007]. C-DYPSA is developed to exploit the additional information given by the availability of microphone arrays. Regarding the third question, the extraction from RIRs of the information related to a specific reflection, typically leads to an issue named as permutation problem. It describes the problem of labelling reflections detected at different microphone positions, but arriving from the same reflector. In fact, some reflections can arrive at different microphones with a different order. Some researchers tried to solve this problem by exploiting matrix operations based on spatial Euclidean distances [Dokmanić et al., 2013]. However, current technologies give the opportunity of creating small, compact arrays of microphones. Therefore, in this thesis, it is assumed the use of one of these arrays. In this way, being the microphones spatially compact, reflections are detected with the same order at every microphone. Finally, a chapter of this thesis (i.e. Chapter 4) will answer the fourth question. There, novel methods following two different conceptual approaches will be presented. The state-of-the-art methods estimating reflector positions in the 3D space will be classified within to main groups, the image source reversion methods and the direct localisation methods. This classification will be also utilised in support of the literature review, that will be reported in Chapter 3.

1.2.2 Spatial Audio

In this thesis, spatial audio is one of the two audio signal processing areas, where the proposed acoustic reflector estimators are applied. The fundamental purpose is to artificially generate or manipulate acoustic environments. Having 3D sound systems available, four main virtual acoustic simulations can be produced for the listener [Begault,
1. Introduction

[2000]: an existing spatial auditory condition can be replicated; a completely unknown spatial auditory experiences can be reproduced; a spatial experience can be transmuted to another; a virtual auditory world can be generated.

The process of characterising enclosed environments through synthetic RIRs, and convolving them with a dry sound\(^1\), is known as auralisation [Hulsebos, 2004]. During the last decades, one of the main focuses in spatial audio has been the creation of new techniques, able to enhance the quality of both capturing and reproducing sounds. From the production side, the current aim is to convey to the listener a sense of being within a specific environment, that can be real, or generated by the imagination of the story writer [Kares and Larcher, 2016]. This distinction between real and imaginary acoustical scenes is important, since different applications can be attributed to them. The purpose of reproducing recording environments usually belongs to cinema production, where spatial audio is used to transport the audience within the story of the projected film [Rumsey, 2001]. Virtual environment production is recently gaining quite of attention [Ohta and Tamura, 2014]. For instance, platforms for the distribution of virtual reality (VR) contents are becoming widely available. Also broadcasters are nowadays focused on studying the VR experience. The BBC Research & Development group has recently created a compelling immersive story using VR, with a strong focus on spatial audio [Pike et al., 2016]. Augmented reality (AR) is another similar application. In contrast to VR environments, where participants are abstracted from the real world, in an AR environment participants interact with virtual elements that are mixed with reality. Spatial audio can be employed to produce this interaction more realistic, for instance, by creating virtual sound objects in the real space, that can be localised by the users [Vazquez-Alvarez et al., 2012].

In conclusion, new methods are currently needed for capturing, producing and reproducing the environmental acoustics. For instance, a final consumer may be interested in having tools to edit their own environmental perception. At the same time, sound engineers may employ these tools to create new acoustic sensations. Within this thesis, a solution is presented to make the acoustic editable, efficient to transmit, and reproducible on different reproduction systems (see Chapter 5).

1.2.3 Source Separation

Source separation is one of the most investigated areas in signal processing, due to its wide range of possible applications. For instance, in biomedical signal processing, it is usually exploited to analyse electrocardiograms, electroencephalograms, or magnetic

\(^1\)A dry sound carries no spatial information. It can be considered as similar to a sound recorded within an anechoic environment.
1.2.3. Source Separation

resonance imaging [Ungureanu et al., 2004]. Generalising the input data, source separation methods may be applied to completely different areas, e.g. helping with the restoration of ancient documents [Tonazzini et al., 2007].

In this thesis, only audio signals are investigated. Acoustic signals recorded in real world scenarios are usually corrupted by additive noise and other sound sources. The principal aim of audio source separation is to characterise the target sound source, given a mixture of sounds, in order to separate it from the interference. This research field can be subdivided into two main branches, depending on the type of signal involved: speech and music [Vincent et al., 2012].

Considering musical signals, source separation is usually described as the ability of separating the different song components, belonging to different instruments or voices. Identification of instruments is, for instance, one of the possible application areas. In that scenario, songs can be classified by considering the instruments that are involved. Useful tools can be created for user searches, for example, by categorising musical genres [Kitaehara et al., 2007]. Another possible application may extract a sequence of frequency values, that corresponds to the pitch of the dominant melody [Salamon et al., 2014]. This process is known as melody extraction, and it is related to automatic music transcriptions. Another way of separating the instruments is done by assuming the musical signals to be sparse in time. Notes can be then associated to the separated instruments and be automatically transcribed [Plumbley et al., 2010]. Repeating structures can also be exploited to separate vocal signals, useful for karaoke gaming track production [Rafii and Pardo, 2013].

Speech source separation is of interest for this thesis. Several applications can be identified for this field. Hearing aids is one of them, where source separation methods are usually applied to provide the hearing-impaired user with higher speech intelligibility. This is done by reducing noise artefacts, while preserving the binaural cues related to the expected target signal, the residual noise and the interferers [Doclo et al., 2015]. As for music transcription, automatic speech recognition is widely investigated, by firstly applying source separation techniques. Automatic speech recognition is usually defined as the process of converting speech signals into their corresponding sequence of words or other linguistic entities [Li et al., 2014]. Therefore, clean signals are required in input, without any additive noise or interfering sources. Creating a connection between source separation and spatial audio, sound sources can be localised and recorded from

\[ \text{The sparsity property defines the characteristic of the signals in a mixture of not being overlapped considering a specific domain (e.g. time, space, or frequency).} \]
1. Introduction

a real environment by applying beamforming\(^3\) algorithms [Coleman et al., 2015a]. Security surveillance systems can also take advantage of source separation methods. For instance, considering defence purposes, they can be used to monitor certain areas, e.g. developing passive sonar systems [Sutin et al., 2010]. Another interesting application can be found within the robotic area. Interaction between men and humanoids is aided by utilising source separation methods, that, reducing the interference, allows robots to easier respond to issued commands [Deleforge and Horaud, 2012].

In this thesis, a novel source separation method is presented, exploiting the proposed reflector localisation method solutions. In particular, in Chapter 6, a model incorporating early reflection cues is proposed. This is an extension to the previous model which was presented in [Mandel et al., 2010]. The results underline the importance of the environmental knowledge, together with the source characterisation, for source separation purposes.

1.3 Contributions and Thesis Outline

1.3.1 Chapter 2: Conceptual Foundations

In Chapter 2, mathematical and physical concepts will be presented, which support the methodologies employed within the rest of the thesis. This will be an important passage, that will allow the reader to understand the paths undertaken during this thesis development. The first concept that will be treated is the sound wave propagation in air. The integral sound wave equation will be provided together with its solution. Furthermore, the interaction of the sound wave with objects encountered during its propagation will be described, showing how a reflection physically happens. The two main models which will be utilised later in the thesis will be then described: the image source model and the RIR model. Finally, a description about the human perception of the first early reflection is reported, together with those general concepts that characterise the acoustic perception of an environment.

1.3.2 Chapter 3: Reflector Localisation and its Applications: a Literature Review

The state-of-the-art of the three main contributions of this thesis will be described within Chapter 3. Precisely, the current literature related to the acoustic reflector localisation, spatial audio, and blind source separation will be reported. Furthermore,

\(^3\)Beamformers are particular source separation algorithms which rely on the spatial sparsity of the sources. They will be better describing in Chapter 4
established algorithms and general tools that will be utilised during the rest of the thesis will be described. Typical algorithms used to estimate the times of arrival (TOAs) of the reflections will be described, together with methods to estimate the directions of arrival (DOAs) and the frequency spectra. Moreover, coordinate systems and geometrical transformations will be presented.

1.3.3 Chapter 4: Acoustic Reflector Localisation Methods

Chapter 4 will introduce the core of this thesis, which is represented by four novel methods for acoustic reflector localisation. Firstly, some of the most influential state-of-the-art works will be categorised between two groups depending on the approach employed: image source reversion and direct localisation. After this, two from these state-of-the-art methods will be selected and presented in detail. The novel methods will be then proposed, dividing them between the two predefined groups. The performance of the reflector localisation will be reported, comparing the proposed methods to the selected state-of-the-art.

The main contributions that will be found within this chapter are simplified using five bullet points:

• An epoch detector named C-DYPSA, employed to calculate the TOAs of each sound wavefront, exploiting multichannel arrays of microphones;

• A method belonging to the image source reversion category, named ISDAR-LIB. It is the fusion between the novel image source localisation method, i.e. the image source direction and ranging (ISDAR), and a reflector localiser algorithm, named loudspeaker image bisection (LIB), that was firstly presented in [Tervo and Tossavainen, 2012];

• Two variants of the ISDAR-LIB method, which utilise the information provided by multiple loudspeakers;

• A method belonging to the direct localisation category, that is named as ellipsoid tangent sample consensus (ETSAC);

• The first comparative evaluation performed among different acoustic reflector localisation methods, utilising both simulated and measured data.

As already discussed, the contributions presented in this chapter represent the core of this thesis. The author of this thesis already published three conference papers and one journal article about them:
1. Introduction


1.3.4 Chapter 5: Parametrisation of Reverberant Spatial Audio Objects

The spatial audio application, to the acoustic reflector localisation solutions presented in Chapter 4, will be presented in Chapter 5. Recently, it was proposed an efficient way of describing room acoustics, both in terms of flexibility and transmission ease: the recorded sounds are treated as spatial audio objects (SAOs) [Herre et al., 2014]. Their coding standard was defined in the moving picture experts group (MPEG)-H standard as spatial audio object coding (SAOC) [Herre et al., 2012].

A novel way of encoding and decoding the parameters characterising SAOs will be proposed in this chapter. Firstly, the encoding part will be introduced, by defining parameters that characterise the SAOs in both frequency and time domain. The specific algorithms to estimate these parameters from RIRs will be also presented. The proposed parametrised sounds will be defined as reverberant spatial audio objects (RSAOs). It will be shown that the ISDAR method, that is presented in Chapter 4, can be employed to localise both sources and image sources. The correspondent decoding part will be also proposed, together with the algorithm that renders the sounds to the available loudspeaker configuration. A major experiment will be presented as a set of subjective assessments. The aim was to evaluate the system, by understanding, from the perceptual point of view, its performance on generating sounds with specific source distances, environmental sizes, and listener envelopment.
With respect to the literature, this chapter proposes three main contributions, that can be summarised as:

- The direct sound and early reflection TOAs are estimated and then used to estimate the related DOAs, by segmenting the RIRs through time windowing techniques;

- The frequency domain of early reflections and direct sound are analysed and parametrised, to provide the rendered sound with a better *colouration*;

- The reverberation part is analysed and parametrised, to be then recreated at the decoding stage with similar frequency characteristics.

The author of this thesis already published two conference papers about RSAOs. In addition, other three conference papers and a journal article were published as result of the cooperative work within the S3A project\(^4\). The list of these publications is as follows:

- L. Remaggi, P. J. B. Jackson and P. Coleman, “Estimation of room reflection parameters for a reverberant spatial audio object”, *138th Audio Engineering Society Convention (AES)*, Warsaw, Poland, 2015 [Remaggi et al., 2015a].


\(^4\)http://www.s3a-spatialaudio.org/
1. Introduction

1.3.5 Chapter 6: Source Separation through Multipath Propagation Analysis

A novel source separation method will be presented in Chapter 6. Binaural audio systems, such as the human auditory system, mainly rely on two cues to localise sound sources: the interaural level difference (ILD) and interaural phase difference (IPD). In [Mandel et al., 2010], both these cues were modelled, to perform source localisation and separation given sound mixtures. However, only the direct sound between source and listener was taken into account. Here, a modified version of that model will be presented, where also the first strong reflection will be included in the IPD model. Probability functions will be defined, together with the source separation problem resolution reached by employing a Gaussian mixture model (GMM) and expectation-maximisation (EM). An initialisation method for EM, i.e. the ISDAR algorithm presented in Chapter 4 will be also described. Finally, experimental results will be shown, to evaluate the novel method against the baseline [Mandel et al., 2010].

In this chapter, two are the main contributions that can be listed as:

- For the first time, a source separation method includes, in its analytic model, knowledge of the early reflections;
- ISDAR is used to initialise the EM algorithm, localising source and image sources.

1.3.6 Chapter 7: Conclusion

This final chapter will draw the overall conclusion, discussing every novel approach introduced by this thesis. For each of the three main contribution chapters, the structure of the discussion will composed by three parts. The first part will briefly summarise the technical description of the method proposed. The second part will state the findings, describing results that were obtained carrying out simulations and experiments. Finally, the third part will analyse possible opportunities for future work.

1.3.7 Appendixes

Appendix A: Multichannel RIR Visualisation. In this appendix, some techniques will be presented, that are able to visualise information about reflection positions. This information is assumed to be carried by multichannel RIRs. These methods were proposed through an Engineering brief (Ebrief), at the 139th AES Convention [Remaggi
et al., 2015c]. Given multichannel RIRs that were recorded in different environments, sound waves, reflections and reflector positions will be visualised.

The choice of including these visualisation techniques within an appendix was made since they are not recognised to be contributions matching the overall flow of the thesis. However, they are also deemed to be important results to show, since they validate the reflection localisation, that is the main objective for this thesis.

Appendix B: The CISPE Algorithm. The cross-correlation based iterative sensor position estimation (CISPE) algorithm will be presented in this appendix. It is an iterative algorithm utilised to estimate the microphone positions given a rough initialisation and multichannel RIR recordings. This algorithm will be presented in this appendix after providing a general background on the microphone calibration literature. Source localisation and reflector localisation methods (similar to the ones presented in Chapter 4) will be then illustrated, before showing some experimental results.

As for Appendix A, it could have been misleading for the reader, to include CISPE within the main flow of this thesis. However, CISPE represents a valuable contribution to the literature. A related conference paper was published at the 22nd ICSV Conference [Remaggi et al., 2015b].

---

5This publication is already listed among the bullet points in Section 1.3.4
6This publication is already listed among the bullet points in Section 1.3.3
Chapter 2

Conceptual Foundations

This chapter focuses on presenting the physical and mathematical foundations of the work proposed by this thesis. Starting from defining the acoustical wave equation and its solution, the theoretical concepts behind sound reflections will be utilised to describe the sound propagation within enclosed environments. A mathematical approach will be developed to formulate the so called image source model [Allen and Berkley, 1979] and the room impulse response (RIR) [Kuttruff, 2009], by firstly defining the Green’s function. Finally, a description of the way the human auditory system perceives the sound will be reported. In particular, the influence of the first early reflection on the sound perception will be elaborated.

2.1 Sound Wave Propagation

A physical event perceived by humans through their auditory system is usually named as sound. It is characterised by oscillations of the air particles, which produce local variations on the atmospheric pressure. Like any other complex field, the sound field can be modelled as superimposition of several waves (a visualisation of propagating sound waves is depicted in Figure 2.1). These waves can be approximated as either plane or spherical waves [Kuttruff, 2009]. Although sound can propagate through any kind of medium, in this chapter propagation in gas (e.g. air) is studied, since this is the medium assumed during the rest of the thesis. Under the assumption of the air being homogeneous and at rest, the speed of sound can be defined as a constant, in both space and time. Its magnitude is given by:

\[ c_0 = 331.4 + 0.6Q, \]  

(2.1)
2.1.1 Acoustical Wave Equation

As already touched upon, a sound wave modifies the air particle locations, vibrating them from their mean position. For this reason, the instantaneous displacements of these particles describe the wave behaviour. These displacements are strictly related to \( \mathbf{v} = [v_x, v_y, v_z] \), that is the velocity of the particle vibrations, with \( v_x, v_y, \) and \( v_z \) being the three components in the Cartesian coordinate system. \( \mathbf{v} \) is considered as one of the fundamental acoustical quantities. This movement of particles, generated by sound waves, also produces variations from the static air density \( \eta_0 \) and the air pressure \( P \). Two are the main physical laws that regulate these phenomena: the conservation of momentum and the conservation of mass. Their mathematical expressions can be combined together to obtain the acoustical wave equation.

The conservation of momentum law describes the fact that, in a system that does not exchange neither energy nor mass with the surrounding environment, the total momentum is constant. This law can be derived starting from Newton’s second law, stating that a force \( F \), applied to an object, is proportional to the applied acceleration \( a \) and

where \( Q \) is the gas temperature in degree Centigrade. The unit of measure of \( c_0 \) is \( \text{m} \cdot \text{s}^{-1} \). For the rest of this section, most of the equations are reproduced or modified from [Kuttruff, 2009] and [Morse and Ingard, 1968].

**Figure 2.1:** Space-time visualisation of sound waves, recorded within three different rooms, by utilising a circular array of 24 microphones. The sound pressure recorded at each microphone is here plotted, vertically adjacent to one another in the time domain. The first wave corresponds to the direct sound coming from the loudspeaker, whereas the later ones are reflections coming from the room boundaries. The methodology employed to generate this kind of images is described in Appendix A. For a better visualisation, the sound pressure between adjacent microphones has been interpolated.
the object mass \( m_a \):

\[
F = m_a a = m_a \frac{\partial v}{\partial t},
\]

where \( t \) is the continuous time variable, and \( \partial (\cdot) \) is the partial differential operator. Considering a unit of spatial volume \( V \), \( F \) is related to differences of \( P \) at the opposite sides of \( V \). Therefore, the force equation can be written as:

\[
F = \nabla P,
\]

where “\( \nabla \)” is the gradient operator. Finally, following the Newton’s third law, which states that a force applied to an object corresponds to an equal force produced by the object having same intensity and opposite direction, the two force definitions, in Equations 2.2 and 2.3, are combined to obtain:

\[
\nabla P = -m_a \frac{\partial v}{\partial t} = -\eta_0 \frac{\partial v}{\partial t}.
\]

\( m_a \) was substituted with the density per unit volume \( \eta_0 \), since the unit of volume \( V \) is considered as reference. Equation 2.4 defines the conservation of momentum law.

On the other hand, considering the same unit of volume \( V \), the conservation of mass law states that in a system that does not exchange neither energy nor mass with the surrounding environment, the mass remains constant. In other words, the variation of mass inside \( V \) is equal to the amount of mass passing through its volumetric surface [Morse and Ingard, 1968]:

\[
- \frac{\partial}{\partial t} \iiint \eta \, dV = \iint \eta_0 (v \mathbf{u}_s) \, dS,
\]

where \( \eta = \eta_0 + \delta \eta \) is the total density within the volume, and it is composed by the static density \( \eta_0 \) and its variable part \( \delta \eta \); the left integral is a volumetric integral, whereas the right one is a surface integral, that is calculated considering the volumetric surface \( S \); \( \mathbf{u}_s \) is a vector that is normal to \( S \). Employing the divergence theorem [Arfken, 1985], it is possible to transform Equation 2.5 into:

\[
- \frac{\partial}{\partial t} \iiint \eta \, dV = \iiint \nabla \cdot (\eta_0 \mathbf{v}) \, dV,
\]

where “\( \nabla \cdot \)” is the divergence operator. Reorganising the equation into:

\[
\iiint \left( \frac{\partial \eta}{\partial t} + \nabla \cdot (\eta_0 \mathbf{v}) \right) \, dV = 0,
\]
and considering the fact that $\eta_0$ is a constant, the conservation of mass law can be finally written as:

$$\eta_0 \nabla \cdot v = -\frac{\partial \eta}{\partial t}. \quad (2.8)$$

By dividing the momentum conservation law in Equation 2.4 by $\eta_0$, and differentiating both the members with respect to the space, it becomes:

$$\frac{\partial}{\partial t} \nabla \cdot v = -\frac{1}{\eta_0} \nabla^2 P, \quad (2.9)$$

where $\nabla^2$ is the Laplacian operator. By then considering the mass conservation law in Equation 2.8, dividing both its members by $\eta_0$, and differentiating them with respect to the time, it results:

$$\frac{\partial}{\partial t} \nabla \cdot v = -\frac{1}{\eta_0} \frac{\partial^2 \eta}{\partial t^2}. \quad (2.10)$$

Finally, Equations 2.9 and 2.10 are combined, to obtain the acoustic wave equation:

$$\frac{\partial^2 P}{\partial t^2} = c_0^2 \nabla^2 P, \quad (2.11)$$

where it is assumed the linear material law (i.e. $P = c_0^2 \eta$).

### 2.1.2 Solution to the Wave Equation

Observing the momentum conservation law in Equation 2.4, it can be noted that sound wave velocity in fluids is parallel to the pressure gradient. This means that, in general, sound waves in fluids are *longitudinal waves* [Kuttruff, 2009]. In other words, the acoustical quantities depend only on the time and direction of propagation. Therefore, by defining the $x$-axis of a Cartesian coordinate system as the propagation direction, the acoustic wave equation can be simplified as:

$$\frac{\partial^2 P}{\partial t^2} = c_0^2 \frac{\partial^2 P}{\partial x^2}. \quad (2.12)$$

The general solution to this differential equation was firstly computed by d’Alambert [Bekefi and Barrett, 1987]:

$$P(x, t) = d^+(c_0 t - x) + d^-(c_0 t + x), \quad (2.13)$$

where $d^+(\cdot)$ and $d^-(\cdot)$ are arbitrary functions, having finite second derivative. $d^+(\cdot)$ represent the wave propagation towards the positive direction of the $x$-axis. Therefore, if a time increase $\delta t$ is evaluated together with a coordinate increase $\delta x = c_0 \delta t$, the value of $d^+(\cdot)$ remains unaltered. The same consideration, but with opposite values of $\delta t$ and
2. Conceptual Foundations

Figure 2.2: Visualisation of the sound propagation. Since the sensor (i.e., a microphone) is drawn in the far field, the incoming waves can be approximated as plane waves.

\( \delta x \), can be done for \( d^-() \), that, hence, describes a pressure wave propagating towards the negative direction of the \( x \)-axis.

As shown in Figure 2.2, the sound waves propagate around the source following a spherical pattern. However, considering the scenario where the sensor is far enough from the source, the sound waves can be approximated as plane waves, as done by Equation 2.13. This scenario is known as far field. In the far field, the sound pressure \( P \) is considered to be constant on any plane that is perpendicular to the direction of propagation [Kuttruff, 2009]. These planes are defined as wavefronts. The spatial limit that separates the far field from the near field condition depends on the signal wave length (thus on the frequency), and it is defined as \( d_\text{ff} = 2d_{\text{mm}}^2 \lambda \) [Balanis, 2005], where \( d_{\text{mm}} \) is the microphone array aperture, and \( \lambda \) is the wavelength. This equation is also known as Fraunhofer rule, and it was first proposed considering radio waves [Rappaport, 2002].

Since sound waves are longitudinal, the solution to the wave equation, along the \( x \)-axis, can be also expressed in terms of particle velocity \( v_x \) as:

\[
v_x(t) = \frac{1}{Z_0} \left[ d^+(c_0t - x) - d^-(c_0t + x) \right], \tag{2.14}
\]

where \( Z_0 = \eta_0 c_0 \) is the characteristic impedance of the medium.

A sound wave propagates alternating, in time, high and low air pressures, at a specific point in space, as shown in Figure 2.3. Due to the periodic nature of this alternation, sound waves are commonly defined as superposition of harmonics. Setting \( d^-() = 0 \), Equation 2.13 can be expressed as a planar, progressive harmonic wave:

\[
P(x, t) = \hat{P} \cos(\omega t - kx), \tag{2.15}
\]

where \( \hat{P} \) is a constant defining the magnitude of the pressure, and \( k \) is a constant defined as angular wavenumber. \( \omega = kc_0 \) is the angular frequency and it is related to the
The temporal period $T$ as $T = 2\pi/\omega$, with $\pi$ representing the “pi” number. Having defined these harmonic constants, the wavelength, i.e. the spatial distance between two points having the same value of pressure, can be then written as:

$$\lambda = \frac{2\pi}{k} = \frac{2\pi c_0}{\omega} = \frac{c_0}{f}, \quad (2.16)$$

where $f$ is the vibration frequency.

Complex number properties can be then utilised to represent harmonic oscillations. Defining $j = \sqrt{-1}$, the complex notation of harmonic vibrations is written as:

$$\exp(jx) = \cos x + j \sin x. \quad (2.17)$$

Therefore, Equation 2.15 can be also defined as:

$$P(x, t) = \hat{P} \exp[j(\omega t - kx)]. \quad (2.18)$$

The harmonic wave that has been derived assumes the medium as being free of losses. However, in the real world, the pressure amplitude does not remain constant for the whole propagation, instead, it decreases following an exponential law. In order to take this into account, Equation 2.18 is modified as:

$$P(x, t) = \hat{P} \exp \left( -\frac{\alpha_m x}{2} \right) \exp[j(\omega t - kx)], \quad (2.19)$$
where $\alpha_m$ is the attenuation coefficient of the medium.

2.2 Acoustic Reflection

The acoustic environments that are of main interest for this thesis are indoor. Therefore, although, for simplicity, in Section 2.1, the wave equation and its solution were derived assuming unbounded medium, in this section, conceptual models are presented, which describe the interaction between propagating sound and acoustic reflectors. However, only the interaction with one reflector will be discussed. The transformation applied to a sound wave, that sequentially interacts with multiple reflectors, will be described later, in Section 2.3.

The main focus of this section is to present the so called specular reflections. These are reflections produced by a sound bouncing off smooth unbounded walls. However, other type of reflections also exist. For instance, a wall edge will diffract the sound, whereas uneven surfaces will scatter the sound towards uneven directions. Whilst they are not going to be considered during the rest of this thesis\(^1\), their presence is important, thus, they will be briefly introduced. To be able to approximate the sound waves as planar, reflectors are assumed to lie in the far field. In addition, phenomena related to the interaction between sound waves and listener’s head will be described, since they represent the fundamental theory behind the source separation model in Chapter 6.

Most of the equations presented in this section are reproduced from [Kuttruff, 2009, Morse and Ingard, 1968] and [Cox and D’Antonio, 2016].

2.2.1 Specular Reflection

Usually, acoustic reflectors do not reflect the whole amount of energy carried by the incident sound wave $\mathbf{v}^{\text{inc}}$; instead, part of it is absorbed. Following the energy conservation law, this absorbed energy is either transferred to an environment that is external to the one under consideration, or it is transformed to heat. This causes the reflected component $\mathbf{v}^{\text{ref}}$ to be attenuated, and to also have a different phase with respect to $\mathbf{v}^{\text{inc}}$.

Four acoustic quantities are generally utilised to define the acoustical properties of a reflector: the impedance $Z$, the admittance $Y$, the pressure reflector factor $R$, and the absorption coefficient $\alpha_p$. Whereas $Z$, $Y$ and $R$ provide information related to both phase and magnitude changes, $\alpha_p$ provides information about magnitude changes only.

\(^1\)To be more precise, although the specular reflections are assumed as predominant to localise reflectors in Chapters 4, 5 and 6, scattering and diffraction are implicitly considered in Chapter 5, where the late reverberation part of the RIR is analysed (RIRs will be defined in Section 2.3.3).
2.2.1. Specular Reflection

Assuming a reflector to be unbounded and smooth, then the related reflection can be assumed to be always specular. In other words, the angle $\xi$ between the incident wave and the vector $\mathbf{u}$, that is normal to the reflector, is equal to the angle between the reflected wave and $\mathbf{u}$. Let us consider a sound wave propagating on the $x$-$y$ plane of a Cartesian coordinate system, and incident, with angle $\xi$, to a planar reflector lying at $x = 0$, as depicted in Figure 2.4. It is possible to transform the coordinate system, by rotating the axis by $\xi$. In particular, it is possible to substitute the spatial variable $x$ with the new variable $x'$, in order to make $\mathbf{v}^{\text{inc}}$ propagating towards the positive direction of $x'$. The new axis is given by:

$$x' = x \cos \xi + y \sin \xi. \quad (2.20)$$

Therefore, by substituting $x$ with $x'$ in Equation 2.18, the pressure related to the incident plane wave is defined as:

$$P^{\text{inc}}(x, t) = \hat{P} \exp[j \omega t - jk(x \cos \xi + y \sin \xi)]. \quad (2.21)$$

To calculate the acoustical quantities of the reflector, the velocity component, $v_x$, that is normal to it is required. The momentum conservation law (in Equation 2.4) can then be written as:

$$v_x = -\frac{1}{j \omega \eta_0} \frac{\partial P}{\partial x}. \quad (2.22)$$

Therefore, from Equation 2.21, it follows that the incident component to the normal velocity vector is given by:

$$v_x^{\text{inc}} = \frac{\hat{P}}{Z_0} \cos \xi \exp[j \omega t - jk(x \cos \xi + y \sin \xi)]. \quad (2.23)$$

Regarding the reflected wave, the signs of both pressure and velocity are opposite to their incident equations. This is due to the fact that the reflection travels towards the opposite direction, with respect to $x$. In addition, the reflector absorption process...
influences the equations by a factor \( R \), that is the pressure reflector factor. Pressure and velocity equations for the reflected sound wave are, respectively:

\[
P_{\text{ref}}(x,t) = R \hat{P} \exp[j\omega t - jk(-x \cos \xi + y \sin \xi)],
\]

\[
v_{x,\text{ref}} = -\frac{R \hat{P}}{Z_0} \cos \xi \exp[j\omega t - jk(-x \cos \xi + y \sin \xi)].
\]

The reflector acoustic impedance \( Z \) is the ratio between the total acoustic pressure at the reflector position and the normal velocity to the reflector. Thus, by setting \( x = 0 \), and dividing \( P_{\text{inc}}(x = 0, t) + P_{\text{ref}}(x = 0, t) \) by \( v_{x,\text{inc}} + v_{x,\text{ref}} \), \( Z \) is obtained as:

\[
Z = \frac{Z_0}{1 + R} \cos \xi \left( 1 - \frac{R}{1 - R} \right).
\]

The admittance \( Y \) is defined as the reciprocal of \( Z \) (i.e. \( Y = 1/Z \)). Solving for \( R \) in Equation 2.26, the pressure reflection factor is obtained as:

\[
R = \frac{Z \cos \xi - Z_0}{Z \cos \xi + Z_0} = \frac{Z_0}{Z_0} \cos \xi - 1
\]

Finally, since the intensity of a plane wave is proportional to the square of the pressure amplitude, the intensity of the reflected wave is smaller by a factor \( |R|^2 \) than that of the incident wave [Kuttruff, 2009]. Hence, the relationship between \( \alpha_p \) and \( R \) can be defined as \( \alpha_p = 1 - |R|^2 \), which leads to the absorption coefficient equation:

\[
\alpha_p = \frac{4 \Re\left( \frac{Z}{Z_0} \right) \cos \xi}{|\xi|^2 \cos^2 \xi + 2 \Re\left( \frac{Z}{Z_0} \right) \cos \xi + 1},
\]

where \( \Re(\cdot) \) is the real part operator.

### 2.2.2 Scattering and Diffraction

Although their properties will not be explicitly investigated during the rest of the thesis, it is important to briefly describe, in this section, phenomena like scattering and diffraction. In fact, they are implicitly related to models and concepts that will be later employed. For instance, scattered reflections are included in the RIR reverberation part, that will be described in Section 2.3.3, whereas diffracted waves wrapping around the human head are utilised by the human auditory system to localise sound sources, as will be described in Section 3.3.

In the real world, reflectors like walls, ceiling, and floor are never completely smooth. Instead, they present irregularities, given by coffers, bumps or other projections. In the
2.2.2. Scattering and Diffraction

Figure 2.5: On the left it is represented a reflector having irregularities on its surface with width $d$ much smaller than the sound wavelength $\lambda$. The central figure shows a configuration where $d \approx \lambda$. The figure on the right shows a reflector having $d >> \lambda$.

Figure printed from [Kuttruff, 2009].

situation where the irregularity width $d$ is much smaller than the sound wavelength $\lambda$, they do not affect the reflected sound, which follows the specular reflection properties, described in Section 2.2.1. However, whenever $d >> \lambda$, each smaller surface can be considered as a plane section, thus, influencing the incident sound wave by specularly reflecting it. The case when scattered reflections are produced is given by $d \approx \lambda$. In fact, when the irregularity width can be considered similar to the sound wavelength, the incident sound is reflected (i.e. scattered) towards every direction [Kuttruff, 2009]. These three cases are graphically depicted in Figure 2.5.

Considering, instead, reflectors having semi-infinite dimensions, the edge which is exposed to the incident sound wave produces an additional sound wave, which propagates with the rest of the sound, interfering with it. This additional wave is produced by diffraction, thus, it is usually defined as diffraction wave [Kuttruff, 2009]. This phenomenon is illustrated in Figure 2.6 (A). At the centre of this figure, it is reported the semi-infinite reflector. On its left, the planar sound wave propagates parallel to the reflector normal direction. Once it reaches the reflector position, the wave: continues its propagation where it does not encounter any obstacle; is reflected where it encounters the semi-infinite reflector; is diffracted behind the reflector where it encounters the edge of the reflector. It is interesting to note that the diffraction waves produced by the semi-infinite reflector edge allow the area that is “behind” the reflector (i.e. the so called shadow zone), to be reached by the propagating sound. This an important physical characteristic, since the same concept is exploited by the human auditory system to localise sound sources.

As it will be described in more detail in Chapter 3, one of the two main cues, employed by the human brain to estimate the direction of arrival of a specific sound, is the time difference between the sound arriving at the nearest ear to the source and the other one [Howard and Angus, 2009]. For instance, by assuming the sound source as being positioned at the left of the listener, the left ear will receive it before than the right ear.

Assuming the human head to be a sphere with radius $r_{\text{head}}$, a mathematical definition, that does not take into account any diffraction phenomenon, would calculate the time
delay between the two ears as:  
\[ \tau = \frac{2r_{\text{head}} \sin \Theta}{c_0} \]  
(2.29) 
where \( \Theta \) is the azimuth DOA of the signal. However, in reality, as depicted in Figure 2.6 (B), the head produces a diffraction, and the time employed by the sound to travel around it must be considered. Therefore, the time delay between the two ears is typically defined as [Kearney et al., 2010]:  
\[ \text{ITD} = \frac{r_{\text{head}}(\Theta + \sin \Theta)}{c_0} \]  
(2.30)

This time delay is known as interaural time difference (ITD). Another way of defining this interaural delay is through the definition of the interaural phase difference (IPD) \( \phi^{\text{IPD}} \). IPD is related to ITD through the equation \( \phi^{\text{IPD}} = \omega \text{ITD} \), with \( \omega \) being the angular frequency.

### 2.3 Room Acoustic Modelling

#### 2.3.1 The Green’s Function

The Green’s functions are mathematical tools that can be employed to solve initial- and boundary-value problems, involving differential equations [Duffy, 2015]. They are also usually defined as impulse responses of homogeneous systems. In this section, they are utilised to solve the planar sound wave equations, considering indoor propagation.
The acoustic wave equation, defined in Equation 2.11, can be also written as [Duffy, 2015]:

\[ \nabla^2 g(X, t|B_0, t_0) - \frac{1}{c_0} \frac{\partial^2 g(X, t|B_0, t_0)}{\partial t^2} = -\delta(X - B_0)\delta(t - t_0), \] (2.31)

where \( X \) is the spatial variable containing the 3D coordinates of a Cartesian system, \( B_0 \) is the source position coordinate vector, \( t_0 \) is the time of arrival (TOA) of the sound propagating between \( B_0 \) and a point of \( X \), \( \delta(\cdot) \) indicates the Dirac function, and \( g(X, t|B_0, t_0) \) is a Green’s function. To respect the causality property, it is defined to be:

\[ g(X, t|B_0, t_0) = 0, \quad \forall t < t_0. \] (2.32)

The Green’s function related to the sound wave equation will be here derived for the 1D space case. Therefore, considering Equation 2.12:

\[ \frac{\partial^2 g(x, t|b_{x,0}, t_0)}{\partial t^2} - c_0^2 \frac{\partial^2 g(x, t|b_{x,0}, t_0)}{\partial x^2} = c_0^2 \delta(x - b_{x,0})\delta(t - t_0), \] (2.33)

where \( b_{x,0} \) is the position of the source in the 1D space. If the general Equation 2.32 is respected, then the Laplace transform of Equation 2.33 can be calculated, to obtain:

\[ \frac{\partial^2 G(x, s|b_{x,0}, t_0)}{\partial x^2} - \frac{s^2}{c_0^2} G(x, s|b_{x,0}, t_0) = -\delta(x - b_{x,0}) \exp(-st_0), \] (2.34)

where \( s \) is the Laplace domain variable, and \( G(x, s|b_{x,0}, t_0) \) is the Laplace transform of \( g(x, t|b_{x,0}, t_0) \). By calculating the Fourier transform of this equation, with respect to the space, the joint Laplace-Fourier transform of the Green’s function is obtained as:

\[ G(k, s|b_{x,0}, t_0) = \frac{\exp(-jkb_{x,0} - st_0)}{k^2 + \frac{s^2}{c_0^2}}. \] (2.35)

By then applying the inverse Fourier transform, it results:

\[ G(x, s|b_{x,0}, t_0) = \frac{\exp(-st_0)}{2\pi} \int_{-\infty}^{\infty} \frac{\exp[jk(x - b_{x,0})]}{k^2 + \frac{s^2}{c_0^2}} dk. \] (2.36)

The integral on the right side of the equation can be solved by applying the residue theorem [Knopp, 1996]. Therefore, Equation 2.36 can be rewritten as:

\[ G(x, s|b_{x,0}, t_0) = \frac{c_0 \exp\left(\frac{-st_0 - s|x - b_{x,0}|}{c_0}\right)}{2s}. \] (2.37)
Finally, by applying the second shift theorem, the Laplace transform can be solved, to obtain:

\[ g(x, t|b_x,0,t_0) = \frac{c_0}{2} U \left( t - t_0 - \frac{|x - b_x,0|}{c_0} \right), \tag{2.38} \]

where \( U(\cdot) \) stands for the unitary step function. This is the Green’s function related to a sound wave propagating in 1D.

### 2.3.2 The Image Source Model

In the previous section, it was demonstrated that a sound wave can be modelled by employing a Green’s function. Considering the sound source as a point in a rectangular cavity, the interaction of the propagating sound with the reflectors can be defined utilising the image source model. Also in 3D spaces, the pressure wave of a sinusoid can be defined, with respect to the source \( B_0 = [b_x,0, b_y,0, b_z,0]^T \) and the sensor \( A = [A_x, A_y, A_z]^T \) positions, where \([\cdot]^T\) indicates the transpose operator [Allen and Berkley, 1979]:

\[ P_{B_0}(\omega) = \frac{\exp \left[ j\omega \left( \frac{|B_0-A|}{c_0} - t \right) \right]}{4\pi |B_0 - A|}. \tag{2.39} \]

The boundary condition for a smooth planar reflector of infinite length is to have zero normal velocity vector. As discussed in Section 2.2.1, such a reflector generates a specular reflection. Therefore, by assuming its absorption coefficient \( \alpha_p = 0 \), the boundary condition is satisfied by defining an additional sound source, i.e. the image source, placed symmetrically to the main source, with respect to the reflector, as shown in Figure 2.7(A). Generalising this definition, given a planar reflector \( p \), the propagating
2.3.3. The Room Impulse Response

The pressure wave equation can be modified as [Allen and Berkley, 1979]:

\[ P(\omega) = \left[ \frac{\exp(jk|B_0 - A|)}{4\pi|B_0 - A|} + \frac{\exp(jk|B - A|)}{4\pi|B - A|} \right] \exp(-j\omega t), \quad (2.40) \]

where \( B = [b_x, b_y, b_z] \) is a vector containing the coordinates of the image source. Assuming the reflector to lie at \( x = 0 \), the coordinates of \( B \) are defined as:

\[ b_x = -b_{x,0}; \quad b_y = b_{y,0}; \quad b_z = b_{z,0}. \quad (2.41) \]

Proceeding towards real world scenarios, six reflectors are now going to be considered, to create a shoebox-like environment. Obviously, the image-source computation becomes much more complicated, since each image source is itself imaged with respect to every reflector, as shown in Figure 2.7 (B). Following the derivation included in the appendix of [Allen and Berkley, 1979], the total pressure wave can be written as:

\[ P_{sp}(\omega) = \sum_{r=1}^{8} \sum_{m_x=-\infty}^{\infty} \frac{\exp(jk|\Delta B_r + \Delta B_{m_x}|)}{4\pi|\Delta B_r + \Delta B_{m_x}|} \exp(-j\omega t), \quad (2.42) \]

where \( m_x = [m_1, m_2, m_3]^T \) is a vector of integer numbers, and \( \Delta B_r \) contains eight vectors given by the eight possible permutations over \( \pm \), as:

\[ \Delta B_r = [b_{x,0} \pm A_x, b_{y,0} \pm A_y, b_{z,0} \pm A_z], \quad (2.43) \]

where \( 1 \leq r \leq 8 \) is the permutation index. On the other hand, \( \Delta B_{m_x} \) is defined as:

\[ \Delta B_{m_x} = [2m_1L_x, 2m_2L_y, 2m_3L_z], \quad (2.44) \]

where \( L_x, L_y \) and \( L_z \) are the room dimensions.

Equation 2.42 is the pressure wave, expressed in the frequency domain, by assuming six reflectors producing specular reflections, a source position \( B_0 \) and a microphone position \( A \). By calculating the inverse Fourier transform of this function, a Green’s function takes shape in the time domain [Allen and Berkley, 1979]:

\[ P_{sp}(t) = \sum_{r=1}^{8} \sum_{m_x=-\infty}^{\infty} \frac{\delta(t - (|\Delta B_r + \Delta B_{m_x}|/c_0))}{4\pi|\Delta B_r + \Delta B_{m_x}|}, \quad (2.45) \]

This equation can be interpreted as spherical pressure waves, which simultaneously start to propagate from every image source (and main source).
2.3.3 The Room Impulse Response

The image source method was firstly introduced in [Allen and Berkley, 1979], to mathematically characterise the sound wave behaviour in a specific environment. Specularity was assumed for the reflections, leading to the Green’s function defined in Equation 2.45. That Green’s function, is also known as RIR (i.e. an approximation of a RIR). In fact, a RIR can be defined as superimposition of several Dirac’s deltas $\delta(\cdot)$, representing the sound arriving at the sensor from the source and the image sources. Therefore, considering each $e$-th reflection having a path dependent attenuation $\alpha_{e}^{\text{path}}$ and TOA in samples $n_{e}$, a RIR can be also written as:

$$I(n) = \sum_{e}^{\alpha_{e}^{\text{path}}} \delta(n - n_{e}) = \sum_{e} h_{e}(n - n_{e}),$$  \hspace{1cm} (2.46)

where $n$ is the discrete time variable\(^2\). A zoom in on the first 20 ms, of a RIR that was simulated through the image source method, is shown in Figure 2.8 (A). The same zoom in on the correspondent recorded RIR, measured in a laboratory at the University of Surrey (i.e. the “Vislab”), is also depicted in Figure 2.8 (B). A part from the first reflections, it is evident that the two signals are different. This is mainly caused by the reflection specularity assumption made by the image source method.

In the real world, scattering and diffraction, produced by non-perfectly smooth reflectors and objects present in the environment, provide additional reflections. Therefore, the approximation made by assuming the reflections to be only specular is correct, but only for the first reflections [Välimäki et al., 2012]. Following this concept, RIRs are commonly known to be composed by three parts: the direct sound $h^{D}(n)$, the early reflections $h^{E}(n)$ (that are specular) and the late reverberation $h^{L}(n)$ (non-specular).

\(^2\)It is possible to go from the continuous time domain, represented by the variable $t$, to the discrete time domain, represented by the variable $n$, by utilising the sampling frequency $f_{s}$, as: $n = t/f_{s}$. 

---

**Figure 2.8:** Absolute value of a RIR simulated through the image source model (A), and absolute value of a RIR recorded in “Vislab” (B), a laboratory at the University of Surrey.
2.3.3. The Room Impulse Response

![Figure 2.9: Representation of the three RIR components. The direct sound is reported in red, the early reflections in green, and the late reverberation in blue. Modified from Välimäki et al., 2012.](image)

reflections are dominant) [Kuttruff, 2009]:

\[
I(n) = h^D(n) + h^E(n) + h^L(n) = h_0(n - n_0) + \sum_{e=1}^{T_m} h_e(n - n_e) + \sum_{e=T_m+1}^{T_L} h_e(n - n_e) + w(n),
\]

(2.47)

where \(T_m\) is the last early reflection, \(T_L\) is the last reflection of the recorded RIR, and \(w(n)\) is the white Gaussian measurement noise. Thus, the direct sound is defined for \(e = 0\), the early reflections for \(1 \leq e \leq T_m\), with \(e \in \mathbb{N}\), where \(\mathbb{N}\) is the natural number set, and the late reverberation for \(T_m + 1 \leq e \leq T_L\). A graphical visualisation of the three RIR parts is reported in Figure 2.9.

Following these definitions, it can be now better understood the brief discussion had in Chapter 1: while the direct path reveals the direction of the source, the early reflections convey a sense of the geometry, whereas the late reverberation is indicative of the size of the environment [Välimäki et al., 2012]. Observing the situation from the perceptual point of view, the late reverberation can also be defined as the perception of many remote reflections from distant surfaces. On the other hand, the early reflections are sonic manifestation of a nearby object [Blesser, 2001].

To sum up, RIRs characterise the acoustic of a specific environment, given source and sensor position. Being \(x(n)\) a sound produced by the source, the resulting sound received at the sensor is defined as:

\[
y(n) = x(n) * I(n),
\]

(2.48)

where the symbol “*” is the convolution operator. This means that, a part from additive noise, the received signal is characterised by attenuated and delayed versions of the same
signal that was generated at the source.

2.4 The Sound Perception in Indoor Environments

The physical and mathematical description of the sound, that was provided within the previous sections, represents the main theoretical foundation of this thesis. However, studies about the psychoacoustical point of view must also be considered, since the final aim of this thesis, is to enhance the sound quality perception of human listeners.

2.4.1 The Perception of the First Reflection

The first early reflection is, usually, a strong specular reflection. For this reason, it is highly correlated to the direct sound, and, by interacting with it, it generates acoustical effects that modify the perception of the produced sound. Here, some of the most important effects are described, which are deemed to be fundamental for the development of the next thesis chapters.

The Precedence Effect. Whenever two highly correlated sounds reach a listener in a short difference of time, they are perceived as a single fused auditory event. The threshold that defines the limit between two sounds being perceived as either fused or distinct is usually set between 5 ms and 40 ms, depending on the sound waveform [Wallach et al., 1973]. The localisation of the fused event is mainly given by the directional cues carried by the earlier sound. This observation is known as the precedence effect [Litovsky et al., 1999, Zurek, 1987]. A practical example of the precedence effect can be provided by the situation when a listener tries to localise a sound source in a reverberant environment. If the path between listener and sound source is not obstructed, the direct sound arrives from the direction of the source. On the other hand, reflections that are coming from nearby surfaces arrive slightly later and from other directions. Because of the precedence effect, localisation of the main source is usually accurate [Zurek, 1987].

The Comb Filter Effect. In environments where the first reflection has a delay between 10 ms and 50 ms with respect to the direct sound, the colouration of the sound perceived is different from the one of the sound produced at the source. This colouration effect is given by the interference between a signal and a delayed version of it, that produces a comb filter [Barron, 1971]. This comb filter effect is more noticeable when the reflection arrives from the same azimuth of arrival of the direct sound [Barron, 1971]. This can be explained by the fact that, for lateral reflections, the precedence effect mitigates the reflection localisation cues [Lokki et al., 2011].
Source Width. Since for lateral reflections the precedence effect suppresses the reflection localisation cues, in the literature, some studies found the reflections produced by floor and ceiling to be providers of most of the perceptual properties [Bech, 1998]. However, also a first reflection which is strong and lateral plays an important role in the sound perception. For instance, in some studies, it was found that, for a delay above 10 ms, the reflection change the source perception, by broadening it. This corresponds to a gain in both body and fullness of the produced sound, improving the listener impression of being in a 3D space [Barron, 1971]. This perceptual property is typically referred to as spatial impression. Some researchers defined the spatial impression as the difference between “feeling inside the music and looking at it through a window” [Marshall, 1967].

Listener Envelopment. The spatial impression is not only characterised by the apparent source width, but also by the listener envelopment (immersion). As described in the previous paragraph, the early lateral reflection energy increases the perceived apparent source width. On the other hand, the listener envelopment is usually high whenever there is significant late arriving lateral energy [Bradley and Soulodre, 1995]. Furthermore, recently, it was demonstrated that the listener envelopment is one of the most important attributes that listeners expect to perceive correctly while exposed to spatial audio reproductions [Francombe et al., 2017a,b].

Distance and Depth Perception. There is a distinction to make between distance perception and the so called depth perception. Distance is the perception related to the spatial range between the listener and the sound source, whereas depth relates to the auditory scene as a whole [Kearney et al., 2012]. Apparent source distance effects can be obtained by modifying the direct to reverberant ratio (DRR) of the sound [Zahorik, 2002]. On the other hand, the early reflection positions have an important role in the depth perception, due to their relation to the acoustic environment. For instance, in a virtual environment, movements of the listener must accordingly correspond to changes in the early reflections, in order to maintain a coherent depth impression [Kearney et al., 2012].

2.4.2 The Mixing Time

The mixing time is the instant that divides the early reflections from the late reverberation in a RIR, and it is represented in Equation 2.47 by the symbol $T_m$. Perceptually, it can also be defined as the instant when the reverberation cannot be distinguished from that of any other position of the listener in the room [Lindau et al., 2012]. The mixing time is an important element, since it plays an important role during artificial
reverberation generation [Schlecht and Habets, 2017]. Different approaches are usually proposed to estimate it: perceptual based approaches and direct measurement based approaches. The perceptual based approaches usually follow results of listening tests, by creating mathematical functions, which relates the environmental acoustic properties to the mixing time [Lindau et al., 2012]. On the other hand, measurement based approaches usually either estimate the *echo density* of a recorded RIR [Abel and Huang, 2006], or the amount of diffuseness of the RIR segments [Stewart and Sandler, 2007].

### 2.4.3 The Reverberation Time

Different parameters can be utilised to evaluate the acoustic properties of enclosed spaces, such as concert halls, classrooms, and stadiums. One of these is the *reverberation time* (RT60) [Kendrick et al., 2008]. RT60 is strictly related to the environment, and it is defined as the time employed by the total energy of the RIR to be reduced of 60 dB from the maximum value. The common way to calculate it is by following the Sabine’s equation, as [Kuttruff, 2009]:

\[
RT60 \approx 0.161 \frac{V_{TOT}}{S_{TOT} \bar{\alpha}},
\]

(2.49)

where \(V_{TOT}\) is the total room volume, \(S_{TOT}\) is the total room reflective surface, and \(\bar{\alpha}\) is the room absorption coefficient, averaged over all the reflectors.
Chapter 3

Reflector Localisation and its Applications: a Literature Review

This chapter aims to provide the reader with knowledge of the current state-of-the-art, related to this thesis. It is divided into three main sections, representing the three contribution areas of this thesis: acoustic reflector localisation, spatial audio, and blind source separation. These sections are shaped following a classical literature survey structure: first they provide a general overview of the research topic; then they analyse the literature in detail, also categorising the different approaches. Beyond these, an additional fourth section will present popular algorithms and methods, that will be employed in the later chapters: epoch detectors for the time of arrival (TOA) estimation; beamformers for the direction of arrival (DOA) estimation; the linear predictive coding (LPC) [Makhoul, 1975] for the frequency domain analysis; the homogeneous coordinates [Hartley and Zisserman, 2003] for the construction of 3D surfaces; some geometric transformations for the 3D space.

3.1 On Acoustic Reflector Localisation

The previous chapters underlined the importance of having knowledge of the acoustic reflector positions, related to recording environments. This is a key passage for this thesis, since it is one of its main objectives.

Algorithms and methodologies presented in different audio research areas are typically vulnerable to both strong early reflections and high level of reverberation. This is due to the usual assumption made about the important signals to be carried by only the direct sound. Therefore, early reflections are usually considered as interfering reflections
3.1.1 Reflector Estimation from RIRs

[Elliot and Nelson, 1993]. Nevertheless, if one would be able to localise the image sources of these early reflections, their information could be constructively exploited, thus considering them as supportive reflections [Krawczyk and Gerkmann, 2014]. This new way of considering acoustic signals within reverberant environments can be exploited in audio forensics [Malik, 2013], simultaneous localisation and mapping [Dokmanić et al., 2016, Naseri and Koivunen, 2017], and spatial audio [Zotkin et al., 2004], by allowing a better source localisation. In addition, new acoustic features can be utilised to enhance target signals in fields such as automatic speech recognition [Yoshioka et al., 2012], music transcription [Plumbley et al., 2002], source separation [Asaei et al., 2014], audio tracking [Öçal et al., 2014], dereverberation [Naylor and Gaubitch, 2010], and microphone array raking [Dokmanić et al., 2015b, Javed et al., 2016]. Finally, it is also important to note that an acoustic reflector localiser can be combined with image processing techniques, to construct a hybrid room geometry estimation method [Hussain et al., 2014, Ye et al., 2015]: acoustic information can aid detection of mirrors and windows, that cannot be identified by a visual sensor.

3.1.1 Reflector Estimation from RIRs

In Section 2.3.3, the room impulse response (RIR) was defined as the Green’s function that characterises the multipath propagation of an acoustic wave within an enclosed environment, given source and sensor position. Furthermore, it was described that the assumption of specular reflections implies the early reflections to be the carriers of information related to the acoustic reflector positions. For this reason, the literature presents several methods, that have been developed during the last decade, to localise acoustic reflectors given recorded RIRs. In this section, most of them will be discussed, classifying them into three groups: 2D methods, half 2D methods, and 3D methods.

As expressed by their category name, the 2D methods localise reflectors by assuming them as lines in the 2D space; half 2D methods still assume reflectors as lines, however, these lines are thought as projections of 3D planes; finally, 3D methods directly consider reflectors as planes in a 3D space.

2D Methods. Several approaches were presented to localise reflectors from RIRs, as the solution of a 2D problem: loudspeaker, microphone and reflector are assumed to lie on the same plane. Most of these works are the pioneers of the reflector localisation research field. Nevertheless, they are limited by the number of spatial dimensions assumed. These approaches, can be subdivided between: methods that exploit a single channel RIR, and methods that exploit multiple channel RIRs.
Starting with the single channel RIR methods, the unique relationship between a recording setup in a room and the respective RIR was exploited in [Dokmanić et al., 2011]. First and second order image sources\(^1\) were combined to infer the reflector positions. In [Marković et al., 2013a], the early reflection TOAs, related to a recorded RIR, were compared to synthetic RIR early reflection TOAs. Minimising a cost function, the room geometry corresponding to the synthetic RIR, that best fitted the recorded one, was defined as the estimated one. The interesting point there was the fact that they also proved this method to work with L-shaped rooms. However, it was not extended to 3D environments. Another method exploiting the first and second order image sources was presented in [Moore et al., 2013]. Different from [Dokmanić et al., 2011], the relationship between adjacent and opposite reflectors was utilised to estimate the 2D room geometry.

Let us now consider the methods exploiting multiple channel RIR information. The work that can be considered as the first one employing geometric figures to calculate reflector positions was presented in [Antonacci et al., 2010]. Ellipses were chosen to exploit their property of having constant the sum of the two Euclidean distances calculated between its foci and every point on its surface, as shown in Figure 3.1 (A). Therefore, by constructing ellipses having foci on source \(B_0\) and sensor \(A\), and major axis \(Q_{maj}\) equal to the reflection path length, each of the infinite point on its surface is a putative point where the reflector is tangent. This reflector, in the 2D space is represented by a line. By constructing one ellipse for every microphone-loudspeaker combination available,
the line which is the common tangent to all the ellipses corresponds to the wanted reflector. In order to find this common tangent line, it was then proposed the common tangent (COTA) algorithm, which is based on a cost function minimisation process. Later, in [Canclini et al., 2011b], an improvement of the cost function used by COTA was presented. An exact solution to the common tangent line problem was proposed, by turning the optimisation problem into a linear least-squares (LS) one. Then, in [Canclini et al., 2011a], the authors applied COTA to find the common tangent line to a set of parabolas. These parabolas were constructed having focus on the source position \( B_0 \) and directrix representing the DOA with respect to the sensors \( A_i \), where \( i \) is the microphone index. The focal property of a parabola states that the line which is tangent to a point on it, also bisects the angle formed between the line joining that point to the parabola focus and the perpendicular to the parabola directrix through the same point. This property is schematically described in Figure 3.1 (B). A further extension of the work proposed in [Antonacci et al., 2010], was presented in [Antonacci et al., 2012]. The results given by COTA were refined, by applying the Hough transform as post-processor. The method that was later presented in [Baba et al., 2016] included the ellipse construction technique previously proposed in [Antonacci et al., 2010]. However, it extended it, also employing the image microphone concept\(^2\). In particular, a reflection point was obtained by calculating the intersection between the ellipse and the line connecting source and image microphone. Finding a point for each source available, the reflector was then estimated as the line fitting all these points.

**Half 3D Methods.** The aforementioned 2D methods have in common the limitation given by the assumed number of spatial dimensions. Therefore, recently, 3D models were also investigated. As intermediate step, half 3D methods were firstly proposed in the literature, starting from the previously presented 2D methods. For instance, in [Filos et al., 2012], the work in [Antonacci et al., 2012] was extended to 3D spaces. However, it was not yet a full 3D estimation, instead combining 2D projections to estimate the position of the surfaces outside of the measurement planes.

**3D Methods.** Methods estimating acoustic reflector positions in the 3D space are here presented. The state-of-the-art 3D reflector localisation methods can approximately be categorised into two groups. The first one is named as *image-source reversion*, where the TOA is used to revert to the image source location [Allen and Berkley, 1979], which is subsequently used to determine the reflector position. The image source reversion group is composed of [Arteaga et al., 2013, Dokmanić et al., 2013, Ribeiro et al., 2012, Tervo and Tossavainen, 2012]. The second group contains those methods directly localising the reflector, without estimating any other element about the room acoustic first, and

\(^2\)Similar to the image source, the image microphone is the virtual microphone calculated by mirroring the real microphone with respect to the reflector position.
3. Reflector Localisation and its Applications: a Literature Review

It is composed of [Kuster et al., 2004, Nastasia et al., 2011, Zamaninezhad et al., 2014]. Accordingly to its definition, this group is named as direct localisation.

The method in [Tervo and Tossavainen, 2012] employed the image-source reversion approach to localise reflectors in 3D. The image source positions were estimated by utilising a maximum-likelihood (ML) approach. More precisely, points in the space were randomly generated, and a cost function was maximised to find that point which produces the time difference of arrival (TDOA), that is the most similar to the TDOA extracted from the RIRs. Then, knowing the position of the source, the reflector was found as the plane perpendicular to the vector $u$, that connects the source and the image source, and passing through their midpoint $M^3$. Whilst it can be considered as the first method reverting the image source model, it needed a large number of generated points to achieve acceptable performance. In [Ribeiro et al., 2012], a LS minimisation was utilised to fit “synthetic” reflections to recorded RIRs. However, it required a large number of RIRs. The synthetic reflections were obtained in an anechoic room, utilising a plastic reflector to simulate a wall. In total, 240 loudspeaker positions were used to collect the recordings, leading to 1440 RIRs, considering the six-element microphone array employed. Since the number of RIRs was deemed not enough to train the model, the RIRs were additionally interpolated in space, in order to have more DOAs. In [Arteaga et al., 2013], firstly the source-sensor distance and the reverberation time (RT60) were calculated, from a single measured RIR, as acoustic features corresponding to the room. Synthetic RIRs were then generated, accordingly to these estimated features, employing the image source method [Allen and Berkley, 1979]. Finally, a cost function was minimised to find the correct room shape, by calculating the cross-correlation between the recorded RIR and the set of synthetic ones. Although this approach proved that also the 3D estimation can be endeavoured to do by using a single RIR, this was, at the same time, a limiting factor to the overall performance. In fact, the source-sensor distance and the RT60 estimation was not robust with respect to errors. By assuming the TOAs to be known a priori, the main contribution of [Dokmanić et al., 2013] was a novel algorithm to label the reflections from a distributed microphone array, where the reflector order would otherwise be ambiguous, if compared among different microphone recordings. To do so, a Euclidean distance matrix (EDM) [Dokmanić et al., 2015a] was constructed considering the distances between the microphones and the source. Then, the respective TOAs were utilised to augment the matrix. The TOAs that did not correspond to the same reflector were identified as the ones producing a non-EDM matrix. The image sources were then localised by employing the multilateration algorithm [Beck

---

$^3$This method of finding the plane that is perpendicular to the source-image source connecting line, and passing through their midpoint, will be utilised also for some of the proposed methods, later in Chapter 4. There, it will be named as loudspeaker image bisection (LIB). This method is schematised in Figures 3.2 (A) and (B).
Figure 3.2: Visualisations of the loudspeaker image bisection (LIB) algorithm. (A) is a 2D scheme showing the midpoint $M$ between source and image source, and the vector $u$ that is normal to the estimated plane $p$. (B) is a 3D representation of source (red circle), image source (blue cross) and estimated reflector (plane).

et al., 2008], whereas, the final reflector localisation was performed through the same algorithm proposed in [Tervo and Tossavainen, 2012], and depicted in Figure 3.2.

One of the first attempts to employ the direct localisation approach was proposed in [Kuster et al., 2004]. Exploiting the inverse wave field extrapolation, the authors mapped reflections from a set of receivers to the related reflecting objects, generating a temporal visualisation of the received reflections. This method provided an accurate analysis, and it was even able to identify small reflecting objects lying between the main reflector and the microphone array. However, the microphone array was assumed to be exactly parallel to the reflector. The method that can be considered as the first one exploiting direct localisation through 3D surfaces, was presented in [Nastasia et al., 2011]. As a general extension of the work that was done for the 2D space in [Antonacci et al., 2010], the COTA algorithm was modified to search the plane (i.e. the reflector) that is common tangent to every ellipsoid having foci on the microphone-source positions. Despite this, the peaks were not automatically identified to extract TOAs from RIRs. In addition, the reflector search was computationally expensive, caused mainly by the optimisation of its cost function. In [Zamaninezhad et al., 2014], reflector positions were estimated by transforming the RIR into the frequency domain, i.e. calculating the room transfer function (RTF). The distance between parallel reflectors was calculated knowing that it is related to the resonant frequency of the RTF. Therefore, a cost function minimisation was employed, comparing the possible resonant frequencies with the measured RTF. In spite of being the first method approaching the problem from the frequency domain prospective, only two parallel reflectors were localised, assuming the other boundaries as being completely absorbent.
3. Reflector Localisation and its Applications: a Literature Review

3.1.2 Literature Limitations

Various limitations can be identified in the literature about 3D reflector localisation given RIRs. First of all, although several algorithms were designed for single channel RIR peak detection [Brookes, 1997, Defrance et al., 2009, Kuster, 2008, Usher, 2010], there is not a peak-picking algorithm specifically designed to automatically extract reflection TOAs from multichannel RIRs, that are recorded using compact microphone arrays. Another limitation is given by the fact that some methods assume specific spatial relationships between the microphone array and the reflectors [Kuster et al., 2004, Zamaninezhad et al., 2014]. Furthermore, other methods may require thousands of RIRs recorded in anechoic rooms [Ribeiro et al., 2012]. Moreover, microphones are often considered to be spatially sparse, introducing the problem of labelling echoes [Dokmanić et al., 2013]. There are also no proposed ways to aggregate measurements from multiple loudspeakers to improve localisation. Finally, classical image-source reversion methods (e.g. [Dokmanić et al., 2013, Tervo and Tossavainen, 2012]) use TOAs to localise the image source without considering other information carried by the RIRs, limiting their robustness.

3.1.3 Reflector Estimation From Continuous Signals

Although the novel methods, that will be proposed in Chapter 4, localise the reflector position given RIRs, it is important to note that, in the literature, some other works are available to estimate the reflector positions given recorded continuous signals.

Assuming knowledge of the microphone and source position, in [Tervo and Korhonen, 2010], reflectors were localised by inverse mapping the acoustic multipath propagation. This was done by maximising a cost function, that took into account every TDOA allowed by the pairs of microphones. The TDOAs were evaluated by calculating the cross-correlation between the signals received at each microphone pair. In [Sun et al., 2011] and [Mabande et al., 2013], the authors proposed a method to localise the image sources, and the related reflector, utilising a spherical array of microphones. A beamformer was employed, working in the eigenbeam domain, to obtain the DOAs of the reflections. From the beamformed signals, obtained by steering the beams towards the reflection directions, the TDOAs were estimated utilising a cross-correlation algorithm. Combining together TDOAs and DOAs, the reflector positions were obtained. Another method was proposed in [Crocco et al., 2014], where a non-linear LS approach was utilised to calculate the TDOAs, by evaluating the continuous signals in the frequency domain. Considering the direct sound TDOAs, microphones and sources were
localised. After having removed possible ambiguities related to the reflection TDOAs, a cost function was then minimised to find the reflector positions.

Although interesting, these kind of approaches based on continuous signals face issues on identifying TDOAs, especially if the signal-to-noise ratio (SNR) is not high. This is the reason why methods exploiting directly RIRs are more present in the literature.

3.2 On Spatial Audio

One of the major aims in spatial audio is to reproduce the acoustical characteristics of an indoor environment, with the intention of providing the listener with the sensation of being in the recorded ambient. As already mentioned in Chapter 1, the research answer to this aim can be defined through the virtual acoustic environment modelling, that can be subdivided into three main tasks [Savioja et al., 1999]: source modelling [Coleman et al., 2014a] (e.g. natural audio, synthetic audio, source directivity), room modelling [Lee et al., 2012] (e.g. modelling of acoustic spaces, artificial reverberation) and listener modelling [Masterson et al., 2012] (e.g. head-related transfer functions (HRTFs), microphone directivity). The novel methods that will be presented in Chapter 5 focus on room modelling, with the aim of spatial audio object (SAO) production and reproduction [Herre et al., 2014].

3.2.1 Spatial Audio Scene Representation

Maintaining the characteristics of an acoustic scene implies to capture and model information about it, in order to finally reproduce it for the listeners. In general, a spatial audio scene can be modelled following three different approaches: channel-based, transform-based, and object-based [Spors et al., 2013]. The channel-based approaches are limited in terms of data transmission, since each loudspeaker feed has to be transmitted. In addition, they do not allow flexibility in terms of loudspeaker configuration, since each target reproduction system must be mixed by the sound engineer. On the other hand, transform-based approaches allow more flexibility in this sense, since the scene, which is mapped into orthogonal bases functions, is sent to the receiver, where it is decoded and mapped to the configuration of loudspeaker that is available. However, the audio elements are fixed at the transmission, thus, not allowing, at the production stage, acoustic scene processing. Instead, object-based approaches rely on the definition of a number of SAOs, each one represented by metadata, including parameters characterising the original acoustic scene, and audio content. Metadata is usually decoded at a renderer, which recreates the audio elements relating them to their position within the
Object-Based Approaches. A SAO is a virtual source that, together with others SAOs, describes a scene. Each SAO is positioned at a certain target position in space, as defined in its related metadata. In contrast to the concept of channels, these object positions can be totally independent from the locations of available loudspeakers. Furthermore, they can vary over time for modelling moving objects, such as a plane flying by over the head of the listener [Herre et al., 2015].

The metadata that characterises SAOs generally describe object properties, such as its position in space and magnitude in time. However, this kind of representation would be limiting the possibility of reproducing reverberation. Therefore, together with this properties, the international telecommunication union (ITU) standard, the audio definition model (ADM), allows a SAO to be diffuse or non-punctiform [ITU-R, 2015]. In addition, spread parameters are included within the moving picture experts group (MPEG)-H standard [Füg et al., 2014, Herre et al., 2015].
However, contemporary, SAO standards do not include the concept of reverberant SAO (RSAO). Instead, a common approach is to use a set of objects placed in certain position in space, as a complement to channel-based or transform-based tracks. These SAOs contain the ambience effect [Herre et al., 2015, Oldfield et al., 2014, Stenzel and Scuda, 2014]. Although being a clever idea, it has limitations: the spatialisation is fixed and assumed to be related to a specific loudspeaker layout; every dialogue is embedded into the reverberation; the reverberation control is limited to a simple “with/without” reverberation.

Instead of these “hybrid” RSAO object schemes, a full object-based representation of reverberation would imply its synthesis directly at the renderer. Therefore, this could allow the renderer to provide the listener with a greater immersion, since it can reproduce the early reflections independently. This is an important point since, as already discussed in Chapter 2, the early reflections, if accurately reproduced, convey a sense of the environmental geometry [Välimäki et al., 2012].

Digital synthesis of reverberation has been investigated by researchers for decades [Blesser, 2001, Välimäki et al., 2017, 2016, 2012]. Current approaches to analyse a recorded reverberation and represent it in an object-based content can be classified into signal-based approaches [Albrecht and Lokki, 2013, Carpentier et al., 2013, Francombe et al., 2015, Rumsey, 2001] and parametric approaches [Jot, 1997, Melchior et al., 2010, Merimaa and Pulkki, 2005, Tervo et al., 2013, Väänänen and Huopaniemi, 2004]. Since the novel approach, that will be presented in Chapter 5, will fall into this second group, the focus of the following discussion will be centred on parametric approaches. In particular, a distinction will be made, by further separating them between methods that employ high-level and low-level parameters.

**High-Level Parameters for Object-based Approaches.** Reverberation can be synthesised utilising parameters that describe a room in either physical or perceptual terms. These kinds of high-level parameters were defined within the standard MPEG-4 [Jot, 1997, Väänänen and Huopaniemi, 2004]. Physical parameters [Väänänen and Huopaniemi, 2004] were specified in terms of transmission paths within the environment, and frequency-dependent directivity models for each sound source. They were rendered by computational acoustic modelling and convolution. On the other hand, the perceptual parameters were defined in [Jot, 1997]. Some of them described the source perceptual characteristics, whereas others described the reverberation perception. To generate RIRs, and control the different portions of it (i.e. direct sound, early reflections, and late reverberation), these high-level perceptual parameters were mapped to low-level feedback delay network (FDN) coefficients [Carpentier et al., 2015, De Sena et al., 2015, Schlecht and Habets, 2017, Väänänen et al., 1997], related to mixing time and RT60.
However, the FDN coefficients were mapped to only a fixed number of loudspeakers, making this approach non-format-agnostic. The direct sound was panned accordingly to the source position; whereas the early reflections were created by panning delayed versions of the direct sound.

**Low-Level Parameters for Object-based Approaches.** Low-level parameters can be also defined to synthesise a reverberant environment. These parameters are directly interpreted by the renderer. A method of efficiently parametrising a RIR is the spatial decomposition method (SDM), and it was presented in [Tervo et al., 2013]. SDM is based on the assumption that the RIR is composed by several reflections, each of them defined by an image source in the far field. Segmenting the RIR in the time domain, the DOA of the most prominent reflection for each segment was calculated. This information was then combined with the RIR recorded by an omnidirectional microphone, placed at the centre of the microphone array, in order to spatialise the signal. This combination was performed during the synthesis step, by the renderer. The spatial impulse response rendering (SIRR) is another state-of-the-art method that analyses RIRs and describes them through low-level parameters [Merimaa and Pulkki, 2005]. SIRR was later employed in both the analysis and synthesis stage of the directional audio coding (DirAC) [Pulkki, 2007]. The analysis part was based on either B-Format or higher-order recordings [Politis et al., 2015]. It represented the time-frequency (TF) spatial response through three parameters: both azimuth and elevation DOA, and a diffuseness coefficient. On the other hand, the synthesis part was constructed in two different ways, by differentiating the diffuse and non-diffuse elements. Whereas the non-diffuse parts were mapped to the loudspeaker setup through VBAP [Pulkki, 1997], the diffuse elements were decorrelated and sent to every loudspeaker [Pulkki and Merimaa, 2006]. Similarly to SDM, SIRR represents a good approach to parametrise a room. However, the parameters employed may not be straightforward to adjust during the editing process. Another approach to analyse recorded RIRs, produce low-level parameters, and send them to a renderer, was presented in [Melchior et al., 2010]. The plane wave description of the sound field was utilised as theoretical background, and wave field synthesis (WFS) was employed as the target rendering approach. The wave field was analysed utilising uniform circular arrays (UCAs) of microphones [Hulsebos, 2004]. RIRs in the plane wave domain were divided into two parts, the early and the late part. Furthermore, the most prominent early reflections were extracted by spatio-temporal windowing. In other words, RIRs were assumed to be composed by two early reflection parts, one characterised by discrete reflections, the other one characterised by the rest of the early reflections, having a higher diffuseness. In general, the discrete reflections were editable based on the position and directivity of the direct sound, whereas the later early reflections and the late reverberation part, were fixed for every room.
3.2.2 Object based Audio Literature Limitations

Several limitations can be found in the current object based audio literature. For instance, by employing high-level parameters [Jot, 1997, Väänänen and Huopaniemi, 2004], it is not possible to directly capture and edit the reverberation. Thus, these parameters do not allow the creation of useful tools to directly modify acoustic perception properties, such as apparent source distance or envelopment. On the other hand, low-level parameters allow the creation of such tools. Hence, their employment is preferable, in order to provide audio engineers and final users with more flexibility in terms of perceptual choices.

Nevertheless, other kind of limitations can be found also within those state-of-the-art systems that employ low-level parameters. For instance, some of these methods calculate DOA parameters to describe the late reverberation part of the RIRs [Tervo et al., 2013]. This kind of approach implicitly assumes the late reverberation as carrier of directional information, opposing its usual definition of being purely diffuse [Välimäki et al., 2012]. Other methods describe the RIRs by specifying parameters for each TF bin [Merimaa and Pulkki, 2005, Pulkki and Merimaa, 2006]. This complicates the editing part at the production and reproduction stage, due to the difficulty on find the correct combination of them. Other methods utilise WFS for the rendering of the recreated sounds [Melchior et al., 2010]. The limitation there is given by the fact that this approach forces potential final users to employ a large number of loudspeakers, in order to achieve high quality reproduction.

3.2.3 Spatial Audio Object Coding

As described above, one issue tackled by object-based approaches is the transmission of high-quality multichannel data through band-limited channels. Parametric coding techniques based on spatial audio coding (SAC) were studied [Goodwin and Jot, 2008] with this regard. The MPEG group firstly defined a standard for spatial audio called MPEG Surroun [Herre et al., 2008]. Recently, their activities turned into the so-called spatial audio object coding (SAOC) [Engdegård et al., 2008, Herre et al., 2012], a coding technique, that exploits rendering of multiple SAOs [Herre et al., 2014].

The idea of defining SAOs for sound scene reproduction was included in the MPEG-4 standard. Specifically, a scene description language called binary format for scenes (BIFS) was defined in [Väänänen and Huopaniemi, 2004]. In [Reiter et al., 2006] the authors introduced a subdivision of the early reflections into two parts, to modify the MPEG-4 BIFS perceptual approach. At the same time, a 3D audio object generation
has been presented in [Potard, 2006] following the physical approach. The MPEG group recently started the MPEG-H Audio Coding development [Herre et al., 2014, Murtaza et al., 2015]. The current status of the standardisation project has been reported in [Bleidt et al., 2017, Herre et al., 2015]. This new standard will allow different input formats. Regarding the SAOs, the decoding part will be the one extended from the previous standards, by expanding the unified speech and audio coding stage (USAC) for 3D audio, and defining vector-based amplitude panning (VBAP) [Pulkki, 1997] as the algorithm used to render SAOs.

3.3 On Blind Source Separation

Source separation is a process usually defined as the extraction of independent acoustic signals from mixtures, containing different interfering sounds. The ability of the human system of focusing on one source only, filtering out the rest, is known as cocktail party effect [Bee and Micheyl, 2008, Cherry, 1953]. In particular, the human auditory system is able to estimate the direction of a specific sound source, allowing the subsequent separation [Middlebrooks and Green, 1991]. Source separation is one of the mostly investigated areas in audio signal processing. In general, arrays of one or more microphones are typically utilised to find possible solutions. A scenario where no prior knowledge is available, related to both sources and mixing process, is usually referred to as blind source separation [Naik and Wang, 2014]. A particular branch of it tries to achieve the separation by approximating human hearing, and it is known as computational auditory scene analysis (CASA) [Brown and Cook, 1994].

In general, the importance of artificially separating signals is remarkable, for several application areas. In biomedical signal processing, it is exploited to analyse electrocardiograms, electroencephalograms, or magnetic resonance imaging [Ungureanu et al., 2004]. Moreover, communication and defence signal processing can benefit from audio source separation methods, e.g. developing passive sonar systems [Sutin et al., 2010]. Source separation methods, firstly developed for audio signals, may be also applied to completely different areas, for instance, it may help with the restoration of ancient documents [Tonazzini et al., 2007].

Audio source separation can be subdivided into different branches, depending on the type of data that is processed. Two of the mostly investigated areas study either musical or speech recordings [Vincent et al., 2012]. Although music source separation represents a key part of the audio community [Ewert et al., 2014], the focus of this thesis is on speech source separation. This is of interest for several audio applications, such as:
3.3.1 Speech Source Separation

During the last twenty years, speech separation has gained quite of attention. Some of the proposed methods achieved source separation exploiting the availability of a single microphone [Jang and Lee, 2003, Radfar and Dansereau, 2007, Schmidt and Olsson, 2006]. However, they were limited by the amount of information utilised. Therefore, other methods attempted the separation process by employing multichannel microphone arrays. These methods are classically categorised into three main groups, depending on the type of approach undertaken [Vincent et al., 2012]: the beamformers [Araki et al., 2003, Coleman et al., 2015a, VanVeen and Buckley, 1988]; the independent component analysis (ICA) based [Bell and Sejnowski, 1995, Cardoso, 1998, Makino et al., 2007]; the time-frequency (TF) mask based [Alinaghi et al., 2014, Deleforge et al., 2015, Mandel et al., 2010, Sawada et al., 2011, Yilmaz and Rickard, 2004]. A visualisation of these three different approach characteristics is reported in Figure 3.4.

**Beamformers.** Beamformers are those methods that, assuming spatial sparsity among the sources, perform a spatial filtering to attenuate the interferer signal directions and enhance the target signal. A beamformer categorisation will be described later, in Section 3.4.2. Here, a high-level analysis is provided, from the blind source separation point of view. In blind source separation, beamformers achieve their purpose by employing space-dependent beam-like functions. The main beam is steered towards the target direction. This is commonly done by calculating the covariance among the signals received by every microphone of the array [Van Trees, 2002]. Nevertheless, methods employing...
TDOAs are also available [VanVeen and Buckley, 1988]. A representation of a frequency dependent beam pattern (i.e. the so called *directivity map*) is reported in Figure 3.4 (A).

Beamformers can be used in every scenario, even the *underdetermined* one (i.e. when the number of sources is greater than the number of microphones). However, decreasing the number of microphones used, the performance dramatically drop. Additionally, they do not take into account the acoustic multipath, thus, they are less effective under reverberant conditions.

**Independent Component Analysis (ICA).** ICA is a statistical based method, formally defined for the first time in [Comon, 1994], and expanded, by introducing new objective functions, in [Hyvärinen, 1997]. Exploiting high-order statistics, it finds a linear transformation that can be applied to the received mixture. A representation of the space defined by independent components (ICs) is depicted in Figure 3.4 (B), together with the observed data projected on it. Following its nature, the two main assumptions are the statistical independence of the source signals, and their linear interaction. ICA has been for many years one of the most important approaches for blind source separation, due to its high quality performance, given highly uncorrelated signals. However, it also has three main limitations which discourage its standalone usage: it can only be applied for mixtures where the number of sources is lower or equal to the number of microphones (i.e. *overdetermined* and *determined* scenario, respectively); it faces the so called permutation problem, in other words, it is not possible to associate a reconstructed signal to a specific original source; it has low performance when applied to reverberant mixtures.

**Time-Frequency Masking.** Following the beamformers and ICA limitations, during the last decade researchers mainly focused on TF mask based approaches [Wang, 2008]. They are based on the approximation of the time variant frequency domain of each source signal, following the human hearing functionalities [Brown and Cook, 1994]. As opposed to ICA, TF masks can be used for any kind of scenarios, including the underdetermined one. Different from beamformers, they have good performance in reverberant conditions, even using only a pair of microphones. TF masking methods assume the spectral sparsity of the signals. This property is also known as W-disjoint orthogonality, and it is verified when the supports of the signal short-time Fourier transforms (STFTs) are disjoint, for a given window function W [Yilmaz and Rickard, 2004]. The STFT of a two speech signal mixture is visualised in Figure 3.4 (C).

Recently, several methods were proposed in literature, employing TF masks. In [Yilmaz and Rickard, 2004], by exploiting a pair of omnidirectional microphones that were placed in anechoic rooms, the authors proposed a mathematical model able to extract from the two mixtures the attenuation coefficients and the time delays belonging to the direct
sound paths. Following the $W$-disjoint orthogonality assumption, a binary mask was then created, having value “one” in every TF bin where the source to separate was considered as predominant, and “zero” otherwise. Although this method provided high separation quality, it was limited to anechoic environments. In fact, where reflections are present, the TDOA-amplitude space, that they estimated to characterise the TF bin probabilities, is highly distorted.

In [Mandel et al., 2010], the authors presented a method based on binaural recordings, employing dummy heads, instead of omnidirectional microphones, within reverberant spaces. Two interaural cues were considered: the interaural level difference (ILD) and the interaural phase difference (IPD) [Hofman and Van Opstal, 1998]. These two cues relate the sound DOA to the head orientation. Soft TF masks were generated, by corresponding to each TF bin the probability of being dominated by the energy of a specific source. These probabilities were calculated through a Gaussian mixture model (GMM). GMM parameters, related to the IPD and ILD models, were initialised and then refined through the expectation-maximisation (EM) algorithm. The method presented in [Sawada et al., 2011] instead, utilised the mixing vector (MV) cue. MV is a vector containing, for each frequency, the time invariant frequency response of the room. [Sawada et al., 2011] was a two-step approach: first they employed the principal component analysis (PCA) to remove additive noise; then, through a Gaussian density function, they clustered the TF bins to determine the probability of each source. The centroids of the clusters were calculated by considering the MV. From this probability, soft TF masks were generated. In [Alinaghi et al., 2014], the two methods firstly proposed in [Mandel et al., 2010] and [Sawada et al., 2011] were combined together. The three cues, i.e. ILD, IPD and MV, were included within a single GMM. In [Deleforge et al., 2015], the STFT of the interaural cues IPD and ILD leaded to a high-dimensional interaural vector for each training data. This data was created recording sources from more than $10^4$ positions around a dummy head. Utilising a manifold learning technique, they employed a non-linear dimensionality reduction to project these vectors into a 2D space [Zhang and Zha, 2005]. The dimensions of this space are the azimuth and elevation DOA of the sound sources. A GMM was used to infer this dimensional transformation. Knowing this transformation, it was then possible to locate a source in space, and consequently, create a binary mask for its recorded binaural spectrogram. However, this method was tested only utilising the same data that was used for the training process.

3.3.2 Source Separation including Early Reflection Information

Every method that has been discussed above considers, to separate sounds from a mixture, only the direct path between source and sensors. However, in real scenarios,
3. Reflector Localisation and its Applications: a Literature Review

sounds interact with the environment during its propagation. Furthermore, information included in the first early reflection is usually important, since it carries only 10-20 dB of energy less than the direct sound [Howard and Angus, 2009]. Therefore, the acoustic multipath should be taken into account during the source separation process [Vincent et al., 2014].

Currently, in the literature, few works can be found that incorporate the multipath propagation into source separation models. In [Huang et al., 2005], the authors proposed a decomposition of the source separation problem into different procedures. The interference was separated from the target source via deconvolving the estimated direct sound RIR segment, then, the same deconvolution approach was applied to each singular echo. However, this method had a high computational cost, thus, in [Rotili et al., 2010] a real-time implementation was presented, where they replaced the inverse filtering, for the deconvolution, with an efficient iterative algorithm. Nevertheless, this method was not robust to low SNR conditions. Similarly, in [Nesta and Omologo, 2012], the authors applied a variation of ICA to estimate the mixing matrix, that is temporal dependent, by considering the multiple source interaction with the environment. However, the permutation problem introduced by ICA was enhanced by the incorrect image source component alignment. Deconvolution of the received signals given an estimation of the RIRs was proposed in [Asaei et al., 2014]. The room geometry of the scene was firstly inferred, by localising the image sources that matched with the temporal support of the recorded RIRs. Nonetheless, with low SNRs, it was hard to estimate the support of the RIRs, and, additionally, the overall model did not consider binaural effects, such as head shadowing.

3.3.3 Literature Limitations

The current literature related to the speech source separation presents several limitations. Some of these limitations have been already discussed in Section 3.3.1, while describing the three main possible approaches to the source separation. It was shown how the TF masking approach is preferable with respect to ICA and beamforming, since it can be applied to underdetermined scenarios, and it is more robust to high level of reverberation.

Looking at the TF approach then, recent methods [Alinaghi et al., 2014, Deleforge et al., 2015, Mandel et al., 2010, Sawada et al., 2011] show better performance with respect to older ones, where anechoic sounds were mainly considered [Yilmaz and Rickard, 2004]. However, also these methods present several limitations. For instance, the work in [Mandel et al., 2010] included into the interaural model only the direct sound information,
without considering effects that are related to the direct sound and early reflection interaction, such as the precedence effect and the comb filter effect. In [Sawada et al., 2011], the TF bins were classified considering the different sources in the mixture. However, this classification was also carried out for those bins that were dominated by reverberation. Since the reverberation is not coherent with the main sounds, this introduced classification errors. The work in [Alinaghi et al., 2014] improved the results with respect to the works in [Mandel et al., 2010] and [Sawada et al., 2011], by combining them. However, by doing so, they did not overcome the respective limitations. By adopting the work proposed in [Deleforge et al., 2015], high quality performance can be achieved only if test and training datasets coincide. This is due to the proposed dimensional reduction that does not take into account possible changes of the environmental geometry.

Methods such the ones described in Section 3.3.2 [Asaei et al., 2014, Huang et al., 2005, Nesta and Omologo, 2012, Rotili et al., 2010] attempted to overcome the problematic fact that, in general, source separation models do not exploit the early reflection information. However, they all followed an approach based on first estimating the RIR, to then deconvolve the target signal. They do not exploit any advantage given by the TF domain. They are, hence, problematic if applied to scenarios with a strong reverberation.

3.4 Fundamental Tools

The novel acoustic reflector localisation methods, that are proposed within this thesis, will be introduced in Chapter 4, whereas its spatial audio and blind source separation applications in Chapters 5 and 6, respectively. Mathematical tools and classical algorithms were exploited during the development of all of these methods. In this section, they are going to be discussed, in order to provide the reader with all the information that is necessary to understand the future concepts.

3.4.1 Epoch Detectors

The word *epoch* is usually used to refer to the instant when a signal starts [Young, 1965]. When a RIR is convolved with a signal produced by a source, this signal is shifted in time at the positions of the direct sound and reflections. Therefore, each reflection (i.e. each peak in the RIR time domain) can be considered as the starting instants of the reflected signals. Thus, in this thesis, the detection of such instants is defined through the words *epoch detection*. The reflection TOA estimation, that will be proposed in Chapters 4, 5 and 6, is done by firstly selecting epochs in the RIR time domain.
In the literature, several algorithms were designed for single channel RIR epoch detection, including some based on spatial pre-processing of high order spherical harmonics [Mabande et al., 2013]. Considering omnidirectional sensors (omnidirectional microphones will be employed during the next chapters), in [Brookes, 1997], the author proposed the so called “FindPeak” algorithm. It was based on the evaluation of the RIR time domain gradient: where a sudden gradient variation was reported, an epoch was detected. Later, an adaptive thresholding technique was employed in [Kuster, 2008]. In particular, the local magnitude mean of the RIR time domain was utilised as threshold. Segments of the signal, having energy magnitude greater than their respective local thresholds, were labelled as epochs. In [Defrance et al., 2008], a matching pursuit based algorithm was employed, by calculating the cross-correlation between the direct sound and the whole RIR. High value of cross-correlation indicated epochs. In [Usher, 2010], the local Kurtosis analysis was used to identify epochs as RIR regions, in time, where the distribution was not normal.

In [Naylor et al., 2007], the dynamic programming projected phase slope algorithm (DYPSA) was designed to estimate glottal closure instances from speech signals. However, in this thesis, it is employed as epoch detector in RIRs, as described in detail in the later chapters. Defining the phase-slope function $S_{gd}(\omega)$ as the opposite of the group delay function of the signal $G_{gd}(\omega) = -S_{gd}(\omega)$, epochs correspond to positive-going zero crossings in $S_{gd}(\omega)$. To reliably select the instants where $S_{gd}(\omega)$ has these zero crossings, $S_{gd}(\omega)$ was smoothed using a Hann window $H_{gd}(n)$ of length $T_{gd}$.

### 3.4.2 Beamformers

Estimation of reflection DOAs will be performed in Chapters 4, 5 and 6 through the employment of beamformers. As already briefly described in Section 3.3.1, beamformers are signal processing tools, that can be used as spatial filters. They select, from a recording, the sound coming to the microphone array from a specific direction, attenuating, instead, all the others [Van Trees, 2002, VanVeen and Buckley, 1988]. Classical beamformers can be divided into two main groups: TDOA-based and eigenspace based.

TDOA based beamformers exploit the TDOA information that is implicit within the recordings. Its principal representative is the delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988]. In DSB, assuming the position of the sensors as known, the array is steered towards a selected direction, by accordingly delaying the received signals. Then, these delayed signals are summed together. The direction producing the highest energy for these summed signals corresponds to the DOA. Expressing this process formulaically, the DOA is calculated as the azimuth $\Theta$ and elevation $\Phi$ that are associated
3.4.3. Linear Predictive Coding

to the delays:

\[ n^{\text{DSB}} = \arg\max \left[ \sum_{i=1}^{M} \sum_{n} y_i(n - n_i^{\text{DSB}})^2 \right], \quad (3.1) \]

where \( n^{\text{DSB}} \) is a vector containing the set of delays that maximise the function above, \( M \) is the number of sensors in the array, \( n \) is the discrete time variable, and \( y_i(\cdot) \) is the signal received at the \( i \)-th sensor position.

The second group is represented by those DOA estimators that evaluate the eigenspace of the received signal, by calculating the eigenvectors of the covariance matrix obtained from the signals received. Example of these algorithms are: multiple signal classification (MUSIC) [Schmidt, 1986]; estimation of signal parameters via rotational invariance techniques (ESPRIT) [Roy and Kailath, 1989]; the Capon’s beamformer [Capon, 1969]; the Bartlett beamformer [Bagheroei et al., 1988].

Recently, spherical microphone arrays, are also employed to evaluate the acoustic properties of environments. New beamforming methods are, thus, proposed in that sense, and they are usually based on the spherical harmonic decomposition [Farina and Tronchin, 2013, Sun et al., 2012, Tervo and Politis, 2015].

3.4.3 Linear Predictive Coding

The frequency domain analysis that will be performed to the RIRs in Chapter 5 is also based on the linear predictive coding (LPC). LPC generates an all-pole filter \( a(n) \), which approximates the spectrum of the arbitrary signal under investigation \( d(n) \) [Makhoul, 1975]. An all-pole model spectrum is defined as:

\[ A(\omega) = \frac{G_a^2}{1 + \sum_{b=1}^{K} q_b \exp(-jb\omega)^2}, \quad (3.2) \]

where \( q_b \) is the \( b \)-th coefficient of the all-pole filter, \( K \) is the order of the all-pole filter, and \( G_a \) the gain. Defining the power spectrum of the signal \( d(n) \) as \( D_{ps}(\omega) \), the approximation error function is defined as the ratio between \( D_{ps}(\omega) \) and \( A(\omega) \) [Makhoul, 1975]:

\[ E(\omega) = D_{ps}(\omega)|1 + \sum_{b=1}^{K} q_b \exp(-jb\omega)|^2. \quad (3.3) \]

Therefore, the total error can be written as:

\[ E_{\text{tot}} = \frac{G_a^2}{2\pi} \int_{-\pi}^{\pi} \frac{D_{ps}(\omega)}{A(\omega)} d\omega. \quad (3.4) \]
Minimising \( E_{\text{tot}} \) is equivalent to minimise the integrated ratio of the signal spectrum \( D_{\text{ps}}(\omega) \) to its approximation \( A(\omega) \). For this reason, a better definition of the LPC can now be provided as: “Given some spectrum \( D_{\text{ps}}(\omega) \), we wish to model it by another spectrum \( A(\omega) \) such that the integrated ratio between the two spectra as in Equation 3.4 is minimised” [Makhoul, 1975].

To obtain the spectrum \( A(\omega) \), both the parameters \( q_b \) and the gain \( G_a \) must be calculated. Minimising \( E_{\text{tot}} \) by setting \( \frac{\partial E}{\partial q_u} = 0 \), with \( 1 \leq u \leq K \), a set of equations are obtained, commonly named as normal equations. Assuming then \( E_{\text{tot}} \) to be minimised over the infinite duration (i.e. for \( -\infty \leq n \leq \infty \)), the normal equations can be written following the autocorrelation function \( Q(u) \) of \( d(n) \):

\[
Q(u) = -\sum_{b=1}^{K} q_b Q(u-b).
\] (3.5)

By solving these equations the \( q_b \) values that minimise \( E_{\text{tot}} \) are obtained. On the other hand, the gain factor \( G_a \) is obtained by equating the total energy of the two spectra. In other words, \( G_a \) is obtained by equating \( Q(u) \) to \( Q_a(u) \), where the latter is the all-pole model autocorrelation function [Makhoul, 1975].

3.4.4 Homogeneous Coordinates

Similar to what was done in [Antonacci et al., 2012], the geometric approach that will be undertaken in Chapter 4, by generating ellipsoids to estimate the reflector position, exploits the so called homogeneous coordinates.

Points in a 3D Euclidean space are represented, in the Cartesian coordinate system, by a vector of three real numbers, \( \mathbf{X} = [x, y, z] \). By utilising this coordinate system, it is usually assumed, that two parallel planes meet at infinity. However, this assumption conflicts with the known dictum that infinity does not exist [Hartley and Zisserman, 2003]. It is possible to bypass this incongruity by enhancing the Euclidean space. This can be done by adding the points at infinity, and defining them as ideal points. Through this process, the Euclidean space \( \mathbb{R}^3 \) is projected to another space, named projective space \( \mathbb{P}^3 \). The new system has four coordinates that represent the same point, but in the projective space, and they can be written as \( \mathbf{X}_{\text{hom}} = [z_2x, z_2y, z_2z, z_2] \) \( (z_2 \neq 0 \) represents the point at infinity). These are named as homogeneous coordinates of a point. Given the four coordinates in \( \mathbf{X}_{\text{hom}} \), it is possible to revert them to the original Cartesian coordinates \( \mathbf{X} \), characterising the \( \mathbb{R}^3 \) space, by dividing them for \( z_2 \), and removing the fourth coordinate. For simplicity, in the literature [Hartley and Zisserman, 2003], and during the rest of this thesis, \( z_2 \) is set to be equal to 1.
Geometrical figures can be defined in the homogeneous coordinates. Planes and quadratic surfaces are figures of main interest for the rest of this thesis, hence, they are going to be defined below. A Plane in $\mathbb{R}^3$ may be written as:

$$p_1 x + p_2 y + p_3 z + p_4 = 0,$$

where $p_1$, $p_2$, $p_3$, and $p_4$ are the plane coefficients. The homogeneous representation of the plane is a 4D vector $\mathbf{p} = [p_1, p_2, p_3, p_4]^T$. Therefore, in the homogeneous coordinates, Equation 3.6 can be rewritten as [Hartley and Zisserman, 2003]:

$$\mathbf{p}^T \mathbf{X}_{\text{hom}} = 0.$$  \hspace{1cm} (3.7)

Similarly, a quadratic surface in $\mathbb{P}^3$ is defined as:

$$\mathbf{X}_{\text{hom}}^T \mathbf{E} \mathbf{X}_{\text{hom}} = 0,$$  \hspace{1cm} (3.8)

where $\mathbf{E}$ is a symmetric $4 \times 4$ matrix, containing the quadratic surface coefficients. It is interesting to note that when $\mathbf{E}$ is a diagonal matrix, with diagonal $[1, 1, 1, -1]$, it represents a unitary sphere centred at the centre of the coordinate system [Hartley and Zisserman, 2003].

### 3.4.5 Linear Transformations

Following general computer graphics definitions, set of points in the 3D space can be transformed, in order to position, shape and rotate the objects that they represent. To do so, Linear transformations are usually utilised, since they preserve the vector addition and the scalar multiplication properties. Three of the most important linear transformations are: translation, rotation and scaling. An example for the translation and rotation transformations is reported in Figure 3.5. Proceeding with defining the
space through homogeneous coordinates, these transformations can be seen as $4 \times 4$ matrices.

Movement of points, in the homogeneous coordinates, from one location to another is represented by a translation matrix, that is defined as [Akenine-Moller et al., 2008]:

$$
T = \begin{bmatrix}
1 & 0 & 0 & \Delta x \\
0 & 1 & 0 & \Delta y \\
0 & 0 & 1 & \Delta z \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

where $\Delta x$, $\Delta y$ and $\Delta z$ are the translation coefficients for the x-, y- and z-axis, respectively. Considering the plane $p$ that was defined in Equation 3.7, the set of points describing it can be translated by factors $\Delta x$, $\Delta y$ and $\Delta z$, by performing the multiplication $Tp$. An important property of $T$ is that the inverse of it (i.e. $T^{-1}$) is given by the same matrix with the three coefficients having the opposite sign (i.e. $-\Delta x$, $-\Delta y$ and $-\Delta z$) [Akenine-Moller et al., 2008].

A rotation transform rotates the vector representing a point in the space by a given angle, around a given axis. Rotation matrices that are widely used in computer graphics are defined as [Akenine-Moller et al., 2008]:

$$
R_x = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & \cos(\alpha) & -\sin(\alpha) & 0 \\
0 & \sin(\alpha) & \cos(\alpha) & 0 \\
0 & 0 & 0 & 1
\end{bmatrix},
$$

$$
R_y = \begin{bmatrix}
\cos(\beta) & 0 & \sin(\beta) & 0 \\
0 & 1 & 0 & 0 \\
-\sin(\beta) & 0 & \cos(\beta) & 0 \\
0 & 0 & 0 & 1
\end{bmatrix},
$$

$$
R_z = \begin{bmatrix}
\cos(\gamma) & -\sin(\gamma) & 0 & 0 \\
\sin(\gamma) & \cos(\gamma) & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix},
$$

where $\alpha$,$\beta$ and $\gamma$ are the angles of rotation for the x-, y-, and z-axis, respectively. These three matrices can be combined together, in order to generate a unique rotation matrix, by multiplying them as: $R = R_x R_y R_z$. Similar to the translation matrix, also the inverse of $R$ has the peculiar property of corresponding to the same matrix, but being characterised by opposite rotation angles (i.e. $-\alpha$, $-\beta$ and $-\gamma$).
A scaling matrix is instead defined to scale an entity with factors $Q_x^{\text{scal}}$, $Q_y^{\text{scal}}$, and $Q_z^{\text{scal}}$, along the $x$, $y$ and $z$ directions, respectively. Thus, a scaling matrix is commonly used to enlarge or shrink objects in the 3D space. It is defined as [Akenine-Moller et al., 2008]:

$$
S = \begin{bmatrix}
Q_x^{\text{scal}} & 0 & 0 & 0 \\
0 & Q_y^{\text{scal}} & 0 & 0 \\
0 & 0 & Q_z^{\text{scal}} & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}.
$$

(3.11)

The multiplication between matrices does not follow the commutative property: the concatenation of transforms is order-dependent [Akenine-Moller et al., 2008]. However, concatenating a sequence of matrices is important to gain efficiency. For instance, if one wants to apply transformations of translation, rotation and scaling, all to the same set of points, it is easier to concatenate all the matrices, instead of multiplying the set by each of them. In general, the scaling matrix should be applied first, hence, appearing at the right of the composition. Referring again to the previous example, a plane $p$ can be translated, rotated and scaled through the order-dependent multiplication: $\mathbf{T} \mathbf{R} \mathbf{S} \mathbf{p}$ [Akenine-Moller et al., 2008]. Similarly, a quadratic surface can be translated, rotated and scaled, in the homogeneous coordinates, through the multiplication: $\mathbf{T}^T \mathbf{R}^T \mathbf{S}^T \mathbf{E} \mathbf{S} \mathbf{R} \mathbf{T}$.
Chapter 4

Acoustic Reflector Localisation given Room Impulse Responses

The creation of an accurate model to identify acoustic reflector positions from room impulse responses (RIRs) is important for several different research areas in audio signal processing. For instance, as it was already outlined in Chapter 3, such a model can provide information about the geometry of an environment, with respect to a listening position, which can be exploited in audio forensics [Malik, 2013], simultaneous localisation and mapping [Dokmanić et al., 2016], and spatial audio [Zotkin et al., 2004]. In addition, acoustical parameters can be estimated to enhance target signals, in fields such as automatic speech recognition [Yoshioka et al., 2012], music transcription [Plumbley et al., 2002], source separation [Asaei et al., 2014], audio tracking [Öcal et al., 2014], dereverberation [Naylor and Gaubitch, 2010], and microphone array raking [Dokmanić et al., 2015b, Javed et al., 2016]. Beyond all of these, an acoustic reflection localiser can be combined with image processing methods, to construct a hybrid room geometry estimation model [Hussain et al., 2014, Ye et al., 2015].

4.1 Introduction

Several approaches have been presented to localise reflectors from RIRs, as solution of a 2D problem, where loudspeaker, microphone and reflector lie on the same plane [Antonacci et al., 2012, Canclini et al., 2011a,b, Dokmanić et al., 2011, Marković et al., 2013b,a, Moore et al., 2013]. For instance, in [Antonacci et al., 2012], the authors exploited time of arrival (TOA) knowledge to localise 2D reflectors, constructing ellipses. These ellipses had their major axis equal to the reflection path length, and foci on the
respective microphone and source positions. The line that was common tangent to all
the ellipses corresponded to the 2D reflector under investigation.

Recently, 3D models were also investigated. In [Filos et al., 2012], the work in [Antonacci
et al., 2012] was extended to 3D spaces. However, it was not yet a full 3D estimation,
instead combining 2D projections to estimate the positions of the surfaces outside the
measurement planes. As already discussed in Chapter 3, the 3D reflector localisation
methods can approximately be categorised into two groups. The first one is the image-
source reversion [Arteaga et al., 2013, Dokmanić et al., 2013, Ribeiro et al., 2012, Tervo
and Tossavainen, 2012], where the TOA is used to revert to the image source location
[Allen and Berkley, 1979], which is subsequently used to determine the reflector posi-
tion. The second group contains, instead, those methods directly localising the reflector,
without first estimating any other feature related to the room acoustic [Kuster et al.,
2004, Nastasia et al., 2011, Zamaninezhad et al., 2014]. Accordingly, this group is named
as direct localisation.

The method in [Tervo and Tossavainen, 2012] used the image-source reversion approach
to localise reflectors in 3D. However, a large number of loudspeaker orientations and
putative image source positions were needed to test. In [Ribeiro et al., 2012], a least-
squares (LS) minimisation was utilised to fit “synthetic” reflections to recorded RIRs.
However, it required a large number of RIRs. The synthetic reflections were obtained
in an anechoic room, with a plastic frame to simulate a wall. In total, 240 loudspeaker
positions were used to collect the recordings, leading to 1440 RIRs, considering the
six-element microphone array employed. Since the number of RIRs was deemed not
enough to train the model, the RIRs were additionally interpolated in space, in order to
have more directions of arrival (DOAs). In [Arteaga et al., 2013], acoustic parameters
were calculated from a single recorded RIR to estimate different image source positions.
Synthetic RIRs were generated from these image sources, and compared to the recorded
one. The most likely room shape was obtained by choosing the best fitting synthetic
RIR. The interesting part of this method was the estimation done using only one single
recorded RIR. However, it was not robust with respect to errors. The main contribution
of [Dokmanić et al., 2013] was an algorithm to label the reflections from a distributed
microphone array, where the reflector order would otherwise be ambiguous if compared
among different microphone recordings. The reflector estimation was performed using
image-source reversion, by assuming that the TOAs were known a priori.

One of the first attempts to employ, instead, the direct localisation approach was pro-
posed in [Kuster et al., 2004]. Exploiting inverse wave field extrapolation, the authors
mapped reflections from a set of receivers to the related reflecting objects. This method
provided an accurate analysis, and it was even able to identify small reflecting objects
lying between the main reflector and the microphone array. However, the microphone array was assumed to be exactly parallel to the reflector. In [Zamaninezhad et al., 2014], reflector positions were estimated by transforming the recorded RIR into the frequency domain. In spite of this, only two parallel reflectors were localised, assuming the other boundaries as completely absorbent. The method that can be considered as the first one exploiting direct localisation through 3D surfaces, was presented in [Nastasia et al., 2011]. Nevertheless, the peaks were not automatically identified to extract TOAs from RIRs. In addition, the reflector search was computationally expensive, caused mainly by the optimisation of its cost function.

Various limitations can be observed in the acoustic reflector localisation literature. First, there is not an epoch detection algorithm that is specifically designed to automatically extract reflection TOAs from multichannel RIRs, recorded by compact microphone arrays, although several algorithms are available in the literature for single channel epoch detection on RIRs [Brookes, 1997, Defrance et al., 2009, Kuster, 2008, Usher, 2010]. Second, some methods assume spatial relationships between the microphone array and the reflectors [Kuster et al., 2004, Zamaninezhad et al., 2014]. Third, other methods may require thousands of RIRs recorded in anechoic rooms [Ribeiro et al., 2012]. Fourth, the microphones are often considered to be spatially sparse, introducing the problem of labelling the echoes [Dokmanić et al., 2013]. Fifth, there are no proposed ways to aggregate measurements from multiple loudspeakers to improve localisation. Finally, classical image-source reversion methods (e.g. [Dokmanić et al., 2013, Tervo and Tossavainen, 2012]) use TOAs to localise the image source without considering any other information carried by the RIRs, thus, limiting their robustness.

To address these issues, the contributions of this chapter are:

- a multichannel version of DYPSA [Naylor et al., 2007], i.e. the clustered DYPSA (C-DYPSA), to automatically extract reflection TOAs from compact microphone array RIRs;
- the image-source reversion method ISDAR-LIB, created by the fusion of the novel image source direction and ranging (ISDAR) algorithm and the loudspeaker-image bisection (LIB) (a reflector localisation algorithm firstly appeared in [Tervo and Tossavainen, 2012]);
- two further novel variants of ISDAR-LIB, exploiting multiple loudspeakers;
- the ellipsoid tangent sample consensus (ETSAC), a direct localisation method;
- a comparative evaluation of the state-of-the-art and the proposed methods, using synthetic and measured RIRs.
The comprehensive comparison presented here is, to the author knowledge, the first that compares image-source reversion and direct localisation methods, as approaches for 3D acoustic reflector localisation. The study also informs the level of estimation accuracy expected from a real-world dataset.

The rest of the chapter is organised as follows: Section 4.2 introduces the underlying theory supporting the presented methods, and the pre-processors that is in common to every method evaluated. The state-of-the-art methods selected for the evaluation are described in Section 4.3, and the proposed methods in Section 4.4. The numerical analysis and results are reported in Section 4.5. Section 4.6 draws the overall conclusions.

4.2 Background and Preliminaries

4.2.1 The Room Impulse Response (RIR)

A RIR is an acoustic signal, carrying information about the environment where it was recorded. It is generally considered as being composed of three parts [Kuttruff, 2009]: the direct sound, revealing the position of the sound source; the early reflections, conveying a sense of the environmental geometry; and the late diffuse reverberation, indicating the size of the environment, and average absorption [Blesser, 2001, Välimäki et al., 2012]. From this classical decomposition, and by defining the discrete time variable $n$, the RIR from source $l$ to sensor $i$ can be defined as superimposition of Dirac deltas, delayed by $n_{e,i,l}$ samples, with $e$ enumerating the reflections (see also Equation 2.47):

\[
I_{i,l}(n) = h_{0,i,l}(n - n_{0,i,l}) + \sum_{l=1}^{T_m} h_{e,i,l}(n - n_{e,i,l}) + w_R(n) + w(n),
\]

(4.1)

where $h_{0,i,l}(n)$ represents the direct sound, $h_{e,i,l}(n)$ the early reflections, and $w_R(n)$ is the late reverberation modelled as exponentially decaying Gaussian noise; $T_m$ is the $e$-th peak corresponding to the last reflection before the mixing time, and $w(n)$ is the additive Gaussian noise.

4.2.2 The Image Source Model

The most prominent early reflections typically have a sizeable specular component. Therefore, one way to localise reflectors is to benefit from the notion of image sources [Allen and Berkley, 1979]. Denoting $B_0 = [b_{x,0}, b_{y,0}, b_{z,0}]^T$, where $[\cdot]^T$ stands for the transpose operation, as a vector containing the three Cartesian coordinates of the sound source and $p[p_1, p_2, p_3, p_4]^T$ as a vector containing the first reflector plane coefficients,
the corresponding image source $B$ is the virtual sound source, constructed as the point $B_0$ mirrored with respect to $p$. The reflector considered in this chapter will always be the one corresponding to the first reflection arriving to the microphone array. Therefore, for simplicity, the index $e = 1$ will be omitted from any variable defined.

It has been already mentioned that the focus of this chapter is the first early reflection, being the reflection carrying most of the energy and having the best perceptual properties [Bech, 1998, Howard and Angus, 2009]. Usually, early reflections are assumed to be completely specular reflections; therefore, it is possible to benefit from the notion of image sources [Allen and Berkley, 1979, Kuttruff, 2009]. Knowing the position in space of $B_0$ and the four coefficients characterising $p$, the related image source can be calculated as:

$$B = B_0 - 2n(nB_0),$$

(4.2)

where $n$ is the vector normal to the plane, defined as $n = [p_1, p_2, p_3]^T$.

### 4.2.3 Method Classification and Overviews

As discussed in Section 4.1, reflector localisation methods can be divided into two main groups: image-source reversion and direct localisation. Table 4.1 summarises these groups and shows the categorisation for the proposed methods, together with the state-of-the-art. Figure 4.1 shows an overview of the structure of the proposed methods, together with two baseline methods (selected to be compared in the experiments in Section 4.5): Tervo et al. [Tervo and Tossavainen, 2012] and Dokmanić et al. [Dokmanić et al., 2013]. The novel methods and algorithms are highlighted in grey (C-DYPSA, ISDAR-LIB, the ISDAR-LIB variants, and ETSAC). To generate methods that are able to automatically extract TOAs from RIRs, C-DYPSA is proposed and employed as a pre-processor to each method tested. A description of this novel algorithm, that is an evolution of the DYPSA algorithm [Naylor et al., 2007], is reported in Section 4.2.5. Different acoustic parameters are exploited by the methods. ISDAR-LIB and its variants exploit DOA, together with TOA, to localise the reflector. Thus, for these methods, the delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988] is employed during the pre-processing stage, together with C-DYPSA.

<table>
<thead>
<tr>
<th>Image-Source Reversion</th>
<th>Direct Localisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Ribeiro et al., 2012]</td>
<td>[Kuster et al., 2004]</td>
</tr>
<tr>
<td>[Tervo and Tossavainen, 2012]</td>
<td>[Zamaninezhad et al., 2014]</td>
</tr>
<tr>
<td>[Dokmanić et al., 2013]</td>
<td>[Nastasia et al., 2011]</td>
</tr>
<tr>
<td>Proposed ISDAR-LIB and variants</td>
<td>Proposed ETSAC</td>
</tr>
</tbody>
</table>
4.2.4 Method Assumptions

The main assumptions in this chapter are as follows:

- knowledge of at least four RIRs;
- the omnidirectional microphone array is *compact*;
- the sources are in the far-field;
- the reflection has a dominant specular component;
- the image source estimation errors are independent and identically distributed.

The assumption concerning the minimum number of RIRs was made due to the fact that, to estimate parameters in the 3D space, at least four positions are needed. The compact microphone array assumption was made to enable the use of classical beam-forming techniques [VanVeen and Buckley, 1988], and avoid erroneous permutations in the labelling of reflections arriving at the microphones [Dokmanić et al., 2013]. Arrays with a maximum microphone displacement from the array centre less than half a 1 kHz wavelength \((d_{cm} < 171\text{ mm})\) are considered to be compact, where 1 kHz is a standard
reference frequency. Note that compactly-arranged microphones are similarly affected by any source directivity. As it was discussed in Chapter 2, by assuming sources and image sources to be in the far-field means that their response at the array can be approximated as plane waves. For the array configuration in the present work, the Fraunhofer rule sets the far-field limit at 2.1 m at 1 kHz [Balanis, 2005], making this a fair assumption for reflector localisation in typical rooms. The fourth assumption, of the specular reflections, enables its detection in the time domain RIR. This assumption excludes scattering, shadowing and diffraction phenomena, and justifies the use of the image source model, which applies to mid-range audio frequencies. No further assumptions regarding reflection, source, and microphone frequency responses are needed. Finally, assuming the image source localisation errors as independent and identically distributed allows the integration of multiple loudspeaker results in a post-processing step. Different reflection signal-to-noise ratios (SNRs) do not influence this, since the dominant specular component of the reflection implies a high SNR.

The first reflection can be considered as the most important one, for two main reasons: subject to masking, it has the most prominent perceptual properties, being a single specular reflection arriving with a limited time delay from the direct sound [Bech, 1998, Lokki et al., 2011]; among the early reflections, it is usually the one that carries the most amount of energy [Howard and Angus, 2009]. In a typical room, the second reflection conveys 20–40 dB less energy than the direct sound. On the other hand, the first reflection is energetically prominent, thus, it modifies the colouration perception and the spatial impression of the produced sound [Barron, 1971]. For these reasons, in this chapter, the focus is the first reflector. Other works that examine later reflections can be found in [Dokmanić et al., 2013, Tervo and Tossavainen, 2012]. Since the first reflection is estimated from multichannel RIRs only, no prior assumption has to be made about the room shape or the reflector orientation.

### 4.2.5 Common Pre-Processing

To obtain the reflection TOAs and DOAs automatically from recorded RIRs, a pre-processing stage was employed consisting of a clustering epoch detector (i.e. C-DYPSA) and a DSB. An overview of the pre-processing steps are depicted in Figure 4.2.
4.2.5. Common Pre-Processing

The Clustered Dynamic Programming Projected Phase-Slope Algorithm (C-DYPSA). The state-of-the-art reflector localisation methods do not provide a specific algorithm to extract TOAs from the multichannel RIRs. Here, an extension of DYPSA [Naylor et al., 2007] is proposed, that clusters TOAs across the microphones of a compact array, to find the reflection TOAs.

DYPSA was initially designed to estimate glottal closure instances from speech [Naylor et al., 2007]. There, the signal that was processed was the estimated vocal tract impulse response, therefore, a similar signal to the RIRs that are under investigation in this thesis. The whole concept of DYPSA is based on assuming an impulse response as a minimum phase signal. For a minimum phase signal, the average of $S_{gd}(\omega)$, that is the slope of the phase spectrum, is zero [Smits and Yegnanarayana, 1995]. The shifted version of a minimum phase signal has a phase spectrum similar to the original but with an average slope proportional to the shift. This means that the average slope of the phase spectrum is dictated by the location of the excitation impulse [Smits and Yegnanarayana, 1995]. Therefore, a RIR is segmented in its time domain with a number of segments equal to the number of samples, and the average phase slope is calculated for each of them. Peaks that occur in the time domain correspond to sudden changes in the average of $S_{gd}(\omega)$. Thus, positive-going zero crossings of this average correspond to the position of the peaks in the signal time domain, as shown in Figure 4.3.

To reliably select the instants where $S_{gd}(\omega)$ has these zero crossings, $S_{gd}(\omega)$ was smoothed using a Hann window $H_{gd}(n)$ of length $T_{gd}$. To adapt the algorithm to this chapter purpose, a threshold $\tau_s$ was defined on $S_{gd}(\omega)$ to take only the most significant peaks of $I_{i,l}(n)$. Another threshold $\tau_A$ was applied on the time domain amplitude, to eliminate the peaks that have much less energy than the main one (i.e. the direct sound). These thresholds were empirically derived.

The proposed C-DYPSA contains two post-processing refinements to DYPSA. First, considering the $e$-th peak for every $i$-th microphone in the compact array, the median $\tilde{n}_{e,i}$
of the estimated TOAs in samples \( \hat{n}_{e,i,l} \) is obtained. The output of DYPSA, considering each \( i \)-th sensor separately, is then observed. If the \((e+1)\)-th reflection TOA \( \hat{n}_{e+1,i,l} \) is closer to the median \( \tilde{n}_{e,l} \) than the \( e \)-th reflection TOA \( \hat{n}_{e,i,l} \), then \( \hat{n}_{e,i,l} \) is treated as a false positive. Consequently, it is replaced with the \((e+1)\)-th reflection TOA value \( \hat{n}_{e+1,i,l} \). Second, the Grubbs’ test [Grubbs, 1969] iteratively identifies the cluster of TOAs related to the \( e \)-th peak considering every microphone in the compact array (see Figure 4.4). The RIRs generating outliers to this cluster are discarded.

**Delay-and-Sum Beamformer (DSB).** The image-source reversion methods, that are proposed in this chapter, also exploit DOAs. To extract them directly from recorded RIRs, the DSB [VanVeen and Buckley, 1988] was used, providing adequate performance for this chapter purposes, and being simple. To apply DSB, the input RIRs were first segmented, as shown in Figure 4.2. To generate these segments without losing the phase differences, the average of the first early reflection TOAs \( \hat{n}_{i,l} \) over the \( M \) microphones was calculated as \( \overline{n}_l = \frac{1}{M} \sum_{i=1}^{M} \hat{n}_{i,l} \). This TOA corresponds to that of a virtual microphone lying at the centre of the array. The segments were obtained by applying a Hamming window \( H_{\text{seg}}(n) \) of length \( T_{\text{seg}} \), for each RIR, centred at \( \overline{n}_l \): \( \overline{I}_{i,l}(n) = I_{i,l}(n) H_{\text{seg}}(n - \overline{n}_l) \).

### 4.3 State-of-the-Art Methods

The two image-source reversion baselines [Dokmanić et al., 2013, Tervo and Tossavainen, 2012], based on single loudspeaker information, are presented in this section. First, their two distinct image source locator algorithms are described. Then, their common reflector position estimation algorithm LIB is presented.
4.3.1 Maximum Likelihood (ML)

The method proposed by Tervo et al. in [Tervo and Tossavainen, 2012] is composed by a maximum likelihood (ML) algorithm to localise the image source, followed by the LIB algorithm to estimate the reflector position (Figure 4.1 (A)).

The ML image source locator exploits the TOAs related to each microphone to generate a probability function [Tervo and Tossavainen, 2012]. First, a uniformly distributed set of \( x^{ML} \) points is generated in the 3D space. These points are represented by the \( x^{ML} \times 3 \) dimensional matrix \( X^{ML} \) containing all the Cartesian coordinates. Considering these points as possible image source positions, and assuming the centre of the microphone array as the origin of the coordinate system, the possible TOAs for the first reflection of the \( l \)-th loudspeaker are obtained, and placed in the vector \( n^{ML}_{l}(X^{ML}) = [n^{ML}_{1,l}(X^{ML}), ..., n^{ML}_{M,l}(X^{ML})]^{T} \), where \( M \) is the number of microphones.

With the TOAs estimated through C-DYPSA \( \hat{n}_l = [\hat{n}_{1,l}, ..., \hat{n}_{M,l}]^{T} \), the multivariate Gaussian probability distribution function can be calculated as:

\[
p^{ML}(E^{ML}_l(X^{ML}), \Sigma) = \frac{\exp\left(-\frac{1}{2} E^{ML}_l(X^{ML})^{T} \Sigma^{-1} E^{ML}_l(X^{ML})\right)}{(2\pi)^{M/2} \sqrt{\det(\Sigma)}},
\]

(4.3)

where \( E^{ML}_l(X^{ML}) = n^{ML}_l(X^{ML}) - \hat{n}_l \), and \( \Sigma = [\text{diag}(J_{i,l})]^{-1} \). \( J_{i,l} \) is the Fisher information, carrying knowledge of the selected frame \( I_{i,l}^{S}(n) \) signal to noise ratio (SNR) [Tervo et al., 2012]. Thus, the image position is given by:

\[
B_l = \arg\max_{X^{ML}} p^{ML}(E^{ML}_l(X^{ML}), \Sigma).
\]

(4.4)

4.3.2 Multilateration

The method presented by Dokmanić et al. [Dokmanić et al., 2013] employs the image source localisation method named as multilateration\(^1\). In addition, the LIB algorithm, which will be introduced in Section 4.3.3, is used as reflector locator (Figure 4.1 (A)).

Having knowledge from C-DYPSA about the first reflection TOAs \( \hat{n}_{i,l} \), and assuming the vector containing the microphone position coordinates \( A_i \) as known, the multilateration generates spheres having radii equal to the TOAs \( \hat{n}_{i,l} \) and centred at the respective sensor positions \( A_i \). Minimising a particular cost function which incorporates each reflection distance [Beck et al., 2008], the image source \( B_l \) related to the \( l \)-th loudspeaker is obtained. However, with traditional multilateration, if microphones were too close to each other, \( B_l \) could not always be localised. Due to small errors during the TOA estimation,

\(^1\)Multilateration is not explicitly presented in [Dokmanić et al., 2013]. However, it can be identified from the authors’ code at http://infoscience.epfl.ch/record/186657/files/.
there were cases where the spheres did not intersect. Therefore, being unreliable with compact microphone arrays, the method was modified to randomly select three spheres, finding the point $B_{l,g}$, where $g$ indicates the selected three-microphone combination, and testing it for 100 combinations. When the algorithm fails, the combination is discarded. Thus, $x_{\text{mul}} \leq 100$ potential image sources are found. The image position is taken as the mean over all the valid combinations: $B_l = \frac{1}{x_{\text{mul}}} \sum_{g=1}^{x_{\text{mul}}} B_{l,g}$.

4.3.3 The Loudspeaker-Image Bisection (LIB) Algorithm

The LIB algorithm was employed to localise the reflector by both [Tervo and Tossavainen, 2012] and [Dokmanić et al., 2013], as shown in Figure 4.1 (A). This algorithm was already briefly described in Section 3.1.1. It is based on the image source model, and it characterises the majority of the image source reversion methods for acoustic reflector localisation. It reverts the image source concept that was firstly proposed in [Allen and Berkley, 1979], by utilising the image source position information to then estimate the reflector location.

In other words, the plane $p_l$, defining the reflector, can be seen as the one bisecting the line $l_l$ from the $l$-th loudspeaker $B_{0,l}$ to the image $B_l$. Their midpoint $M_l$ lies on the plane. In order to find $p_l$, first, the unit vector normal to $p_l$ is defined as:

$$u_l = \frac{B_{0,l} - B_l}{\|B_{0,l} - B_l\|} = [u_{1,l}, u_{2,l}, u_{3,l}]^T,$$

where $\| \cdot \|$ stands for the Euclidean norm. Consequently, $p_l$ is defined in homogeneous coordinates as:

$$p_l = [u_l^T, -M_l^T u_l]^T,$$
where the midpoint is:

\[
M_t = \frac{B_{0,t} + B_t}{2} = [M_{1,t}, M_{2,t}, M_{3,t}]^T.
\]  

(4.7)

A schematic representation of this algorithm is depicted in Figure 4.5.

4.4 Proposed Methods

4.4.1 Image-Source Reversion Methods

The proposed method ISDAR-LIB (Figure 4.1 (A)) utilises the same algorithm as in [Tervo and Tossavainen, 2012] and [Dokmanić et al., 2013] for the reflector estimation part (i.e. LIB), together with the novel image source locator ISDAR.

The two novel ISDAR-LIB variants (Figure 4.1 (B)) are mean-ISDAR-LIB and median-ISDAR-LIB. Their main novelty is the integration of multiple loudspeakers. The “multiple loudspeaker combination” block, in Figure 4.1 (B), represents mean and median, respectively. These two averages provide insight to the error types that most degrade localisation performance: median is more robust to outliers rejecting all samples except the central one; whereas the mean is more robust to additive noise, reducing noise variance by \(L\) for \(L\) estimates.

**Image Source Direction and Range (ISDAR) - LIB Method.** The selected baselines exploited information given by TOAs only. To improve the image source localisation part, it was necessary to introduce an algorithm that exploits information from both TOAs (from C-DYPSA) and DOAs (from DSB). ISDAR is based on the idea of combining TOA and DOA to localise the image sources with the following novel aspects. The first aspect is the use of a compact array of non-coincident omnidirectional microphones. The second aspect concerns the TOA estimation. In the here proposed approach, TOAs are estimated for each microphone channel by DYPSA [Naylor et al., 2007], and then clustered. This allows for the development of more robust algorithms, by detection and correction of gross errors, such as the proposed C-DYPSA. Correlations between each pair of microphone RIRs were utilised in [Tervo et al., 2013], where a squared-error cost function was then minimised to find DOAs from estimated time differences of arrival (TDOAs). Third, in [Tervo and Politis, 2015], a probabilistic approach was proposed to find DOA by steering towards the signals recorded through spherical microphone arrays. In case of the proposed ISDAR, the DSB, which is a TDOA-based approach, was employed to obtain the DOA as that giving the maximum response. Finally, whereas direct sound cancellation can avoid swamping the reflection signal [Robinson and Xiang, 2010],
Algorithm 4.1 The ISDAR-LIB method

| **Input** | TOAs $\pi_l$ and DOAs $\Gamma_l = [\Theta_l, \Phi_l]$ w.r.t. source $B_{0,l}$ and the microphone array centre; source position |
| **Output** | Plane $p_l$ (reflector) |

/* ISDAR */
1: Calculate the radial distance $\rho_l$ from $n_l$
2: Localise the image source $B_l$ through Equation (4.8)
/* LIB */
3: Calculate the unit vector $u_l$ through Equation (4.5)
4: Calculate the midpoint $M_l$ of $B_{0,l}$ and $B_l$ (Equation (4.7))
5: Calculate the position of $p_l$ through Equation (4.6)

for ISDAR a time window is applied around the reflection TOA, to gate the reflection signal and extract the related time domain segment, as in [Tervo and Politis, 2015]. Methods to localise sources that combine TOA with TDOA can be also found in the literature. For instance, in [Tervo et al., 2012], the authors provided an evaluation over several localisation methods, exploiting the TOA and TDOA probability density functions. Then, they fused together these two densities, improving robustness. However, with ISDAR, TOA and DOA are directly combined, in spherical coordinates.

Given the radial distance $\rho_l = \frac{\pi_l c_0}{f_s}$, where $c_0$ is the sound speed and $f_s$ the sampling frequency, the image position $B_l = [b_{x,l}, b_{y,l}, b_{z,l}]^T$ can be written as:

\[
\begin{align*}
    b_{x,l} &= \rho_l \cos(\Theta_l) \cos(\Phi_l), \\
    b_{y,l} &= \rho_l \sin(\Theta_l) \cos(\Phi_l), \\
    b_{z,l} &= \rho_l \sin(\Phi_l),
\end{align*}
\]

where $\Theta_l$ and $\Phi_l$ are the azimuth and elevation, respectively.

The reflector locator exploited by ISDAR-LIB is the same as the one utilised by the image-source reversion baselines [Tervo and Tossavainen, 2012], [Dokmanić et al., 2013], i.e. the LIB algorithm, already described in Section 4.3.3. Therefore, the plane estimated is $p_l$, from Equation (4.6). The pseudocode of ISDAR-LIB is reported in Algorithm 4.1.

**Mean-ISDAR-LIB.** To improve the results provided by ISDAR-LIB, the information about the reflector position carried by multiple loudspeakers can be exploited. For this reason, mean-ISDAR-LIB is also proposed. It applies a multiple loudspeaker mean based post-processing algorithm to ISDAR-LIB, to refine the results from the fine error point of view, by reducing the variance. Considering $L$ loudspeakers, the related $L$ midpoints $M_l$ and normal vectors $u_l$ are obtained by applying ISDAR-LIB to each of them, as shown
Algorithm 4.2 The Mean-ISDAR-LIB method

**Input** TOAs $\bar{\tau}_l$ and DOAs $\Gamma_l = [\Theta_l, \Phi_l]$ w.r.t. $L$ sources and the microphone array centre; the source positions

**Output** Plane $\bar{p}_M$ (reflector)

1: for $l \leftarrow 1, L$ do
   /* ISDAR-LIB */
2:   Points between 1 and 4 in Algorithm 4.1
   /* Multiple loudspeaker based post-processing */
3:   Calculate unit vector mean $\bar{u}$ from Equation (4.9)
4:   Calculate midpoint mean $\bar{M}$ from Equation (4.9)
5:   Calculate the plane $\bar{p}_M$ utilising $\bar{u}$ and $\bar{M}$

---

**Figure 4.6:** Two examples of different loudspeaker midpoints (blue circles). The green stars are the loudspeakers, the red circle is the microphone array, whereas the lines represent the groundtruth for the shoebox room dimensions.

in Figure 4.6. The mean midpoint and the mean normal vector are then calculated as:

$$\bar{M} = \frac{1}{L} \sum_{l=1}^{L} M_l; \quad \bar{u} = \frac{1}{L} \sum_{l=1}^{L} u_l.$$  (4.9)

By substituting $\bar{M}$ and $\bar{u}$ into Equation (4.6), the plane estimated through mean-ISDAR-LIB $\bar{p}_M$ is obtained. The pseudocode of mean-ISDAR-LIB is reported in Algorithm 4.2.

**Median-ISDAR-LIB.** Another way to exploit multiple loudspeaker information is to calculate the median midpoint and median normal vector, from the ones obtained by applying ISDAR-LIB to $L$ loudspeakers (see Figure 4.6). This method is named as median-ISDAR-LIB, and it reduces the estimation error by avoiding gross errors that would otherwise affect the overall result. These median midpoint $\tilde{M}$ and normal vector $\tilde{u}$ are found as the closest with respect to $\bar{M}$ and $\bar{u}$ (from Equation (4.9)):

$$\tilde{M} = \arg\min_{M_l} ||\bar{M} - M_l||; \quad \tilde{u} = \arg\min_{u_l} ||\bar{u} - u_l||.$$  (4.10)

The plane estimated through this method $\tilde{p}_{MED}$ is obtained by applying $\tilde{M}$ and $\tilde{u}$ in Equation (4.6). The pseudocode of median-ISDAR-LIB is given in Algorithm 4.3.
Algorithm 4.3 The Median-ISDAR-LIB method

Input TOAs \( \bar{\tau}_l \) and DOAs \( \Gamma_l = [\Theta_l, \Phi_l] \) w.r.t. \( L \) sources and the microphone array centre; the source positions

Output Plane \( \hat{p}_{MED} \) (reflector)

1: for \( l \leftarrow 1, L \) do
2: /* ISDAR-LIB */
3: Points between 1 and 4 Algorithm 4.1
4: /* Multiple loudspeaker based post-processing */
5: Points 3 and 4 in Algorithm 4.2
6: Calculate median midpoint as in Equation (4.10)
7: Calculate median normal vector as in Equation (4.10)
8: Calculate the plane \( \hat{p}_{MED} \) utilising \( \hat{u} \) and \( \hat{M} \)

**Figure 4.7:** Illustration of a floor (brown plane) estimated through ETSAC, the floor groundtruth (blue plane), and ellipsoids constructed for three loudspeakers (red, green, blue). Every loudspeaker-microphone combination produces one ellipsoid, thus, having employed 48 microphones, in this example, there are 48 ellipsoids for each colour.

### 4.4.2 Direct Localisation Method

**Ellipsoid Tangent Sample Consensus (ETSAC).** The proposed ETSAC (Figure 4.1 (C)) is a direct localisation method: it only has a reflector localisation step. It uses the information extracted from multiple loudspeakers. RIRs that were recognised by C-DYPSA as outliers are not included by ETSAC in the analysis for the reflector localisation. ETSAC first generates ellipsoids in the homogeneous coordinates (see Chapter 3 for further details regarding the homogeneous coordinates), having major axis equal to the respective reflection path, and foci on the loudspeaker-microphone combination. Then, RANSAC searches for the reflector [Fischler and Bolles, 1981]. In other state-of-the-art methods based on ellipsoids (e.g. [Nastasia et al., 2011]), no specific TOA estimators were identified, and the reflector was informed through a computationally expensive cost function minimisation. An example of the ETSAC output is shown in Figure 4.7. Although, in general, ETSAC provides a unique solution, the particular case shown, with every microphone and loudspeaker at the same height, produces an
4.4.2. Direct Localisation Method

up-down ambiguity. Prior knowledge may be used to constrain the solution (e.g., the floor is closer than the ceiling).

The ellipsoid has the property that the sum of the distances between a random point on its surface and its foci is constant. The TOA of the reflection yields the length of the reflection path. For this reason, ellipsoids are constructed having major axis equal to the reflection paths and foci on the respective microphone and source positions. By finding their common tangent plane, the reflector position is estimated. The parameters characterising a general quadratic surface (i.e. an ellipsoid is a quadratic surface) can be placed in a $4 \times 4$ matrix $E$, to define it in homogeneous coordinates:

$$E = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{12} & h_{22} & h_{23} & h_{24} \\ h_{13} & h_{23} & h_{33} & h_{34} \\ h_{14} & h_{24} & h_{34} & h_{44} \end{bmatrix}, \quad \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{12} & h_{22} & h_{23} \\ h_{13} & h_{23} & h_{33} \end{bmatrix} > 0.$$ (4.11)

To represent a valid quadratic surface it has to satisfy:

$$\det(E) \neq 0, \quad \det(E)/(h_{11} + h_{22} + h_{33}) < 0.$$ (4.12)

A unit sphere centred on the origin of the system is defined as the matrix $E_I$, obtained from Equation (4.11) by choosing $h_{11} = h_{22} = h_{33} = 1$, $h_{44} = -1$, and setting all the other coefficients equal to 0 [Hartley and Zisserman, 2004]. The linear transformations that were defined in Chapter 3, i.e. translation, rotation and scaling, are applied to model the ellipsoid with the required focus positions, axis directions and lengths [Akenine-Moller et al., 2008]. The sphere centre is translated to the point $\Delta X_{i,l} = (B_{0,l} + A_i)/2$. The major axis is defined as the one having the same orientation of the $x$-axis, thus $Q_{x,i,l}^{scal} = \hat{\rho}_{i,l}$, whereas the two minor axes are identical and coincide with $Q_{y,z,i,l}^{scal} = \sqrt{\hat{\rho}_{i,l}^2 - \hat{\rho}_{0,0,i,l}^2}$, where $\hat{\rho}_{i,l}$ is the estimated path length related to the reflection, and $\hat{\rho}_{0,0,i,l}$ is the distance between microphone and main source, that is known a-priori. The scaling matrix $S_{i,l}$ enlarges or shrinks the sphere utilising $Q_{x,x,i,l}^{scal}$, $Q_{y,y,i,l}^{scal}$ and $Q_{z,z,i,l}^{scal}$. Finally, a 3D rotation matrix $R_{i,l}$ is generated utilising the angles of rotation $\alpha_{i,l}^{rot} = \arctan \left( \frac{x_{0,l} - A_{x,i}}{y_{0,l} - A_{y,i}} \right)$, $\beta_{i,l}^{rot} = \arctan \left( \frac{z_{0,l} - A_{z,i}}{x_{0,l} - A_{x,i}} \right)$, and $\gamma_{i,l}^{rot} = \arctan \left( \frac{y_{0,l} - A_{y,i}}{z_{0,l} - A_{z,i}} \right)$. Therefore, the matrix defining the $i$-th microphone and $l$-th loudspeaker ellipsoid is:

$$E_{i,l} = T_{i,l}^T R_{i,l}^T S_{i,l}^T E_I S_{i,l} R_{i,l} T_{i,l}.$$ (4.13)
Algorithm 4.4 The ETSAC method

Input TOAs $\hat{n}_{0,i,l}$ and $\hat{n}_{i,l}$ (direct sound and reflection paths, respectively) w.r.t. the $L$ sources and the $M$ microphones; every source $B_{0,l}$ and microphone $A_i$ positions

Output Plane $p_{\text{ETSAC}}$ (reflector)

1: for $i \leftarrow 1, M$ do
2: \hspace{1em} for $l \leftarrow 1, L$ do
3: \hspace{2em} The unit sphere $E_l$
4: \hspace{2em} The distances $\hat{p}_{0,i,l}$ and $\hat{p}_{i,l}$ from $\hat{n}_{0,i,l}$ and $\hat{n}_{i,l}$
5: \hspace{2em} The parameters $\Delta X_{i,l}, Q_{l,l}^{\text{maj}}, Q_{l,l}^{\text{min}}, \alpha_{i,l}^{\text{rot}}, \beta_{i,l}^{\text{rot}}, \gamma_{i,l}^{\text{rot}}$
6: \hspace{2em} The matrices $T_{i,l}, R_{i,l}$ and $S_{i,l}$
7: \hspace{2em} The ellipsoid through Equation (4.13)
8: for $p \leftarrow 1, C$ do
9: \hspace{1em} The random unit vector $u_p$
10: \hspace{1em} The plane $p_p$ through $u_p$ and Equation (4.14)
11: for $i \leftarrow 1, M$ do
12: \hspace{2em} for $l \leftarrow 1, L$ do
13: \hspace{3em} The tangency coefficient $\zeta_{i,l,p}$
14: \hspace{3em} if $\zeta_{i,l,p} < \tau_l$ then
15: \hspace{4em} The ellipsoid is considered tangent
16: \hspace{3em} End if
17: End for
18: End for
19: End for
20: End for

Once all the $N = LM$ ellipsoids are defined, where $L$ is the number of loudspeakers and $M$ the number of microphones in the array, the next step is to find their common tangent plane. The approach chosen is based on RANSAC [Fischler and Bolles, 1981]. It randomly selects a certain number of points $C$ on the ellipsoid with coefficients $i = 1$ and $l = 1$, and it verifies, by setting a threshold, which one generates the plane that is the closest to the required one. To do so, for each point, it randomly generates a normal vector $u_p = [u_{1,p}, u_{2,p}, u_{3,p}]^T$, that is related to the $p$-th plane tried during the algorithm. This plane can be defined in homogeneous coordinates as $p_p = [u_{1,p}, u_{2,p}, u_{3,p}, p_{4,p}]^T$. Then, the coefficient $p_{4,p}$ can be calculated by considering the general property of tangency between a plane and an ellipsoid, $p_p^T E_{1,1}^* p_p = 0$ [Hartley and Zisserman, 2004], where $E_{1,1}^*$ is the adjoint matrix of $E_{1,1}$. A system of four random ellipsoid equations is, hence, constructed, to obtain $p_{4,p} = (-w_2 + \sqrt{w_2^2 - 4w_1w_3})/2w_1$, with:

\[
\begin{align*}
  w_1 &= h_{44}^*, \\
  w_2 &= 2(h_{14}^* u_{1,p} + h_{24}^* u_{2,p} + h_{34}^* u_{3,p}), \\
  w_3 &= h_{11}^* u_{1,p}^2 + h_{22}^* u_{2,p}^2 + h_{33}^* u_{3,p}^2 + \\
  &\quad + 2(h_{12}^* u_{1,p} u_{2,p} + h_{23}^* u_{2,p} u_{3,p} + h_{13}^* u_{1,p} u_{3,p},)
\end{align*}
\]

(4.14)

where $h_{11}^*, h_{22}^*, h_{33}^*, h_{12}^*, h_{23}^*, h_{13}^*, h_{14}^*, h_{24}^*$ and $h_{34}^*$ are the elements of the matrix $E_{1,1}^*$, organised in the same order as for the general matrix in Equation (4.11). To verify if the plane is tangent to the $N$ ellipsoids, the tangency coefficient is calculated for each
Table 4.2: Parameter values used for the experiments.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope function threshold (C-DYPSA)</td>
<td>$\tau_S$</td>
<td>0.2</td>
</tr>
<tr>
<td>Amplitude threshold (C-DYPSA)</td>
<td>$\tau_A$</td>
<td>25 dB</td>
</tr>
<tr>
<td>Group-delay window length (C-DYPSA)</td>
<td>$T_{gd}$</td>
<td>$3.5 \cdot 10^{-3}$ s</td>
</tr>
<tr>
<td>Segmentation window length (DSB)</td>
<td>$T_{seg}$</td>
<td>$2.7 \cdot 10^{-4}$ s</td>
</tr>
<tr>
<td>Space samples (ML)</td>
<td>$x_{ML}$</td>
<td>$10^4$</td>
</tr>
<tr>
<td>RANSAC samples (ETSAC)</td>
<td>$C$</td>
<td>$10^4$</td>
</tr>
<tr>
<td>RANSAC threshold (ETSAC)</td>
<td>$\tau_t$</td>
<td>$1.4 \cdot 10^{-3}$</td>
</tr>
</tbody>
</table>

randomly generated plane as $\zeta_{i,l,p} = |p_p^T E_{i,j}^* p_p|$, where $| \cdot |$ indicates the absolute value. Since the $p$-th plane is perfectly tangent to the ellipsoid if $\zeta_{i,l,p} = 0$, a threshold $\tau_t$ is empirically set, depending on the dataset used, and, when $\zeta_{i,l,p} > \tau_t$, the ellipsoid related to the $i$-th microphone and $l$-th source is considered as non-tangent. The plane with the fewest non-tangent ellipsoids is selected as the estimated $p_{ETSAC}$. In the scenario where more than one plane have the fewest non-tangent ellipsoids, the plane with the lowest sum of tangency coefficients $\zeta_p = \sum_{i=1}^{M} \sum_{l=1}^{L} \zeta_{i,l,p}$ is selected, among them, as $p_{ETSAC}$. The ETSAC pseudocode is shown in Algorithm 4.4.

4.5 Experimental Evaluation and Discussion

In this section, the proposed methods are evaluated and compared with the baseline methods [Dokmanić et al., 2013, Tervo and Tossavainen, 2012], fulfilling the last contribution of this chapter, defined in Section 4.1. First, it is described how the data used in the experiments were either generated or recorded, then the performance metrics are described, before presenting the comparative studies in terms of reflector localisation accuracy and computational cost. The C-DYPSA performance is also compared to DYPSA [Naylor et al., 2007], and the additional state-of-the-art epoch detector that was proposed in [Kuster, 2008].

As no other RIR dataset was publicly available for microphone arrays that can be defined as compact (see Section 4.2.4), the data for the experiments was both simulated and recorded. A 48-channel bi-circular array with a typical microphone spacing of 21 mm (spatial aliasing half wavelength at 8 kHz) and an aperture of 212 mm (wavelength at 400 Hz) was deployed in four rooms with different sizes and reverberation times (RT60s). This data, recorded with a sampling rate of $f_s = 48$ kHz, is available online at [Coleman et al., 2015b][2]. The experiments were run on MATLAB R2014b on Intel(R) Core(TM)i7-2600 CPU @ 3.40GHz, 16GB RAM PC. To aid the reproducibility, the values of the

Table 4.3: Room dimensions (m), and volumes (m$^3$) in brackets, for the 10 simulated rooms. When the absorption coefficient $\alpha = 0.5$, 2 medium-sized rooms were simulated.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.2$</th>
<th>$\alpha = 0.5$</th>
<th>$\alpha = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>6.0, 4.3, 2.3 (59)</td>
<td>2.4, 4.0, 2.4 (23)</td>
<td>4.1, 5.0, 2.1 (43)</td>
</tr>
<tr>
<td>Medium</td>
<td>7.4, 5.7, 2.5 (105)</td>
<td>7.4, 5.7, 2.5 (105)</td>
<td>7.4, 5.7, 2.5 (105)</td>
</tr>
<tr>
<td>Large</td>
<td>19.7, 24.3, 6.0 (2872)</td>
<td>14.6, 17.1, 6.5 (1623)</td>
<td>6.6, 8.8, 4.0 (232)</td>
</tr>
</tbody>
</table>

parameters used, empirically obtained for the employed datasets, are reported in Table 4.2.

4.5.1 Datasets

Simulated Datasets. Ten rooms were simulated, with varying dimensions and absorption coefficients covering a typical range. They are classified by size and average absorption coefficient $\bar{\alpha}$ in Table 4.3. Inside each room, ten different loudspeaker and microphone array configurations were randomly chosen, leading to a total of 100 different setups. The image source model was employed to generate RIRs, through a Matlab toolbox [Habets, 2006]. The maximum order of the reflections was set to 5, and the high-pass filter, employed to eliminate the artificial energy at the low frequencies, was enabled. The loudspeakers were randomly positioned on a circle around the centre of the microphone array, following a uniform distribution over azimuth angles, with the only condition of not allowing interspaces between the loudspeakers of less than 5 degrees. Two radii of the circle were chosen: 1.00 m for the small sized rooms, and 1.68 m for the medium and large rooms. Their height was the same as for the microphone array, i.e. 0.90 m. The simulated microphone array was composed by 48 evenly spaced microphones placed in two concentric circles, with the inner circle of radius 0.083 m, and outer circle of 0.104 m radius, similar to the prototype designed for the experimental apparatus. Its circular configuration was chosen since it has been proved to be effective for analysing acoustic 3D information [de Vries et al., 2007]. The centre of the microphone array on the horizontal plane was randomly chosen. However, a limit was set to maintain the loudspeakers at a minimum distance from the reflectors. For the small rooms this distance was set to be 0.22 m, whereas for the medium and large rooms 0.36 m. Two noise regimes were imposed on the simulated RIRs, to examine the effects of microphone misplacement and additive measurement noise, respectively. For the first regime, spatial vectors were generated and applied to modify the original microphone positions. They had random directions and amplitudes: the maximum amplitude was
4.5.1. Datasets

Figure 4.8: Floor plan view of the four measured rooms (different scales). The loudspeaker (loudspeaker symbol) and microphone array (red circle) positions are also illustrated. The ceilings are at 2.10 m in AudioBooth, 3.98 m in Vislab, 2.42 m in VML, and 6.50 m in Studio1.

7 mm, i.e., 1 sample at sampling frequency $f_s = 48$ kHz. This regime also models systematic bias in propagation uncertainty. Independent white Gaussian noise was added to each RIR, providing a direct-to-noise ratio (DNR) of 70 dB.

For the second regime, extra datasets were generated with the same ten rooms, randomly choosing ten loudspeaker and microphone array configurations for each, as described above. In this case, the goal is to observe the performance of the methods in the possible scenario where the acoustic channel is estimated [Naylor and Gaubitch, 2010]. The maximum amplitude for the microphone displacement was 1 mm, and the DNR was set to be either 30 dB, 40 dB, or 50 dB. Therefore, having 100 room setups for every DNR, there were a total of 300 additional simulated datasets.

Recorded Datasets. The recorded rooms are named as “Vislab”, “Studio1”, “AudioBooth” (ABooth) and “VML”. Plan view of each of them is shown in Figure 4.8, whereas pictures of them are in Figure 4.9. Table 4.4 reports their general acoustic characteristics, including the average absorption coefficient $\alpha$ in third octave bands. Their $\alpha$s are also shown over the range between 100 Hz and 10 kHz, in Figure 4.10. This is calculated
4. Acoustic Reflector Localisation given Room Impulse Responses

![Figure 4.9: Pictures of the four recorded rooms.](image)

**Table 4.4**: RIR dataset room properties: reverberation time $RT60$, Dimensions, volume $V_{TOT}$, average absorption coefficient $\bar{\alpha}$, and number of loudspeakers used $L$.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$RT60$ (ms) $0.5–1–2$ kHz</th>
<th>Dim. (m), $(V_{TOT})$ (m$^3$)</th>
<th>$\bar{\alpha}$ $0.5–1–2$ kHz</th>
<th>$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABooth</td>
<td>158–110–109</td>
<td>4.1, 5.0, 2.1 (43)</td>
<td>0.55–0.79–0.80</td>
<td>9</td>
</tr>
<tr>
<td>Vislab</td>
<td>385–286–306</td>
<td>7.8, 6.1, 4.0 (189)</td>
<td>0.38–0.51–0.50</td>
<td>12</td>
</tr>
<tr>
<td>VML</td>
<td>505–499–330</td>
<td>2.4, 4.0, 2.4 (23)</td>
<td>0.15–0.15–0.22</td>
<td>22</td>
</tr>
<tr>
<td>Studio1</td>
<td>894–901–945</td>
<td>14.6, 17.1, 6.5 (1623)</td>
<td>0.32–0.32–0.30</td>
<td>4</td>
</tr>
</tbody>
</table>

![Figure 4.10: The average absorption coefficient $\bar{\alpha}$ of the recorded rooms, evaluated in $\frac{1}{3}$-octave bands. 500 Hz, 1 kHz and 2 kHz are highlighted with dotted vertical lines.](image)

by the inverse of the Sabine’s equation:

$$\bar{\alpha} \approx \frac{0.161 \cdot V_{TOT}}{S_{TOT}RT60},$$

(4.15)
where $V_{TOT}$ is the room volume, and $S_{TOT}$ is the total reflective surface area [Kuttruff, 2009]. These rooms were chosen since they cover ranges between small and large $V_{TOT}$, and between small and large $\alpha$ [Lindau et al., 2012]. “Vislab” can be considered as characterised by a medium $V_{TOT}$ (suitable for 20 people) and large $\alpha$, “Studio1” by a large $V_{TOT}$ (for 200 people) and medium $\alpha$, whereas “VML” has both small $V_{TOT}$ (for 2 people) and $\alpha$. In addition, “AudioBooth” is characterised by a small $V_{TOT}$ (for 2 people), and a peculiar $\alpha$, which is very large for high frequencies and medium for low frequencies. Every dataset was recorded using the swept sine RIR method [Farina, 2000], and the sound speed was assumed to be $c_0 = 343.1 \text{ m} \cdot \text{s}^{-1}$.

To analyse the methods varying only parameters like size and RT60, similar $M$ and $L$ must be chosen, therefore, subsets of these datasets were selected. In addition, to be uniform across the datasets, for every room except “VML”, loudspeakers were selected in the horizontal plane only, with the same height as the microphones. In every room the same 48 channel bi-circular compact uniform array of Countryman B3 omni microphones was used, similar to the design of the simulated array. However, there is a small discrepancy in the size of the array used for generating the simulated data and real recordings. The array was simulated by considering the original design, although, due to manufacturer tolerances, the real one has the two radii 2 mm wider. A picture of the employed microphone array is reported in Figure 4.11. Genelec 8020B loudspeakers were used.

The “AudioBooth” is an acoustically treated room at the University of Surrey. $L = 9$ loudspeakers were selected for this chapter, lying around the equator of a truncated geodesic sphere, at 1.68 m radius, at $0^\circ$, $\pm 30^\circ$, $\pm 70^\circ$, $\pm 110^\circ$ and $\pm 155^\circ$ in azimuth relative to the centre channel. The microphone array was positioned at the centre of the sphere at a height of 1.02 m. “Vislab” is another acoustically treated room at the University of Surrey, where the “Surrey Sound Sphere”, having radius of 1.68 m, was assembled. $L = 12$ loudspeakers clamped on the sphere equator, at a height of 1.62 m, with azimuth $0^\circ$, $\pm 30^\circ$, $\pm 60^\circ$, $\pm 90^\circ$, $\pm 110^\circ$, $\pm 135^\circ$ and $180^\circ$, were selected for this work.
4. Acoustic Reflector Localisation given Room Impulse Responses

The microphone array was placed at the centre of the sphere. “VML” is a mock room built within a lab at the University of Surrey, with one wall and ceiling missing like a film set [Liu et al., 2015b]. The microphone array was hanging, at the height of 2.20 m, at the centre of the room. \( L^{\text{TOT}} = 24 \) loudspeakers were laid equispaced around the array with 1 m radius, and facing the centre. The two loudspeakers equidistant from the two walls were discarded, introducing ambiguities with C-DYPSA, the pre-processor that is in common to every tested method. Thus, in VML, the selected loudspeakers were \( L = 22 \). “Studio1” is a large recording studio at the University of Surrey. \( L = 4 \) loudspeaker positions were used, at a height of 1.50 m (the same used for the microphone array). Three of them were placed at a distance of 2 m from the microphone array and azimuth of 0° and ±45°, whereas the fourth one was at 0° azimuth and 3 m distant.

It is important to note that all the microphone and loudspeaker positions, together with the room dimensions, were physically measured in situ, through a laser distance meter. This means that some of the errors, that will be reported during the acoustic reflector localisation method evaluation, are also dependent on the measurement inaccuracies. The smallest component of this error is given by the laser distance meter uncertainty. The largest components of the measurement error are: the manufacturer tolerances assumed during the microphone array assembling; the inaccuracy given by the manual positioning of the laser distance meter in the recorded environment.

4.5.2 Evaluation Metrics

Large errors, generated during the process of finding either the image source or the reflector position, can highly influence result averages. This may not allow discrimination of smaller error behaviours. Therefore, a distinction was made between gross and fine errors, defining a threshold at 500 mm, as in [Omologo et al., 2006].

**TOA Estimation.** For consistent evaluation in spatial terms, the TOA was evaluated as the corresponding propagation distance \( \hat{\rho}_{i,l} = \hat{n}_{i,l}c_0/f_s \), where \( \hat{n}_{i,l} \) is the TOA in samples of the reflection path between the \( i \)-th microphone and the \( l \)-th loudspeaker, \( c_0 = 343.1 \text{ m} \cdot \text{s}^{-1} \) is the sound speed, and \( f_s \) the sampling frequency. The error \( \epsilon^\text{TOA}_{i,l} \) is thus calculated as the distance (in mm) between \( \hat{\rho}_{i,l} \) and its groundtruth. With \( L \) loudspeakers and \( M \) microphones, the overall error is then obtained as the root mean square error (RMSE):

\[
\text{RMSE}_{\text{TOA}} = \sqrt{\frac{1}{LM} \sum_{i=1}^{M} \sum_{l=1}^{L} (\epsilon^\text{TOA}_{i,l})^2}.
\] (4.16)
4.5.2. Evaluation Metrics

**Image Source Localisation.** The image source localisation errors $\epsilon_l$ were evaluated as the Euclidean distance between each single estimated image $B_l$ and its own groundtruth $B_{Gl}$, averaged over all the $L_I \leq L$ loudspeakers giving fine errors:

$$
\mu_\epsilon = \frac{1}{L_I} \sum_{l=1}^{L_I} \epsilon_l, \quad \text{for } \epsilon_l < 500 \text{ mm.} \tag{4.17}
$$

**Reflector Localisation.** To obtain the error in reflector positioning, $x^{ev} = 5$ equi-spaced points were selected between each of the $C_1^{M(L-1)} = N$ source-sensor combinations. The projections of all these points on the estimated plane $p$ and the related groundtruth $p_G$ were then calculated. The Euclidean distance between each pair of points is obtained to give the errors $\epsilon_{i,l,q}^{ref}$. To provide a reliable measure, the RMSE was calculated over all the points as $\text{RMSE}_c = \sqrt{\frac{1}{x^{ev}N} \sum_{i=1}^{M} \sum_{l=1}^{L} \sum_{q=1}^{x^{ev}} (\epsilon_{i,l,q}^{ref})^2}$, indicating the error of each estimated plane. Then, to be coherent with the image source evaluation and provide a summary value for each dataset, the average was calculated over every plane. To do this, for all the reflector localisation methods exploiting multiple loudspeakers (i.e. mean-ISDAR-LIB, median-ISDAR-LIB and ETSAC), the leave-one-loudspeaker-out (LOLO) method was applied. It consists of selecting $L-1$ loudspeakers, where $L$ are the loudspeakers in the dataset. All the combinations $C_{L-1}^{L} = L$ were tested, and the average over the $L_R \leq L$ ones with fine errors provided:

$$
\mu_{\text{RMSE}} = \frac{1}{L_R} \sum_{c=1}^{L_R} \text{RMSE}_c, \quad \text{for } \text{RMSE}_c < 500 \text{ mm.} \tag{4.18}
$$

**Confidence Interval and Gross Error.** The gross error rates were evaluated as $G_\epsilon = (1 - L_I/L)100$ and $G_{\text{RMSE}} = (1 - L_R/L)100$, together with their average over the different datasets and related confidence interval. In contrast to the outlier thresholds defined in Equations (4.17) and (4.18) for image source and reflector evaluation, the threshold separating gross and fine TOA estimation errors was set to match the maximum microphone distance from the array centre, i.e., the array radius of 106 mm. These gross TOA errors were named $G_{\text{TOA}}$.

To provide a better statistical evaluation of the results, the confidence interval of the average across the datasets was calculated as [Rao, 2009]:

$$
\text{CI}_c = \psi \sqrt{\frac{1}{D} \sum_{o=1}^{D} (\mu_{\epsilon,o} - \frac{1}{D} \sum_{o=1}^{D} \mu_{\epsilon,o})^2}, \tag{4.19}
$$

where $\psi = 1.96$ is the critical value for a confidence interval of 95%, $D$ is the number of datasets available, and $o$ is the dataset index. The confidence interval for the RMSEs
4. Acoustic Reflector Localisation given Room Impulse Responses

Table 4.5: The top part shows the gross error $G_{\text{TOA}}$ for the first reflection. The bottom part shows RMSE$_{\text{TOA}}$ and CI$_{\text{TOA}}$ of the reflection path length calculated using DYPSA, C-DYPSA and [Kuster, 2008], for the four recorded datasets, expressed in mm.

<table>
<thead>
<tr>
<th>$G_{\text{TOA}}$ (%)</th>
<th>ABooth</th>
<th>Vislab</th>
<th>VML</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Kuster, 2008]</td>
<td>16.9</td>
<td>20.8</td>
<td>57.5</td>
<td>26.7</td>
<td>30.5 ± 9.0</td>
</tr>
<tr>
<td>DYPSA [Naylor et al., 2007]</td>
<td>2.3</td>
<td>15.5</td>
<td>27.1</td>
<td>0.5</td>
<td>11.4 ± 10.6</td>
</tr>
<tr>
<td>C-DYPSA</td>
<td>0.7</td>
<td>11.5</td>
<td>21.1</td>
<td>0.0</td>
<td>8.3 ± 8.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE$_{\text{TOA}}$ (mm)</th>
<th>ABooth</th>
<th>Vislab</th>
<th>VML</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Kuster, 2008]</td>
<td>89</td>
<td>94</td>
<td>227</td>
<td>194</td>
<td>151 ± 34</td>
</tr>
<tr>
<td>DYPSA [Naylor et al., 2007]</td>
<td>54</td>
<td>110</td>
<td>194</td>
<td>100</td>
<td>115 ± 50</td>
</tr>
<tr>
<td>C-DYPSA</td>
<td>48</td>
<td>95</td>
<td>192</td>
<td>99</td>
<td>109 ± 51</td>
</tr>
</tbody>
</table>

(CI$_{\text{RMSE}}$) was calculated, by substituting into Equation (4.19), $\mu_e$ with $\mu_{\text{RMSE}}$; similarly for the TOA estimation confidence interval CI$_{\text{TOA}}$, by substituting in Equation (4.19) $\mu_e$ with RMSE$_{\text{TOA}}$.

4.5.3 C-DYPSA Evaluation

The novel TOA estimator C-DYPSA (Section 4.2.5) was evaluated and compared against its previous version, the DYPSA algorithm [Naylor et al., 2007], on the four recorded datasets. In addition, experiments to compare DYPSA and C-DYPSA against the state-of-the-art algorithm in [Kuster, 2008] were performed, applying the same datasets. The peak detector utilised in [Kuster, 2008] employed an adaptive threshold based on the time domain amplitudes averaged over neighbouring samples. This is an interesting approach, thus, it was included within the C-DYPSA evaluation section. However, the main scope of this chapter is presenting acoustic reflector localisation methods.

The fine errors produced were calculated as RMSE$_{\text{TOA}}$ (Equation (4.16)), and CI$_{\text{TOA}}$ (Equation (4.19)). These results are reported in the bottom of Table 4.5. C-DYPSA performed better in every dataset. This was expected, since outliers produced by DYPSA for single RIRs are discarded in C-DYPSA, generating a final estimate that is more robust and accurate. The top part of Table 4.5 shows the gross errors $G_{\text{TOA}}$ decreasing for every dataset, applying clusters to the DYPSA outputs.

Regarding Kuster’s method [Kuster, 2008], preliminary results showed a gross error rate close to 100%. Therefore, a couple of improvements were applied on it: the first peak detected was forced to correspond to the RIR peak having greater energy (in other words all the peaks detected before the direct sound were deleted); since the Kuster’s method observes the RIRs by dividing them in temporal windows, it was improved by not allowing multiple peaks inside the same time interval (the only peak selected for each window is the one corresponding to the local maximum of energy). However, as
Table 4.6: Reflector localisation gross errors $G_{\text{RMSE}}$, averaged RMSE $\mu_{\text{RMSE}}$, and confidence interval $\text{CI}_{\text{RMSE}}$, for the simulated dataset grouped by room size (Small, Medium, Large) and absorption coefficient $\bar{\alpha}$, with overall values.

<table>
<thead>
<tr>
<th>$G_{\text{RMSE}}$ (%)</th>
<th>Size</th>
<th>$\bar{\alpha}$</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>ISDAR-LIB</td>
<td>29.7</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Median-ISDAR-LIB</td>
<td>14.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Mean-ISDAR-LIB</td>
<td>9.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>ETSAC</td>
<td>8.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu_{\text{RMSE}}$ (mm)</th>
<th>Size</th>
<th>$\bar{\alpha}$</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>ISDAR-LIB</td>
<td>61</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Median-ISDAR-LIB</td>
<td>27</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>Mean-ISDAR-LIB</td>
<td>207</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>ETSAC</td>
<td>145</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

can be seen in Table 4.5, C-DYPSA still generated better results than [Kuster, 2008], both in terms of fine and gross errors.

As already mentioned, this chapter concerns reflector localisation methods with the first reflection; yet for other early reflections, the clustering of responses across microphones by C-DYPSA can exploit the array’s compactness to reduce errors in the association of epochs to a reflection. For later higher-order reflections, C-DYPSA fails predictably as the level of the reflection energy falls towards the noise floor. A further study could investigate how the number of detectable early reflections varies with the quality of the recordings, and the room properties. Here, C-DYPSA is used to clean up the input to the reflector localisation methods.

### 4.5.4 Simulated Experiments

Experiments were performed considering the simulated datasets introduced in Section 4.5.1. The aim of these simulations was to evaluate the proposed reflector localisation methods, over a wide variety of controlled scenarios, highlighting potential strengths and weaknesses. The metrics utilised were $\mu_{\text{RMSE}}$ (Equation (4.18)) and $\text{CI}_{\text{RMSE}}$ (Equation (4.19)) to evaluate the fine errors, and $G_{\text{RMSE}}$ (Section 4.5.2) for the gross errors. Two different sets of simulations were performed. First, the 100 datasets produced by varying size and $\bar{\alpha}$, with direct sound 70 dB louder than the additive noise, and microphone perturbation of 7 mm maximum, were evaluated, with results reported in Table 4.6. Then, the 300 datasets obtained by varying the DNR were considered, and the results are shown in Table 4.7.
Table 4.7: Reflector localisation gross error $G_{RMSE}$, averaged RMSE $\mu_{RMSE}$, overall for the simulated dataset, varying the DNR (30 dB, 40 dB and 50 dB).

<table>
<thead>
<tr>
<th>Method</th>
<th>30 dB</th>
<th>40 dB</th>
<th>50 dB</th>
<th>30 dB</th>
<th>40 dB</th>
<th>50 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISDAR-LIB</td>
<td>11.8</td>
<td>9.4</td>
<td>9.3</td>
<td>41 ± 2</td>
<td>28 ± 2</td>
<td>30 ± 2</td>
</tr>
<tr>
<td>Median-ISDAR-LIB</td>
<td>19.3</td>
<td>5.1</td>
<td>5.7</td>
<td>45 ± 2</td>
<td>33 ± 1</td>
<td>33 ± 1</td>
</tr>
<tr>
<td>Mean-ISDAR-LIB</td>
<td>3.3</td>
<td>3.9</td>
<td>3.6</td>
<td>128 ± 6</td>
<td>107 ± 6</td>
<td>109 ± 6</td>
</tr>
<tr>
<td>ETSAC</td>
<td>3.7</td>
<td>2.1</td>
<td>2.2</td>
<td>65 ± 1</td>
<td>80 ± 1</td>
<td>80 ± 1</td>
</tr>
</tbody>
</table>

Starting from the first set of simulations (Table 4.6, top), the direct localisation ETSAC gives the best performance, with the lowest $G_{RMSE}$ over the 100 datasets. The multiple-loudspeaker methods (i.e. the mean-ISDAR-LIB and median-ISDAR-LIB) outperformed the single-loudspeaker method (i.e. ISDAR-LIB). Mean-ISDAR-LIB was the better image-source reversion reflector locator, among those tested. Grouping by room size (see Table 4.3) ETSAC, the direct locator, is still better. However, it is possible to note that, in general, every method suffers when the room dimensions become too small. This is due to the fact that, in really small environments, the loudspeakers, which are perfectly omnidirectional for the simulated datasets, can happen to be closer to different reflectors, raising an ambiguity on which reflector is under investigation. This issue is also highlighted by organising the results considering the three different $\alpha$. In fact, although there is no clear trend while observing the results for $\alpha = 0.2$ and $\alpha = 0.8$, when $\alpha = 0.5$ all the methods seem to deteriorate. This sudden increase in the error rate is due to the fact that, as shown in Table 4.3, the smallest room generated has been coincidentally selected to have $\alpha = 0.5$. This means a high probability of having different loudspeakers close to different walls.

To conclude, all the methods are more affected by the room size rather than $\alpha$. Again, the direct localisation ETSAC is the best method under every condition. The $\mu_{RMSE}$ reported on the bottom of the table, should be read with the related $G_{RMSE}$, as the RMSE of the fine error values depends on the amount of gross errors eliminated from the calculation. First, median-ISDAR-LIB has consistent results over all the conditions: although it produces gross errors with more datasets than mean-ISDAR-LIB, if the setup gives fine errors it is more robust on identifying outliers over the estimated image sources. Compared to the image-source reversion method with lowest $G_{RMSE}$, ETSAC’s fine error is better. There is also a tendency for higher $\alpha$s to produce higher fine errors with every method.

For the second set of simulations, observing the $G_{RMSE}$ reported in the top part of Table 4.7, the only two methods that are not strongly affected by lower DNRs are mean-ISDAR-LIB and ETSAC. ETSAC is, in general the best method here tested, however,
it faces small issues with $\text{DNR} = 30\, \text{dB}$. Here, mean-ISDAR-LIB has comparable performance, showing a high robustness over DNR variations. Nevertheless, looking at the fine errors on the right side of the table, a general trend of improving performance with increasing DNR can be noted. The only one that does not follow that trend is ETSAC. However, it has the lowest $G_{\text{RMSE}}$ for $\text{DNR} = 40\, \text{dB}$ and $\text{DNR} = 50\, \text{dB}$, which includes more samples in its $\mu_{\text{RMSE}}$ calculation. Compared to mean-ISDAR-LIB, ETSAC has lower $\mu_{\text{RMSE}}$, showing ETSAC to be the best method tested in these simulations. Given that the first reflection can be $10$-$20\, \text{dB}$ down from the direct sound [Howard and Angus, 2009], reflector estimation may be expected to degrade at DNRs below $20\, \text{dB}$.

### 4.5.5 Comparative Evaluation with Recorded Data

In this work, only a subset of the 3D methods presented in Table 4.1 are evaluated. The other methods are based on assumptions which tend to be too restrictive: in [Kuster et al., 2004] it was assumed that a uniform linear array of microphones was placed parallel to the reflector; in [Zamaninezhad et al., 2014] they assumed only two surfaces in the room to be reflective; and the method in [Ribeiro et al., 2012] required large datasets. Consequently, the methods compared here are Tervo et al. [Tervo and Tossavainen, 2012], Dokmanić et al. [Dokmanić et al., 2013], and the proposed ones. As described in Section 4.2.5, C-DYPSA and DSB were applied as pre-processors, to provide coherent input to every method.

The comparative evaluation was performed in three main parts. First, the three image-source reversion methods, based on a single loudspeaker were evaluated (i.e. Tervo et al. [Tervo and Tossavainen, 2012], Dokmanić et al. [Dokmanić et al., 2013], and the proposed ISDAR-LIB). Keeping in common the reflector localisation part of these methods (i.e. LIB), their image source locator algorithms were assessed. Second, the ISDAR-LIB variants, together with the single loudspeaker version, and the direct localisation method ETSAC, were compared, to determine which conceptual approach is the better to perform the reflector location. The third experiment observes, given the plane generated by the better method, whether the corresponding image source is closer to the groundtruth, compared to the one localised with the three image locator algorithms.

Although most of the method parameters, that were shown in Table 4.2, were empirically derived for the experiments, the number of RANSAC samples $C$, used in ETSAC, was heuristically found. By varying $C$ between $10$ and $10^5$, ETSAC localisation errors were calculated for all the datasets. The results are reported in Figure 4.12. For this specific heuristic test, no discrimination was done between fine and gross errors, i.e. all the errors were included within the RMSE calculation. From the results it is clear, in
4. Acoustic Reflector Localisation given Room Impulse Responses

Both the axes are in logarithmic scale.

**Figure 4.12:** Mean and confidence interval of the ETSAC reflector localisation error.

**Table 4.8:** Image source localisation gross error $G_\epsilon$, averaged error $\mu_\epsilon$, and confidence interval $\text{CI}_\epsilon$, related to the four recorded datasets, and their weighted average (W-AVG). The Maximum Likelihood and Multilateration methods are reported using the reference of the articles where they were firstly presented, i.e. [Tervo and Tossavainen, 2012] and [Dokmanič et al., 2013], respectively.

<table>
<thead>
<tr>
<th>$G_\epsilon$ (%)</th>
<th>ABooth</th>
<th>Vislab</th>
<th>VML</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Tervo and Tossavainen, 2012]</td>
<td>67.7</td>
<td>70.3</td>
<td>89.0</td>
<td>66.0</td>
<td>73.3 ± 9.0</td>
</tr>
<tr>
<td>[Dokmanič et al., 2013]</td>
<td>18.8</td>
<td>25.8</td>
<td>100.0</td>
<td>5.8</td>
<td>37.6 ± 36.0</td>
</tr>
<tr>
<td>ISDAR</td>
<td>0.0</td>
<td>0.0</td>
<td>68.2</td>
<td>0.0</td>
<td>17.1 ± 28.9</td>
</tr>
<tr>
<td>Mirrored ETSAC</td>
<td>0.0</td>
<td>0.0</td>
<td>50.0</td>
<td>0.0</td>
<td>12.5 ± 21.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu_\epsilon$ (mm)</th>
<th>ABooth</th>
<th>Vislab</th>
<th>VML</th>
<th>Studio1</th>
<th>W-AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Tervo and Tossavainen, 2012]</td>
<td>323</td>
<td>328</td>
<td>342</td>
<td>331</td>
<td>334 ± 6</td>
</tr>
<tr>
<td>[Dokmanič et al., 2013]</td>
<td>265</td>
<td>263</td>
<td>–</td>
<td>296</td>
<td>267 ± 10</td>
</tr>
<tr>
<td>ISDAR</td>
<td>208</td>
<td>239</td>
<td>352</td>
<td>232</td>
<td>245 ± 4</td>
</tr>
<tr>
<td>Mirrored ETSAC</td>
<td>82</td>
<td>163</td>
<td>438</td>
<td>100</td>
<td>220 ± 8</td>
</tr>
</tbody>
</table>

general, that with $C$ greater than $10^3$ ETSAC seems to stabilise its performance. In particular, considering all the four datasets $C = 10^4$ was identified as the better value, in terms of mean and CI, and selected to perform the later experiments.

**Image Source Localisation.** The single loudspeaker image-source reversion methods are compared using their gross $G_\epsilon$ and fine $\mu_\epsilon$ errors. The TOA estimator C-DYPSA is used as a pre-processor by every tested method. Therefore, although the performance of the ML algorithm and the multilateration are lower than the proposed methods, this difference cannot be attributed to a large error variance produced by C-DYPSA. As in Table 4.2, $x^{ML} = 10^4$ sample points were used for the ML algorithm, 10 times more than what was originally done in [Tervo et al., 2012]. The results are reported within the first three rows of Table 4.8, showing that ISDAR performs much better than the two baselines, benefiting from two acoustic parameters (i.e. TOA and DOA), rather than only TOA. “VML” appears as a problematic dataset for every algorithm tested, due to its high reverberance at the middle-high frequencies. Furthermore, although the ML algorithm [Tervo and Tossavainen, 2012] and ISDAR provide some fine errors as
Table 4.9: Reflector localisation averaged RMSE $\mu_{RMSE}$, and confidence interval CI$_{RMSE}$, related to the four datasets, and their average (AVG).

<table>
<thead>
<tr>
<th>$\mu_{RMSE}$ (mm)</th>
<th>ABooth</th>
<th>Vislab</th>
<th>VML</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISDAR-LIB</td>
<td>86</td>
<td>47</td>
<td>148</td>
<td>46</td>
<td>102 ± 20</td>
</tr>
<tr>
<td>Median-ISDAR-LIB</td>
<td>92</td>
<td>70</td>
<td>120</td>
<td>54</td>
<td>96 ± 10</td>
</tr>
<tr>
<td>Mean-ISDAR-LIB</td>
<td>56</td>
<td>59</td>
<td>127</td>
<td>49</td>
<td>90 ± 12</td>
</tr>
<tr>
<td>ETSAC</td>
<td>21</td>
<td>30</td>
<td>82</td>
<td>17</td>
<td>52 ± 2</td>
</tr>
</tbody>
</table>

output, the multilateration fails [Dokmanić et al., 2013] with $G_\epsilon = 100\%$. In Table 4.8, the weighted average (W-AVG) error over all the rooms is also reported. W-AVG was calculated by taking into account the amount of fine errors $\epsilon_j$ provided by each dataset.

Reflector Localisation. Having identified the novel ISDAR-LIB as the best single-loudspeaker image-source reversion method, it is then compared with its two novel variants (i.e. mean-ISDAR-LIB and median-ISDAR-LIB), that utilise the information from multiple loudspeakers, and the novel direct localisation method ETSAC. The results are reported in Table 4.9, where the $\mu_{RMSE}$ values are calculated following Equation (4.18). For every dataset and every method $G_{RMSE} = 0\%$.

Results show that ETSAC performs much better than the other methods. This indicates that the better approach to localise reflectors, for these compact microphone array RIRs, is the direct localisation rather than the image-source reversion. On the other hand, it is not possible to distinguish which method is the best among the image-source reversion methods. Every dataset provides different results. However, observing the $\mu_{RMSE}$ averaged over all the datasets, mean-ISDAR-LIB performs best. For the “Vislab”, the introduction of multiple loudspeakers did not have a noticeable effect on the image-source reversion method results (even though it reduced the CI$_{RMSE}$). This is due to the fact that LIB performs similarly with every loudspeaker in this room. “Studio1” is a dataset including four loudspeakers. Due to this small number, methods that use multiple loudspeaker information do not obtain improvement. With the “AudioBooth” there are problems in LIB with the correct identification of the normal vectors $u_l$, however, the midpoints $M_l$ are finely localised. The median-ISDAR-LIB method, which exploits the median of $M_l$, gave lower performance than the others, since it is not robust to fine errors. Finally, “VML” is again the most problematic dataset. However, even with this dataset ETSAC has better performance. Nevertheless, for one loudspeaker, the current implementation of ISDAR-LIB has a run time of 11 ms, whereas ETSAC requires 2.1 s, making ISDAR-LIB appealing for fast processing purposes.

In addition to mean- and median-ISDAR-LIB, two other ISDAR-LIB variants were tested, fitting a plane to the $L$ midpoints. The first used least square (LS), the second used RANSAC. Although they improved over ISDAR-LIB, their performance was
lower than mean- and median-ISDAR-LIB, thus, we decided to not present them within this chapter.

**Image Source Localisation, a Cross-Check.** To evaluate the ETSAC performance directly together with [Tervo and Tossavainen, 2012] and [Dokmanić et al., 2013], images were calculated from the estimated planes. In particular, having all the $p_c$ from running ETSAC, where $c$ is the index of the loudspeaker combination, and the $L$ loudspeaker positions $B_{0,l}$, the $L$ images $B_l$ were localised through the image source method [Allen and Berkley, 1979]. Then, exploiting Equation (4.7), the midpoints $M_l$ were obtained. This method is named as Mirrored ETSAC. Figure 4.13 shows circles to mark reflection positions in the plane of the reflector estimated through ETSAC, with a shoebox outline of each room.

The image localisation errors for Mirrored ETSAC, calculated as before by Equation (4.17), are reported in the last row of Table 4.8. The fine error results indicate that the images generated via the ETSAC-estimated reflector are consistently more accurate than those from the other methods in AudioBooth, Vislab and Studio1. In VML, it is evident an increment on the level of fine errors since all the methods find this dataset challenging, as seen in the high levels of gross error. The key result here, therefore, is the reduction in gross error rate, from over two thirds down to one half using Mirrored ETSAC. The gross error reduction by Mirrored ETSAC however comes with sacrifice in its fine error score in VML. Nevertheless, the W-AVG shows an overall improvement in the performance. In addition, comparing only those cases that both ISDAR-LIB and Mirrored ETSAC successfully resolved (i.e., discarding any cases of gross error), the fine error average are $248\pm5$ mm and $160\pm7$ mm respectively, which confirms the superior performance of Mirrored ETSAC. In spite of this, with a gross error rate of 12.5% across all four datasets, an average error of 22 cm can be observed in the image source location, whereas it is 5 cm in the reflector location. It is also interesting to note that the scale of
these localisation errors is comparable to the limits of human perception [Makous and Middlebrooks, 1990].

4.5.6 Computational Complexity

To assess the computational complexity of image-source reversion and direct localisation methods, a rough calculation of the number of linear and non-linear operations is reported, considering ISDAR-LIB and ETSAC. ISDAR-LIB needs a total of 21 linear operations and 8 non-linear ones to find the reflector. On the other hand, ETSAC employs 83 linear and 9 non-linear operations to generate each of the $N = ML$ ellipsoids, plus 93 linear and 2 non-linear operations for each $p$-th plane generated. In addition, ETSAC uses, in the reflector search step, a sampling method based on RANSAC, which tests $C = 10^4$ planes before finding the best one. On the other hand, ISDAR-LIB, once it localises the image source, it estimates the position of the related plane once. As a result, ISDAR-LIB had a run time approximately 200 times faster than ETSAC (i.e. the run times are 0.011 s for ISDAR-LIB, and 2.123 s for ETSAC).

4.6 Conclusion

Four novel reflector localisation methods have been presented: three image-source reversion (ISDAR-LIB, mean-ISDAR-LIB, median-ISDAR-LIB), and a direct localisation (ET SAC). To automatically extract TOAs from multichannel RIRs, the novel C-DYPSA was also introduced. The proposed methods were compared with two baselines, to discover the better approach for reflector localisation given compact microphone array RIRs.

Simulations of recording conditions, with background noise and microphone position displacements, were used to test the methods by varying the room size, absorption coefficient and DNR. Results showed that ETSAC performed better than the other methods tested, in every condition. All methods were affected by gross errors for small environments, whereas fine errors increased with the increasing of the absorption coefficient. Furthermore, mean-ISDAR-LIB and ETSAC were robust to low DNR conditions. Experiments with recorded RIRs were divided into three main tasks. Firstly, the image localisation algorithms proposed in [Tervo and Tossavainen, 2012] and [Dokmanić et al., 2013] were compared to ISDAR. Results show that the novel ISDAR provided the best performance. The second part of the experiments compared ISDAR-LIB, with mean-ISDAR-LIB, median-ISDAR-LIB, which are novel image-source reversion methods exploiting multiple loudspeakers, together with the novel direct localisation method
ETSAC. Results show that ETSAC localised the reflector with an average 5 cm RMSE, i.e., 42% lower than the best alternative method, here tested. In the last experiment, the reflectors estimated through ETSAC were converted into their corresponding image sources, and compared with the image locators in [Tervo and Tossavainen, 2012] and [Dokmanić et al., 2013]. This showed the percentage of gross errors dropping drastically from 38% (multilateration [Dokmanić et al., 2013]) to 13% (ETSAC). To sum up, these experiments showed that the direct localisation gave better reflector localisation accuracy than image-source reversion across the evaluated RIR datasets. Results also showed that image sources located by mirroring sources with respect to the estimated reflector benefited from the improved reflector estimation. However as ISDAR-LIB ran 200 times faster than ETSAC, it has an advantage for fast processing applications, as well as single-source measurements, which may be useful in tracking.

Further improvements may in the future be found by exploring alternative microphone array arrangements over a large set of rooms, optimal beamformer designs for DOA estimation, and robust methods for multiple loudspeaker ISDAR-LIB post-processing.
Chapter 5

Parametrisation of Reverberant Spatial Audio Objects

As already discussed in previous chapters, knowledge of the acoustic reflector positions can be exploited to create robust methods, to be applied within different audio signal processing areas. In this chapter, one of the reflector localisation methods that was proposed in Chapter 4, i.e. the image source direction and ranging-loudspeaker image bisection (ISDAR-LIB), is employed for spatial audio purposes. ISDAR is utilised to localise source and image sources, in order to parametrise the environmental acoustics given recorded room impulse responses (RIRs). Therefore, together with other algorithms able to analyse the frequency domain of the acoustic field, ISDAR can be considered as one of the foundations of the here proposed reverberant spatial audio object (RSAO).

This chapter is structured as follows: in Section 5.1, a brief description is reported about the spatial audio problems that are tackled by the method proposed in this chapter; in Section 5.2 the encoding part of the RSAO is presented; Section 5.3 reports the RSAO decoding and rendering part; in Section 5.4 the experiments performed are described as a set of subjective assessments\(^1\); finally Section 5.5 draws the overall conclusion.

5.1 Background

One of the main objectives of the spatial audio research field is to reproduce, in a plausible manner, the acoustical characteristics of indoor environments. The intention is to

\(^1\)Results from the subjective assessments, that will be presented in Section 5.4, were produced as part of the collaborative journal article [Coleman et al., 2017], related to the S3A project.
provide consumers with the sensation of being physically present within the recorded environment. This research area can be defined as virtual acoustic environment modelling [Savioja et al., 1999]. It is subdivided into three main tasks: source modelling [Coleman et al., 2014a] (e.g. natural audio, synthetic audio, source directivity), room modelling [Lee et al., 2012] (e.g. modelling of acoustic spaces, artificial reverberation) and listener modelling [Masterson et al., 2012] (e.g. head-related transfer functions (HRTFs), microphone directivity). This chapter focuses on the room modelling, with the aim of spatial audio objects (SAOs) production [Herre et al., 2014].

5.1.1 Auralisation

The process of characterisation of indoor environments to generate synthetic RIRs, followed by convolution with a “dry” signal, is named as auralisation [Hulsebos, 2004]. Initially, researchers attempted to approximate a recorded RIR using multirate systems and discrete-time wavelet transform (DTWT) [Shoenle et al., 1993]. Later, others tried to recreate the sound field of a given room employing the so called “plenacoustic function” [Ajdler and Vetterli, 2003]. This function was generated by utilising RIRs produced through the image source method. However, it conceptually required an infinite number of RIRs in the continuous time domain. The authors tried to achieve the best approximation by choosing a high sampling frequency, and using a large finite number of source-sensor positions. The image source theory was also used in more recent works to synthesise RIRs. For instance, in [Tervo et al., 2013], their method, that was named as spatial decomposition method (SDM), was based on analysing recorded RIRs to localise the image sources, and extract the related pressure signals, giving good results. However, the performance of the image source localisation technique was directly proportional to the signal to noise ratio (SNR).

5.1.2 Spatial Audio Object Coding

Although it was already discussed in Chapter 3, it is important to remark the importance of spatial audio object coding (SAOC) [Engdegård et al., 2008]. This technique tackles a recent issue that concerns spatial audio, and it is related to the parametric model proposed in this chapter. This issue regards the amount of data that is required to be transmitted through the channel, to the renderer, in order to achieve high quality 3D rendering. At the same time, the intention of providing the listeners with specific acoustic sensations, for instance belonging to the recorded environment, has recently started to be investigated considering domestic environments [Bleidt et al., 2017]. In this sense, one challenge is that the band limit of the transmission channel, usually do
not allow multichannel data to be easily sent. Therefore, parametric coding techniques, able to encapsulate acoustic information within a small amount of data, are studied. The moving picture experts group (MPEG) defined a standard for SAOC [Herre et al., 2012]. Coding techniques were suggested, which exploited analysis, parametrisation, and rendering of multiple SAOs [Herre et al., 2014]. An overview of the SAOC standard conceptual structure is shown in Figure 5.1.

Regarding possible ways to parametrically describe the acoustic of a specific environment, a scene description language called Binary Format for Scenes (BIFS) was defined within the MPEG-4 standard, defining the parameters from a “perceptual” point of view [Väänänen and Huopaniemi, 2004]. Later, in [Reiter et al., 2006] the authors introduced a subdivision of the early reflections into two parts, modifying the MPEG-4 BIFS approach. Contemporary, a 3D SAO generation was presented in [Potard, 2006] following a “physical” approach to define the parameters. The MPEG group is currently working on a new standard, the MPEG-H Audio Coding [Herre et al., 2014, Murtaza et al., 2015]. The current status of the standardisation project was reported in [Bleidt et al., 2017, Herre et al., 2015]. This new standard will allow different loudspeaker configurations at the production side, without having knowledge of the microphone array chosen for the recordings.

5.1.3 Spatial Audio Object Parametrisation

In the literature, methods to analyse and parametrise RIRs are recently started to be intensively investigated. As explored in Chapter 3, they can be categorised depending on the type of parameters defined. In other words, whether they employ either high-level [Jot, 1997, Väänänen and Huopaniemi, 2004] or low-level [Melchior et al., 2010, Merimaa and Pulkki, 2005, Tervo et al., 2013] parameters. The method, that is going to be proposed later in this chapter, belongs to the class of the low-level parameters. This choice was made since, although high-level parameters may be useful for applications such as sonic art, where reverberation is predominantly used as a creative effect, they do not allow the producer to directly capture and edit the reverberation parameters.
Table 5.1: Differences and similarities between the proposed RSAO and the state-of-the-art, considering the different stages (from the capture, through the RIR analysis and editing, to the rendering). The three state-of-the-art methods considered are the ones employing low-level parameters, i.e. SIRR [Merimaa and Pulkki, 2005, Pulkki and Merimaa, 2006], R-WFS [Melchior et al., 2010], and SDM [Tervo et al., 2013]. The acronym UCA stands for uniform circular array, DS stands for direct sound, ER means early reflection, LR late reverberation, and TF time-frequency.

<table>
<thead>
<tr>
<th></th>
<th>RSAO</th>
<th>Differences</th>
<th>Similarities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capture</strong></td>
<td>- UCA</td>
<td>- B-Format (SIRR)</td>
<td>- UCA (R-WFS)</td>
</tr>
<tr>
<td></td>
<td>- RIR divided in DS, ERs and LR</td>
<td>- DOA image sources and diffuseness in TF (SIRR)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- TOA, DOA and colouration for DS and ERs</td>
<td>- DOA LR (SDM)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Decay for LR in octave bands</td>
<td>- RIR division in DS, strong ERs, ERs and LR (R-WFS)</td>
<td></td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td>- Source and image sources through TOA, DOA and colouration</td>
<td>- TF bin parameters (SIRR)</td>
<td>- DOA image sources (SDM)</td>
</tr>
<tr>
<td></td>
<td>- LR envelope in octave bands</td>
<td>- Omnidirectional RIR (SIRR, SDM)</td>
<td></td>
</tr>
<tr>
<td><strong>Editing</strong></td>
<td>- Convolve RIRs with anechoic audio objects</td>
<td>- Convolve target DOA with omni (SIRR, SDM)</td>
<td>- VBAP DS and ERs (SIRR)(SDM)</td>
</tr>
<tr>
<td></td>
<td>- VBAP for DS and ERs</td>
<td>- WFS (R-WFS)</td>
<td>- Decorrelate diffuseness (SIRR)</td>
</tr>
<tr>
<td></td>
<td>- Decorrelate LR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Differences and similarities of the proposed method with respect to the state-of-the-art are emphasised through Table 5.1.

One of the first low-level parameter approaches, i.e. the spatial impulse response rendering (SIRR) [Merimaa and Pulkki, 2005, Pulkki and Merimaa, 2006], analysed the time-frequency (TF) spectrum, calculating, for each TF bin, DOA and diffuseness of B-Format RIRs. In particular, for each TF bin, two components were estimated: an impulsive directional one and a diffuse one. Thereafter, directional audio coding (DirAC) [Pulkki, 2007] was developed, where multiple sources were allowed. In both SIRR and its evolution DirAC, the rendering part of the directional component was implemented by generating virtual sources from the related DOA and amplitude parameters. This was done by steering the omnidirectional microphone signal towards the related direction. A later version of DirAC reduced the synthesis artefacts, by utilising virtual microphones instead of steering omnidirectional ones [Vilkamo et al., 2009]. To pan the signals to the available loudspeaker setup, they employed the so-called vector base amplitude panning.
5. Parametrisation of Reverberant Spatial Audio Objects

(VBAP) [Pulkki, 1997]. On the other hand, the diffuse component of the signal TF bins was sent to every channel. Before doing so, if \( L \) channels were available, \( L \) decorrelated versions of each TF diffuse component were generated.

Another approach [Melchior et al., 2010], here named as reverberant wavefield synthesis (R-WFS), utilised, to parametrise RIRs, a method based on the analytical solution to the sound wave equation. The direct sound and the strongest early reflections were observed within spatio-temporal windows, based on the plane wave domain of the signal. They were then reproduced through wavefield synthesis (WFS) [Berkhout, 1988]. On the other hand, the rest of the early reflections and the reverberation were synthesised as diffuse. They considered any point on the sound wave front as a secondary source, like it was firstly done in [Berkhout et al., 1993]. To implement this theoretical concept, a uniform circular array (UCA) of microphones recorded multichannel RIRs [Hulsebos, 2004]. This technique allows the spatial reproduction of sound images for large areas [Horbach et al., 2002].

SDM [Tervo et al., 2013] exploited the same approach employed by SIRR, for the rendering stage. However, different from SIRR, parameters representing time dependent DOAs were calculated from RIRs recorded through a multichannel microphone array. This calculation was based on the assumption that the whole RIR can be represented as superimposition of signals arriving from image sources. Therefore, for each time segment, the DOA of the most prominent image source was estimated and encapsulated as parameter.

In [Li et al., 2006], a model was presented to render spatial sound using multiple RIRs. They approximated the direct sound and early late reflections through the image-source method and the reverberation using filters derived from the recorded signals. In this chapter, these ideas are applied into the context of SAOs, to create the proposed RSAO.

5.1.4 The Reverberant Spatial Audio Object Novelties

In this chapter, a low-level parametric model is presented, to transmit RIR information as part of a SAO. This is referred to as a RSAO, and it is based on the physical representation of the sound scene. The proposed method synthesises RIRs from measured ones, recorded through a UCA of microphones. The novelties are given by three main points:

- the new combination of methods for the parameter estimation;
- direct sound, early reflections, and late reverberation are singularly analysed and synthesised;
• the early reflections and late reverberation frequency domain is also considered.

Regarding the first point, the novel clustered dynamic programming projected phaselapse algorithm (C-DYPSA) (an evolution of DYPSA [Naylor et al., 2007]), the delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988] and the linear predictive coding (LPC) [Makhoul, 1975] are the three methods chosen for the RIR analysis. The three RIR components described in Chapter 2 are individually analysed. Direct sound and early reflections are parametrised with respect to: their source (or image source) ranges and amplitudes exploiting C-DYPSA; their DOAs using DSB; their colourations\(^2\) through LPC. The combination of C-DYPSA and DSB was defined in Chapter 4 as ISDAR. These parameters are then used, on the rendering side, to pan an anechoic signal (i.e. a “dry” audio object) through VBAP. On the other hand, the late reverberation is parametrised by fitting its exponential decay, by looking at different octave bands. This decay is then used by the renderer to recreate the reverberation effect, by multiplying it with Gaussian noise. The resulting signals are sent to every loudspeaker available, to provide diffuseness. An overview of the described components is reported in Figure 5.2.

5.1.5 RIR Definition

A signal \(x(n)\) sent by a source is received by the \(i\)-th sensor as \(y_i(n) = x(n) * I_i(n) + w(n)\), where \(I_i(n)\) is the \(i\)-th RIR, \(w(n)\) is the assumed Gaussian measurement noise and the symbol “\(*\)” represents the convolution operator. In general, as it was already described through Equation 2.47, a RIR is formed of infinite replicas of the source signal, with some additive noise. Each \(e\)-th replica has a path dependent attenuation \(\alpha_{\text{path}}^{\text{path}}\) and time of arrival (TOA), in samples, \(n_{e,i}\) [Kuttruff, 2009]:

\[
I_i(n) = \sum_e (1 - \alpha_{e,i}^{\text{path}}) \delta(n - n_{e,i}) = \sum_e h_{e,i}(n - n_{e,i}), \tag{5.1}
\]

where \(\delta(n)\) is the discrete-time \(n\) dependent Dirac function, and \(n_{e,i}\) is the TOA related to the \(e\)-th peak and \(i\)-th microphone. The RIR \(I_i(n)\) can be decomposed into direct sound \(h^D(n)\), early reflections \(h^E(n)\) and late reverberation \(h^L(n)\) [Kuttruff, 2009]:

\[
I_i(n) = h_i^D(n) + h_i^E(n) + h_i^L(n) = \\
= h_{0,i}(n - n_{0,i}) + \sum_{e=1}^{T_m} h_{e,i}(n - n_{e,i}) + w_R(n), \tag{5.2}
\]

\(^2\)The colouration is a common word utilised in acoustic to define the frequency spectrum of an audio signal.
Figure 5.2: System diagram overview of the RSAO encoder and the combination of RSAO decoder and renderer. The proposed encoder is composed by segmentation, parametrisation and dry object encapsulation; $W^B$ subbands are used to parametrise the late reverberation. The decoder converts dry objects into wet objects before generating $L$ signals ($L$ is the number of loudspeakers) per each part (direct sound, early reflections and late reverberation). Then, the renderer pans the signals $\hat{y}_l(n)$, where $l$ indicates the $l$-th loudspeaker. “wgn” denotes white Gaussian noise, “D”, “E” and “L” stand for direct, early and late respectively.
where $T_m$ is the last peak before the mixing time [Kuttruff, 2009] and $w_R(n)$ is the late reverberation modelled as exponentially decaying noise. The direct sound is defined for $e = 0$, whereas the early reflections for $1 \leq e \leq T_m$.

In the method proposed here, the RIR analysis is performed extracting TOAs, amplitude parameters, DOAs and frequency content for the direct sound and early reflections; whereas for the late reverberation a global frequency dependent analysis in the time domain is performed. This different approach is due to early reflections appearing as delayed impulses in the RIR, whereas the late reverberation appears as a continuum. Furthermore, it is important to note that the energy of the reverberation decays at an exponential rate [Kuttruff, 2009], related to the reverberation time (RT60).

## 5.2 Reverberant Spatial Audio Object Encoding

Recordings of signals produced by different sources within the same environment can generate a huge amount of data, which can be difficult to transmit. Furthermore, they do not allow the final user to interact with the virtual scene. For this reason, SAOC was defined in MPEG-H, treating different audio signals as different SAOs [Herre et al., 2014]. In the proposed approach, the main contribution is the creation of a metadata package (bitstream) defining parameters representing the room acoustics, hence, locating SAOs in space. This package, sent through the transmission channel, together with the anechoic signals, defines the RSAOs.

As starting point, deconvolution can be applied to the $M$ recorded signals, where $M$ is the number of microphones in the UCA. In this way, RIRs are obtained and separated from the anechoic signal. In the experiments performed for this chapter, RIRs are directly recorded from the field. Specific parameters are then extracted from them, packed and sent as bitstreams.

### 5.2.1 TOA Estimation and RIR Segmentation

To extract TOAs from RIRs, a method for selecting the peaks of the signal was developed based on DYPSA [Naylor et al., 2007]. DYPSA was originally designed to estimate glottal closure instances, from speech signals, and it was modified, here in this thesis, to make it applicable to RIRs.

The phased-slope function $S_{gd}(\omega)$ is the average slope of the unwrapped phase spectrum of the short-time Fourier transform (STFT) of the linear prediction residual [Naylor et al., 2007]. In other words, it is the group delay function $G_{gd}(\omega)$ of the signal, but
with the opposite sign \( S_{gd}(\omega) = -G_{gd}(\omega) \). Variations in the time domain (i.e. peaks) correspond to positive-going zero crossings in \( S_{gd}(\omega) \). To reliably select the instants where \( S_{gd}(\omega) \) has these zero crossings, in DYPSA it was smoothed using a Hann window of length \( T_{gd} \). To adapt the algorithm to the purposes of this chapter, a threshold \( \tau_S \) was defined on \( S_{gd}(\omega) \), in order to take only the most significant peaks of \( I_i(n) \). Another threshold \( \tau_A \) was applied on the time domain amplitude, to eliminate the peaks that are much quieter that the main one. These thresholds were empirically derived. The DYPSA output is a sequence of non-zero values placed on the time samples corresponding to the peaks of \( I_i(n) \). TOAs \( \hat{n}_{e,i} \) were estimated by observing the position of these peaks in the DYPSA output time domain. In this chapter, the TOA correspondent to the \( e \)-th non-zero sample is defined by the \( e \) index.

In measured RIRs, after a certain time depending on the measurement setup, peaks are often not distinguishable from the measurement noise. For this reason, as it was already described in Chapter 4, DYPSA was improved, by exploiting the information provided by a multichannel array of microphones. C-DYPSA is proposed in this thesis with this aim. It contains two post-processing refinements to DYPSA. First, the median \( \tilde{n}_{e,l} \) of the estimated TOAs in samples \( \hat{n}_{e,i,l} \) is calculated. If the \( (e+1) \)-th reflection TOA \( \hat{n}_{e+1,i,l} \) is closer to the median \( \tilde{n}_{e,l} \) than the \( e \)-th reflection TOA \( \hat{n}_{e,i,l} \), then \( \hat{n}_{e,i,l} \) is treated as a false positive. Thus, it is replaced with the \( e+1 \)-th reflection TOA value \( \hat{n}_{e+1,i,l} \). After this step, the Grubbs’ test [Grubbs, 1969] identifies the cluster of TOAs related to the \( e \)-th peak considering every microphone in the compact array. RIRs generating outliers to this cluster are discarded. Figure 5.3 shows an example of the output of the C-DYPSA algorithm for a measured multichannel RIR.

Regarding the RSAO parametrisation, C-DYPSA is used to detect the direct sound and the first early reflections, obtaining their TOAs. Finally, the mean of the TOAs over the \( M \) microphones in the array, \( \bar{n}_e \), is calculated and encapsulated to be transmitted.

To maintain the reflection energies as they were in the recorded RIRs, a segmentation of the RIRs is performed, placing Hamming windows having size \( T_{seg} \). The energy is
5.2.2. Direction of Arrival Estimation

Another set of parameters extracted from RIRs is the DOA of the direct sound and early reflections. Several beamforming algorithms can be used to reach this goal [Van Trees, 2002], however, DSB was chosen [VanVeen and Buckley, 1988], since it gives adequate performance for this chapter purposes, and it is simple. To estimate the DOAs for the direct sound and the early reflections, the segmented RIRs $I_{e,i}^S(n)$ are evaluated.

DSB calculates the DOA of a signal exploiting the TOA between the related source (or image source) and each microphone in the array. A phase shift was applied to the signals received by each microphone, to obtain the shifted signals $I_{e,i}^S(n - n_i^{DSB})$, where $n_i^{DSB}$ is the microphone-dependent time shift. Only signals from a particular direction are aligned in time when they are finally summed. The square of this sum is, thus, calculated to obtain the power related to a specific angle. With these prototype delays related to every angle under investigation, the angle that yields the maximum power in output corresponds to the DOA. In other words, DOAs (both azimuth $\Theta_e$ and elevation $\Phi_e$) are calculated for the $e$-th reflection as the angles that are associated to the delays calculated for these time intervals as:

$$U_{e,i}^{EN} = \frac{1}{T_{seg}} \sum_{n=\hat{n}_{e,i} - \frac{T_{seg}}{2}}^{\hat{n}_{e,i} + \frac{T_{seg}}{2}} ||I_{i}(n)||^2 = \frac{1}{T_{seg}} \sum_{n=1}^{T_{seg}} ||I_{e,i}^S(n)||^2,$$  \hspace{1cm} (5.3)

where $I_{e,i}^S(n)$ represents the RIR segments. These energies are used to obtain the pressure amplitude $P_e$ of the output pulses using the equation:

$$P_e = \frac{1}{M} \sum_{i=1}^{M} \sqrt{U_{e,i}^{EN}}.$$  \hspace{1cm} (5.4)

This is another parameter that is included in the metadata.

**Figure 5.4**: Power in output of the DSB for a source positioned at 83° azimuth and 11° elevation, with respect to the centre of the microphone array.
Figure 5.5: Example of a frequency spectrum magnitude for a RIR direct sound (A) and first reflection (B), compared with the respective approximations made by LPC.

(see Equation 3.1):

\[ n^{\text{DSB}} = \arg\max_n \left[ \sum_{i=1}^{M} \sum_{n=1}^{T_{\text{seg}}} r_{S,i}(n) - n_i^{\text{DSB}} \right]^2, \]  

(5.5)

where \( n^{\text{DSB}} \) is a vector containing the \( M \) delays that maximise the function above. A 3D version of DSB was used, exploiting two UCAs, having same radius, lying on two planes parallel to the floor, and having the centre at points with the same \( x-y \) Cartesian coordinates, but a different \( z \). In this way, azimuth and elevation were both estimated using DSB, without ambiguities. In Figure 5.4, an example of angle-dependent power in output of a DSB is reported.

5.2.3 Colour Estimation

Perception of a room acoustic is not only provided by time delays, as the frequency content plays another important role on it. For this reason, the frequency domain of the direct sound and the early reflections was also analysed. This analysis was based, as for the DSB, on the C-DYPSA-based RIR segmentation, by using Hamming windows, as already explained in Section 5.2.2. In this way, the analysis in the frequency domain was done for direct sound and early reflections singularly.

The well-established LPC [Makhoul, 1975] was the method chosen to estimate the spectral envelope. Applying it to every acoustic event in \( h^D_i(n) \) and \( h^E_i(n) \), \( K \)-th order finite impulse response (FIR) filters \( a_{e,i}(n) \) were generated. The \( K+1 \) filter coefficients were averaged over the \( M \) microphones. Results were those parameters providing information about the frequency content in the bitstream. An example of spectrum estimation through LPC for a RIR direct sound and first reflection is shown in Figure 5.5.
5.2.4 Late Reverberation Parametrisation

In the human auditory system, sound cues are processed on a non-uniform frequency scale [Moore, 2012]. Hence, it is important to transform RIRs into a representation that resembles this non-uniform scale by using an appropriate filter bank [Herre et al., 2008]. $h_i^L(n)$ was divided into $W^B$ subbands, through the implementation of a filter bank composed by octave band FIR filters.

The time domain analysis is then performed. The extracted parameter is the energy decay $e_{i,h}^D(n)$, where $i$ indicated the microphone under investigation and $h$ is the subband index. The Schroeder’s algorithm is used to estimate the RT60 given a RIR [Schroeder, 1965]. The envelopes are then averaged over the $M$ microphones:

$$e_h^D(n) = \frac{1}{M} \sum_{i=1}^{M} e_{i,h}^D(n). \quad (5.6)$$

Each $h$-th envelope is encapsulated to be sent through the bitstream as fitting exponential coefficient.

The late reverberation onset time is named as late reverberation TOA (LR-TOA), and each $i$-th microphone has its own defined as $\hat{n}_{T_m+1,i}$. Even in this case, the LR-TOAs are averaged over the $M$ microphones as:

$$\pi_{T_m+1} = \frac{1}{M} \sum_{i=1}^{M} \hat{n}_{T_m+1,i}. \quad (5.7)$$

$\pi_e$ is the parameter sent to the decoder.

5.3 Reverberant Spatial Audio Object Decoding and Rendering

Once the parameters are estimated from the measured RIRs, they are transmitted as a bitstream, together with the anechoic signal $x(n)$ defined in Section 5.1.5. The decoder, using specific algorithms for rendering point sources and diffuseness, converts them into RSAOs. Every signal is directed to the correct loudspeaker through an integration system, i.e. a renderer, which is responsible of handling a surround reproduction system composed of $L$ loudspeakers.

The parametric components received by the decoder were $T_m + 2$, for each audio object: the first component contained the direct sound part $\hat{h}_0(n)$; the next $T_m$ are composed of parameters that approximate the early reflections $\hat{h}_e(n)$ (with $1 \leq e \leq T_m$); the last
one approximates the late reverberation part \( \hat{h}_L(n) \). Each SAO was coded as the combination of the related anechoic signal and packages of metadata, containing information regarding its TOA, DOA, frequency content in the form of LPC coefficients, and energy decays. Due to the different nature of the parameters, two different approaches are used, depending on whether the decoded part represents the source position (direct sound and early reflections) or the room effect (late reverberation) [Jot, 1999]. Direct sound and early reflections are rendered as independent sources (main and image sources), whereas the late reverberation is reproduced as a diffuse source.

5.3.1 Direct and Early Spatial Audio Objects

Direct sound and early reflections are treated at the same way. Using the amplitude of the peaks \( \mathcal{P}_e \) extracted from C-DYPSA, impulses are generated by the decoder and filtered using the filters \( a_e(n) \) given by the LPC parameters. The resulting signals are the estimated RIR parts related to direct sound and early reflections:

\[
\hat{h}_e(n) = \mathcal{P}_e[\delta(n) * a_e(n)] \quad \text{for} \ 0 \leq e \leq T_m \tag{5.8}
\]

where the symbol “*” stands for convolution.

At this point, the dry audio objects are converted to “wet” direct and early SAOs, by convolving the signal \( x(n) \), that was downmixed before being transmitted, with the synthesised direct and early RIR parts in Equation 5.8:

\[
\hat{y}_e(n) = x(n) * \hat{h}_e(n) \quad \text{for} \ 0 \leq e \leq T_m. \tag{5.9}
\]

Since \( L \) loudspeakers in the 3D space do not always coincide with the directions indicated by the DOA parameters, the VBAP algorithm [Pulkki, 1997] is exploited to create virtual sources for the source and image sources. In VBAP, the idea is based on panning between the three closest channels to the intended DOA of the source, and leaving out the others. Different weights \( O_l \) are applied to the amplitude of the sound produced by these loudspeakers to create the impression of the virtual source [Pulkki, 1997]. The \( e \)-th source output is defined as the panned signal \( \hat{y}^{P}_{e,l}(n) = O_{e,l} \hat{y}_e(n) \), where \( \hat{y}_e(n) \) was defined in Equation 5.9, and \( 0 \leq O_{e,l} \leq 1 \) indicates the weight applied to the \( l \)-th channel for the \( e \)-th source, given \( 0 \leq e \leq T_m \). It is important to note that for each source, just three values of \( l \) give \( \hat{y}^{P}_{e,l}(n) \neq 0 \).
5.3.3. Mixer

Figure 5.6: Spectrogram of three late reverberation octave bands, (A) the first (0-88 Hz), (B) the sixth (1.4-2.8 kHz) and (C) the eighth (5.7-11.3 kHz). The top figures represent the bands related to one of the recorded RIRs, the bottom ones the respective bands for the decoded late reverberation.

5.3.2 Late Diffuse Spatial Audio Object

The \( W^B \) exponentials \( e^{\text{D}h(n)} \) are received and multiplied by Gaussian noise, that was produced by a filtered pseudo-random sequence generator. Defining the \( h \)-th subband of the Gaussian noise as \( w_h(n) \), the resulting subband signal is given by \( \hat{h}_h^L(n) = e^{\text{D}h(n)}w_h(n) \). The octave bands are then summed together as:

\[
\hat{h}^L(n) = \sum_{h=1}^{W^B} \hat{h}_h^L(n).
\] (5.10)

In Figure 5.6, three late reverberation frequency bands, related to one of the measured RIRs (top), are compared to the respective bands for the decoded late reverberation (bottom). At this point, the audio objects were converted to a “wet” late diffuse SAO:

\[
\hat{y}^L(n) = x(n) * \hat{h}^L(n).
\] (5.11)

The signal \( \hat{y}^L(n) \) was then sent to a \( L \)-channel decorrelator, which generates \( L \) different signals by convolving \( \hat{y}^L(n) \) by \( L \) all pass filters, i.e. having poles randomly distributed within the unit circle [Zölzer, 2011]. In this way, the reverberation signals sent to each loudspeaker are not correlated among each other, providing the listener with a higher envelopment sensation. The \( l \)-th output signal is so defined as \( \hat{y}^D_l(n) \). In contrast to source and image sources, in this case for every \( 1 \leq l \leq L, \hat{y}^D_l(n) \neq 0. \)
Table 5.2: Parameter values used for the listening tests.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-DYPSA slope function threshold</td>
<td>$\tau_S$</td>
<td>0.2</td>
</tr>
<tr>
<td>C-DYPSA amplitude threshold</td>
<td>$\tau_A$</td>
<td>25dB</td>
</tr>
<tr>
<td>C-DYPSA group-delay window length</td>
<td>$T_{gd}$</td>
<td>$3.5 \cdot 10^{-3} \text{s}$</td>
</tr>
<tr>
<td>Segmentation window length</td>
<td>$T_{seg}$</td>
<td>$2.7 \cdot 10^{-3} \text{s}$ ($e = 0$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$5.3 \cdot 10^{-3} \text{s}$ ($e &gt; 0$)</td>
</tr>
<tr>
<td>LPC filter order</td>
<td>$K$</td>
<td>16</td>
</tr>
<tr>
<td>Subbands for late reverberation</td>
<td>$W^B$</td>
<td>9</td>
</tr>
<tr>
<td>Cut off frequency for subbands</td>
<td>$f_c$</td>
<td>88 Hz (low pass)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.3 kHz (high pass)</td>
</tr>
</tbody>
</table>

5.3.3 Mixer

The last step in the MPEG standard [Herre et al., 2012] is the mixing block. This block receives the signals decoded and rendered by the decoding algorithms, and combines them to create the right connections with the available loudspeakers. Depending on the implemented surround system, $L$ is the number of channels and loudspeakers available (e.g. for a 22.2 surround system $L = 24$ [Hamasaki et al., 2005]). It also receives the TOAs $\hat{n}_e$ extracted from the metadata by the decoder and it uses them to give the right time shift to each part of the RIR. The virtual sources produced by VBAP and the $L$ decaying noises are combined into the final signals sent to the $L$ loudspeakers as:

$$\hat{y}_l(n) = \sum_{e=1}^{T_m} [\hat{y}^P_{e,l}(n - \hat{n}_e)] + \hat{y}^D_{l}(n - \hat{n}_{T_m+1}), \quad (5.12)$$

where $1 \leq l \leq L$ indicates the $l$-th channel. Potential low frequency channels (i.e. subwoofers) are not defined following the low frequency effect (LFE) properties. Instead, the full spectrum signal is post-processed, in order to match the cut-off frequency related to the specific subwoofer that is employed.

5.4 Experimental Evaluation

The sound quality of the proposed RSAOs was evaluated through a set of experiments regarding formal subjective assessments. Parameters related to four recorded rooms were estimated and manually edited. The different edited versions were reproduced to some listeners, asking to evaluate three different sensations: the apparent source distance, the apparent room size, and the listener envelopment. The system thresholds and parameters that were manually set to perform these tests are reported in Table 5.2.
5.4.1 Subjective Assessments

To test the system performance, subjective assessments were performed\(^3\). Four rooms with different dimensions and RT60s were recorded using a double concentric uniform circular array (UCA) of microphones, composed of 48 capsules. A picture of the UCA is reported in Figure 5.7. The recorded RIRs were parameterised and, by modifying these parameters, different virtual rooms were produced. Listeners were then asked to evaluate these new rooms in terms of different perceptual properties, such as source distance, room size and envelopment. For these tests, a slightly improved version of the RSAO model introduced in Section 5.2 was employed: the mixing time \( \hat{\tau}_{n+1} \), that is the instant of separation between early reflections and late reverberation, was calculated following a perceptual based regression approach. In other words, it was calculated utilising the equation [Lindau et al., 2012]:

\[
\hat{\tau}_{n+1} = \hat{\tau}_{P,n+1} = 0.020 \frac{V_{TOT}}{S_{TOT}} + 0.012,
\]

in seconds, where \( \hat{\tau}_{P,n+1} \) is the perceptually evaluated mixing time, \( V_{TOT} \) is the total volume of the room and \( S_{TOT} \) is the total reflective surface.

**Recorded Datasets.** The acoustics of four rooms were measured, by recording RIRs through a multichannel array of microphones. This microphone array consisted of two concentric UCAs, each with 24 omnidirectional capsules (Countryman B3) evenly spaced around the circles. The two radii were 85 mm and 106 mm. This configuration was adopted to allow for robust beamforming with equal resolution in all azimuths. To perform 3D beamforming avoiding elevation ambiguity, two different heights for the microphone array were used for the recordings, at 1.50 m and 1.54 m. A photograph of the recording setup with the double UCA is shown in Figure 5.7. Level calibration was performed by recording a 1 kHz tone at 94 dB sound pressure level (SPL), and scaling

\(^3\)As already mentioned at the beginning of the chapter, the results of these subjective assessments were produced as part of a collaborative work within the S3A project [Coleman et al., 2017].
the recordings for each channel in software. These RIRs are available online at [Coleman et al., 2015b]4.

The room names are “Vislab”, “Studio1”, “Church I”, and “Church II”. Vislab is an acoustically treated laboratory at the University of Surrey ($V = 240 \text{ m}^3; \text{RT60} = 0.80 \text{ s}; \rho = 1.70 \text{ m}$); Studio1 is a classical recording studio at the University of Surrey ($V = 1615 \text{ m}^3; \text{RT60} = 1.11 \text{ s}; \rho = 3 \text{ m}$); Church I is a Victorian church with thick concrete walls and thinly carpeted floor ($V = 1027 \text{ m}^3; \text{RT60} = 1.31 \text{ s}; \rho = 3 \text{ m}$); whereas Church II is a modern church with brick walls, large wooden roof and carpeted floor ($V = 2857 \text{ m}^3; \text{RT60} = 1.41 \text{ s}; \rho = 5 \text{ m}$). The dimensions, RT60s and distance between the loudspeaker and the listener, for every room, are reported in Table 5.3. Genelec 8020B (Vislab), 8030A (Church I and Church II), and 1030A (Studio1) loudspeakers were used. Every source used for this experiment was placed facing towards the microphone array.

### Parameter Estimation and Synthesis

The parameter estimation was performed following the encoder description in Section 5.2. C-DYPSA selected the $T_m = 6$ strongest peaks (ranked by amplitude) detected across all the 48 RIR channels. No frequency filtering was applied to the direct sound in this test implementation. Furthermore, beyond to the RSAO model presented in Section 5.2, the mixing time $\hat{n}_{T_m+1}$ was calculated utilising the perceptual approach defined in Equation 5.13 [Lindau et al., 2012]. Following this, also the reverberation decay calculation was slightly modified. After the subband filtering described in Section 5.2.4, the exponential decay was estimated using the first 20 dB of decaying energy after $\hat{n}_{T_m+1}$.

Following what was defined in Section 5.3, the following parameters were sent to the renderer as a Javascript Object Notation (JSON)5-formatted text string over the user datagram protocol (UDP): direct sound level and position; six early reflection levels, delays, positions, and sets of LPC coefficients factorised into second-order filter sections; late delay; and, nine late subband onset ramp lengths, levels, and exponential decay constants. The parameters were received by an object-based renderer developed as part

---

4 Accessible via http://cvssp.org/data/s3a/

5 JSON is a data-interchange format standard, based on a subset of the JavaScript Programming Language [ECMA-404, 2013].
5.4.1. Subjective Assessments

Figure 5.8: Recorded RIR (A) and representation of the parametric RIRs (B), prior to rendering, showing the squared magnitude of the direct pulse, the delayed, filtered and attenuated specular reflections (black) and the diffuse late reverberation filter (grey).

of the S3A project. RIRs constructed from the parameters for each tested room are shown in Figure 5.8, before being spatialised by VBAP. From this figure, it is possible to note that, in general, the estimated RIRs are similar to the recorded ones, in both early and late parts. However, regarding Studio1, the reverberation tail results to be underestimated. This is due to an overestimation of the mixing time.

Listening Tests. The potential of the RSAO framework, of being easily editable and format-agnostic, from the reproduction system point of view, was demonstrated carrying out a set of pilot listening tests, in the ITU-R BS.1116 standard listening room at the University of Surrey (named “TB7” and described in [Mason, 2016]). Nine listeners were tested: five experienced listeners and four inexperienced listeners. In each of the three tests, described below, listeners were presented with a MUSHRA-style interface, and used multiple sliders to rank each stimulus against the attribute under test. The scales were unmarked and listeners were asked to rate at least one item on each page at the bottom of the scale (0) and at least one item at the top of the scale (100). Three programme items were used: an anechoic hand clap from the “Freesound project” 6, an anechoic male speech, and guitar recordings from the Bang & Olufsen “Music for Archimedes” CD [Bang and Olufsen, 1992]. Four possible types of rendering were used: “22chan”, object rendering to ITU 22:0 loudspeakers; “stereo”, object rendering to stereo; “mono”, the stereo object render summed to mono (centre channel); and, “meas”, a timbral reference created by convolving one of the original omnidirectional microphone recordings, with the source signal, and replaying it in mono. In every case, the direct sound DOA was at 0° azimuth and elevation with respect to the listening position. In reproduction, all channel layouts used a bass management system, using two frontal subwoofers, for bass content. In addition to the parameters estimated from the real room, edited versions of the parameters were used in some of the tests. The

6http://www.freesound.org/people/Anton/sounds/345/
5. Parametrisation of Reverberant Spatial Audio Objects

Table 5.4: P-values of the paired t-test related to the apparent source distance experiment results.

<table>
<thead>
<tr>
<th></th>
<th>Mono</th>
<th>Stereo</th>
<th>22chan</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 0.4 \text{ m vs } r = 0.6 \text{ m} )</td>
<td>1.43%</td>
<td>0.01%</td>
<td>0.01%</td>
</tr>
<tr>
<td>( r = 0.6 \text{ m vs } r = 1.0 \text{ m} )</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.99%</td>
</tr>
<tr>
<td>( r = 1.0 \text{ m vs } r = 2.0 \text{ m} )</td>
<td>0.23%</td>
<td>0.05%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Programme items were manually loudness matched prior to processing. This means that the volume of the items was modified before the convolution with RIRs, in order to have all the same loudness\(^7\). Any loudness differences for different test stimuli were due to the reverberation rendering under test.

**Apparent Source Distance.** The first test was performed asking to the listeners: "Please, rate the following stimuli according to how far they appear to be from you, rating at least one stimulus as farthest and at least one stimulus as nearest". For this test, the guitar and speech programmes were used, together with the stimuli comprised the original and three modified versions of the Church II parameters. They were rendered over mono, stereo and 22chan, together with the original Meas\(^*\) recording. Parameters were altered based on a set of simple rules and a relative distance coefficient \( \rho_e' \), as: \( \overline{P}_0' = \overline{P}_0 / \rho'_0 \), \( \overline{P}_e' = \overline{P}_e / \sqrt{\rho_e'} \), where \( e = 0 \) refers to the direct sound, and \( 1 \leq e \leq T_m \) is the general index for the early reflections. It was chosen to vary the source relative distance \( \rho'_0 \) between 0.4 m and 2.0 m, to avoid underestimation errors, usually made by humans, for greater distances [Kearney et al., 2012]. Figure 5.8 (B) shows the resultant RIRs. Although not edited here, the scheme equally allows for adjustment of reflection delay and direction to account for source and receiver position changes.

Results are shown in Figure 5.9, top. Overall, the listeners were most uncertain about rating the original distance (between \( \rho'_0 = 0.6 \) and \( \rho'_0 = 2.0 \)), yet overall it is clear that by modifying the direct to reverberant ratio (DRR) in the parameter domain the listeners’ perception of distance was altered. In general, the listeners rated the source to appear to be at the same distance across the three reproduction systems. There is also good agreement between the distance ratings for meas\(^*\) (horizontal lines, Figure 5.9, top) and the original parameters over 22chan, although meas\(^*\) was rated closer than the mono and stereo object renders.

To provide a statistical analysis of the results, the paired t-test was performed. It was run considering the pairs of distances that are the most similar to each other, in order to determine if their results are significantly different. The pairs of results that were tested are: \( \rho'_0 = 0.4 \) versus \( \rho'_0 = 0.6 \), \( \rho'_0 = 0.6 \) versus \( \rho'_0 = 1.0 \), and \( \rho'_0 = 1.0 \) versus \( \rho'_0 = 2.0 \).

\(^7\)“Loudness is a psychological term used to describe the magnitude of an auditory sensation” [Fletcher and Munson, 1933].
5.4.1. Subjective Assessments

The results are reported in Table 5.4. It is possible to note that, for every tested pair and across all the reproduction systems, the p-value percentage is always lower than 5%. This proves that, with a significance level of 5%, the increasing of the listeners’ score is consistent with the increasing of $\rho'_0$.

**Apparent Room Size.** The second test was made requesting the listeners to: <<Please, rate the following stimuli according to how large the room appears to be, rating at least one stimulus as largest and at least one stimulus as smallest.>> Here, only the speech programme was used, and the original sets of parameters for Vislab, Studio1 and Church II were rendered over stereo and 22chan together with meas* for each room. In addition, a modified version of Church II, where the late energy decayed 20% slower and $t_{mix}$ was 50ms later, was used (this is named as Church II').

Results (Figure 5.9, middle) show that the listeners were able to rank the parametric rooms in the same order as the real rooms, with no significant differences between the
ratings. This implies that the parameters can properly convey the sense of the size of the target room. Similarly, the edited parameters were able to increase the perceived size of the target space. However, looking at the results related to Studio1, they seem to be reverted with respect to the other datasets'. Meas\(^*\) presents higher rating than both Stereo and Mono. This is related to the fact that the room size perception depends on the reverberation tail. Therefore, the underestimation of the Studio1 reverberation, that was shown in Figure 5.8, causes this perceptual inaccuracy.

The paired t-test was also performed for this experiment. To determine if the results of the rooms with the most similar size are significantly different, the pairs of datasets that were tested are: Vislab versus Studio 1, Studio1 versus Church II, and Church II versus Church II\(^\prime\). The results are reported in Table 5.5. As it was for the apparent source distance experiment, for every tested pair and across all the reproduction systems, the p-value percentage is always lower than 5%. This proves that, with a significance level of 5%, the increasing of the listeners’ score is consistent with the increasing of the actual room size.

**Table 5.5:** P-values of the paired t-test related to the apparent room size experiment results.

<table>
<thead>
<tr>
<th></th>
<th>Meas(^*)</th>
<th>Stereo</th>
<th>22chan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vislab vs Studio 1</td>
<td>0.00 %</td>
<td>0.05 %</td>
<td>0.01 %</td>
</tr>
<tr>
<td>Studio 1 vs Church II</td>
<td>0.10 %</td>
<td>0.02 %</td>
<td>0.01 %</td>
</tr>
<tr>
<td>Church II vs Church II(^\prime)</td>
<td>-</td>
<td>0.03 %</td>
<td>0.01 %</td>
</tr>
</tbody>
</table>

**Listener Envelopment.** Finally, listeners were asked to: <<<Please, rate the following stimuli according to how surrounded you feel by them, rating at least one stimulus as most enveloping and at least one stimulus as least enveloping>>>. For this test, the clap and guitar programmes were used, and the original sets of parameters for all rooms were rendered over stereo and 22chan and compared with the measured reference meas\(^*\).

In general, listener envelopment increased with room size, although the ratings were similar between the two church buildings (Figure 5.9, bottom). This similarity might be explained by Church I having stronger early reflection parameters, yet Church II being slightly more reverberant (see Figure 5.8 (A)). The listener envelopment results were similar across all the reproduction systems for the smaller rooms, but for the larger spaces listeners rated the 22chan reproductions to be more enveloping than the mono or stereo. This result is not entirely surprising; nevertheless, it illustrates that, by using an object-based approach to reverberation, the renderer can improve the listener envelopment where more loudspeakers are available. When combined with the results above, the ratings suggest that this increase in envelopment may not in general come at the cost of altering the apparent size of the reverberant space.
Table 5.6: P-values of the paired t-test related to the listener envelopment experiment results.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Meas*</th>
<th>Stereo</th>
<th>22chan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vislab vs Studio 1</td>
<td>2.54%</td>
<td>0.09%</td>
<td>0.15%</td>
</tr>
<tr>
<td>Studio 1 vs Church I</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Church I vs Church II</td>
<td>0.01%</td>
<td>68.51%</td>
<td>52.41%</td>
</tr>
</tbody>
</table>

The paired t-test was performed also for this experimental results. The pairs of datasets that are supposed to have the most similar level of listener envelopment are compared, in order to determine if their results are significantly different. The pairs of dataset tested are: Vislab versus Studio 1, Studio1 versus Church I, and Church I versus Church II. The results are reported in Table 5.6. Here, almost every tested pair has a p-value percentage lower than 5%, across all the reproduction systems. This proves that, in general, with a significance level of 5%, the increasing of the listeners’ score is coherent with the level of listener envelopment that was a-priori expected for the rooms. However, as it was also noted in Figure 5.9, for both Stereo and 22chan, there is no statistical difference between the results given by Church I and Church II. As already mentioned, this is given by the fact that Church I presents strong early reflections, whereas Church II has a longer reverberation tail. Both are factors that regulate the spatial impression, a perceptual property that is similar to the listener envelopment [Bradley and Soulodre, 1995], thus, it can create a confusion during the listening tests.

5.5 Conclusion

A model to generate RSAOs, composed by an encoding and a decoding part, has been presented. Novelties for the parameter extraction have been introduced: the C-DYPSA algorithm estimated the TOAs and, hence, the energies of the RIR components; the DSB estimated the DOAs; LPC analysed the frequency content. The late reverberation was analysed in octave bands, and parametrised considering the related exponential decays.

The experiment performed was a set of formal listening tests. The listeners were asked to evaluate the parametrisation system regarding three different aspects: the perception of the source distance, the room size perception, and the listener envelopment. Results showed that the parametrisation algorithms maintain the sense of room size across reproduction systems and with respect to a measured reference case. They also showed that a source could be made to appear more distant just by adjusting the signal level. Furthermore, the space could be made appear larger by modifying the subband exponential decays related to the late reverberation. Moreover, although the listener envelopment resulted to be partially related to the room size, these experiments also
demonstrated that, where more loudspeakers were available, the RSAO render could improve the listener envelopment.

Future works may be focused on finding additional parameters to refine the RIR analysis process. This would allow a higher perceptual accuracy for the rendered sounds. In addition, production tools may be developed, also formally relating perceptual properties to specific parameter setups.
Chapter 6

Source Separation through Multipath Propagation Analysis

Blind source separation is one of the most investigated fields in the audio signal processing community. Several application areas can benefit from it. For instance, blind source separation methods are often applied to communication systems: defence signal processing employs audio source separation methods to develop passive sonar systems [Sutin et al., 2010]. Moreover, in a different research area, biomedical engineers usually apply source separation to analyse electrocardiograms, electroencephalograms, or magnetic resonance imaging [Ungureanu et al., 2004]. Noteworthy are also projects on ancient document restoration, which utilised similar techniques [Tonazzini et al., 2007]. However, due to its large range of audio applications that better fit the purpose of this thesis (e.g. speech enhancement [Mohammadiha et al., 2013], crosstalk cancellation [Akeroyd et al., 2007], hearing aids [Healy et al., 2013], and automatic speech recognition [Li et al., 2014]), speech source separation will be investigated within this chapter.

6.1 Introduction

During the last two decades, the speech separation problem has gained quite of attention, and different approaches have been proposed to solve it. Some of them performed separation focusing on the challenging problem of having availability of one single recorded channel [Jang and Lee, 2003, Radfar and Dansereau, 2007, Schmidt and Olsson, 2006]. However, the small amount of information carried by one signal limited these models. For this reason, more accurate methods usually employ multichannel microphone arrays. As described in Chapter 3 (Section 3.3), these methods are classically categorised into
three main groups [Vincent et al., 2012]: beamforming based [Araki et al., 2003, Coleman et al., 2015a, VanVeen and Buckley, 1988], independent component analysis (ICA) based [Bell and Sejnowski, 1995, Cardoso, 1998, Makino et al., 2007], and time-frequency (TF) masking based [Alinaghi et al., 2014, Deleforge et al., 2015, Mandel et al., 2010, Sawada et al., 2011, Yilmaz and Rickard, 2004].

In general, beamformers can be used in many scenarios, from the one with more sources than microphones (i.e. the underdetermined case) to the one with more microphones than sources (i.e. the overdetermined case). However, they require a large number of channels to achieve high quality separation. In addition, they do not take into account the multipath propagation, making them not robust, when utilised within reverberant environments.

The same issue with reverberant environments is also faced by ICA. Additionally, the ICA based methods have two more limitations: they cannot be applied to underdetermined scenarios; they produce permutation ambiguities over the separated signals given in output. However, when it is applicable, ICA is known to provide high quality performance in source separation, assuming uncorrelated signals.

Due to the beamformer and ICA limitations, recently, the most adopted approach is the one based on TF masking [Wang, 2008]. Following the human hearing functionalities [Brown and Cook, 1994], the TF domain approximates the time variant frequency domain of each source signal. This approach can be used for different scenarios, including the underdetermined one. Furthermore, it provides good performance in reverberant conditions, even using a pair of microphones. In [Yilmaz and Rickard, 2004], a pair of omnidirectional microphones was employed within anechoic rooms. The authors estimated from the two recorded mixtures the air attenuation coefficients and the times of arrival (TOAs), related to the direct sound produced by each source. Assuming the spectral sparsity\(^1\), hard binary masks in the TF domain were created to separate the sources. Despite its high performance, this method was only tested with anechoic mixtures. Later evaluation of it, using reverberant environments, showed lower performance with respect to the current state-of-the-art [Mandel et al., 2010]. Similar ideas were later proposed, to study reverberant mixtures [Mandel et al., 2010, Sawada et al., 2011]. In [Mandel et al., 2010], the authors presented a method based on binaural recordings, utilising dummy heads instead of omnidirectional microphones in the open space. In this way, two interaural cues could have been exploited, relating the azimuthal sound direction of arrival (DOA) to the head orientation: the interaural level difference (ILD) and the

\(^1\)The spectral sparsity assumption is also known as \(W\)-disjoint orthogonality of the source signals in the mixture. In other words, two source signals are \(W\)-disjoint orthogonal, if the supports of their short-time Fourier transforms (STFTs) are disjoint, for a given window function \(W\) [Yilmaz and Rickard, 2004].
interaural phase difference (IPD) [Hofman and Van Opstal, 1998]. TF soft masks were, thus, generated, considering a Gaussian mixture model (GMM), that took into account both the ILD and IPD probabilities. These probabilities defined the correspondence between each TF bin and a specific sound source in the mixture. Due to the employment of the expectation-maximisation (EM) algorithm, this method was named as model-based expectation-maximisation source separation and localisation (MESSL) [Mandel et al., 2010]. The method presented in [Sawada et al., 2011], utilised, instead, the so-called mixing vector (MV). This is a vector containing the time invariant frequency response of the room. Through a Gaussian density function, the clustering approach was utilised to determine the probability of each TF point to belong to a specific source in the mixture. This was done for every sensor available, and, from this probability, TF masks were generated, as for [Mandel et al., 2010]. Although it was the first approach considering the room acoustics as a cue for the source separation, it did not consider it from an interaural point of view, hence, limiting the performance. In [Alinaghi et al., 2014], the two methods firstly proposed in [Mandel et al., 2010] and [Sawada et al., 2011] were combined together. In particular, the three cues ILD, IPD and MV were utilised to create a probability distribution, able to generate the required soft masks. In [Deleforge et al., 2015], the interaural cues IPD and ILD produced a high-dimensional vector, in the TF domain. Utilising a manifold learning technique, the authors employed a non-linear dimensionality reduction to project these vectors into a 2D space. This space had as dimensions the azimuth and elevation of the sound sources. Regression was then used to infer the dimensional transformation, locating source in space. Soft masks were finally generated. This is a new way of approaching the problem, via first training a model, to then test it. However, in [Deleforge et al., 2015], the results were produced by employing as test dataset, the same one that was utilised for the training session.

In reverberant scenarios, a sound produced by a source interacts with the environment during its propagation, before reaching the listening position. Therefore, the acoustic multipath properties have to be taken into account during the source separation [Vincent et al., 2014]. In literature, yet few works can be found that consider both direct sound and reflections. In [Huang et al., 2005], the source separation problem was divided into different procedures, by applying a deconvolution to each individual echo. Due to its high computational cost, in [Rotili et al., 2010], a real-time implementation was presented, where they replaced the inverse filtering with an efficient iterative algorithm. However, this method performance degraded with low signal-to-noise ratio (SNR) conditions. In [Nesta and Omologo, 2012], a variation of ICA estimated the temporal dependent mixing time, considering the multipath propagation of the sounds. However, ICA introduced the permutation problem, and its effect was also enhanced by the incorrect source component alignment. Deconvolution of the received signals was
proposed in [Asaei et al., 2014], where simulated RIRs matching the temporal support of recorded ones were estimated. Nevertheless, low SNRs may prevent to correctly observe the temporal support of the recorded RIRs. Additionally, the overall model did not consider binaural recording effects (e.g. head shadowing), making it not applicable for computational auditory scene analysis (CASA)-based studies [Brown and Cook, 1994].

In this chapter, a novel blind source separation method is presented, which takes into account interaural cues related to both the direct sound and the first early reflection. In fact, the information carried by the first early reflection carries only 10-20 dB of energy less than the direct sound [Howard and Angus, 2009], and it is the most highly contributing to the spatial effect [Bech, 1998]. The proposed method is an extension of the MESSL method, previously presented by Mandel et al. [Mandel et al., 2010]. Therefore, it is named as early reflection MESSL (ER-MESSL), since it extends MESSL by including early reflection information. A second novelty is given by the application of the novel source and image source localisation algorithm, named as image source direction and ranging (ISDAR), that was proposed in Chapter 4, to initialise the EM algorithm. The new method was tested by utilising recorded datasets, from four rooms, with different size and reverberation time (RT60).

The overall structure of this chapter is as follows: Section 6.2 will define the theoretical foundations of the proposed approach; in addition, it will also state the overall assumptions that have been made. In Section 6.3, the proposed interaural cue model including the first reflection information will be presented. Section 6.4 will describe the source separation algorithm, based on GMMs and EM algorithm. In Section 6.5, the experiments, together with the related results and discussion, will be provided. Finally, Section 6.6 will draw the conclusion.

6.2 Background Definitions

6.2.1 Proposed Method Assumptions

During the development of the proposed blind source separation method, some assumptions were made, defining the scientific boundaries of the work. These assumptions can be listed as:

- The number of sources in the mixture $L$ is known a priori;
- The source signals are sparse in the TF domain ($W$-disjoint orthogonal sound sources [Yilmaz and Rickard, 2004]);
6. Source Separation through Multipath Propagation Analysis

- The mixing system is time invariant;
- The early reflections have a dominant specular component;
- The sources are distant enough from the reflectors;
- The sources are in the far-field;
- For initialisation purposes only, multichannel room impulse responses (RIRs) are available.

Although the number $L$ has to be known a priori, there are no restrictions on that number with respect to the number of microphones $M$, thus, the method can be also applied to underdetermined scenarios. Sparsity over the TF domain corresponds to the assumption of having, for each TF bin, only one of the sources in the mixture. Sources and microphones are assumed to be static within a static environment, thus, the mixing system is time invariant. Having the early reflections dominant specular components, they can be detected in the RIRs during the EM initialisation process. Sources have to be distant enough from the reflectors, in order to have the first reflection arriving 6 ms later than the direct sound. 6 ms is the just noticeable difference (JND) for humans discriminating two sounds in time [Friberg and Sundberg, 1995]. The far-field assumption allows to consider the signals as plane waves. Finally, the algorithm chosen to initialise the EM relies on multichannel RIR knowledge.

6.2.2 Binaural Room Impulse Response

As already discussed in previous chapters, RIR is a peculiar signal which characterises the acoustic of an environment with respect to source and sensor positions. The acoustic signals characterising the recording setup, recorded by utilising microphones placed in a dummy head ear canals, are usually named as binaural RIRs (BRIRs). Similar to what was described for a generic RIR in Equation 2.47, they are composed of the superimposition of direct sound and multiple reflections:

$$I_{i,l}(n) = h_{0,i,l}(n - n_{0,i,l}) + \sum_{e=1}^{T_m} h_{e,i,l}(n - n_{e,i,l}) + w_R(n), \quad (6.1)$$

where $i \in \{1, 2\} \in \mathbb{N}$ and $l$ are the microphone and source indexes, respectively; $n$ is the discrete temporal variable, $T_m$ indicates the last early reflection, $w_R(n)$ is the reverberation tail, whereas $e$ is the reflection index ($e = 0$ indicates the direct sound). Following the assumption of having dominant specular components, the early reflections are approximated by Dirac deltas, having different amplitudes.
2.3. Interaural Spectrogram

Figure 6.1: Representation of an ideal BRIR, zoomed into its direct sound and first reflection components (depicted as red Dirac pulses). The top figure shows the RIR related to the sensor labelled as 1, whereas the bottom one shows the RIR recorded at sensor $i = 2$. The amplitudes and delays between Dirac deltas are defined in Equation 6.2.

In this chapter, the direct sound and the first early reflection (i.e. $e = 0$ and $e = 1$) components are considered for blind source separation purposes. The second reflection is not considered since it already conveys 20-40 dB less than direct sound [Howard and Angus, 2009]. Direct sound and first reflection can be defined as:

\begin{align}
  h_{0,1,l}(n) &= P_{0,1,l} \delta(n - n^D_D); \\
  h_{1,1,l}(n) &= P_{1,1,l} \delta(n - (n^D_D + n^{DF}_D)); \\
  h_{0,2,l}(n) &= P_{0,2,l} \delta(n - (n^D_D + n^{DS}_D)); \\
  h_{1,2,l}(n) &= P_{1,2,l} \delta(n - (n^D_D + n^{DS}_D + n^{ST}_D));
\end{align}

(6.2)

where $P_{e,i,l}$ is the pressure amplitude related to the $i$-th sensor, $l$-th source, and $e$-th reflection; $\delta(n)$ represents the Dirac delta in the time domain; $n^{DF}_D$, $n^{DS}_D$, and $n^{ST}_D$ are, respectively: the time delay between direct sound and first reflection, recorded at the sensor labelled as 1; the delay between direct sound recorded at the second sensor and direct sound recorded at the first microphone; the delay between the first reflection recorded at the second sensor and the first reflection recorded at the first microphone. Additionally, $n^D_D$ is defined as the direct sound TOA, related to channel $i = 1$. A visualisation of these signal delays is presented in Figure 6.1.
6.2.3 Interaural Spectrogram

Following the definition of BRIR in Equation 6.1, the mixtures received at the \( i \)-th sensor can be written as:

\[
y_i(n) = \sum_{l=1}^{L} x_l(n) \ast I_{i,l}(n) \ast w_{i,l}(n),
\]

(6.3)

where \( x_l(n) \) is the signal generated by the \( l \)-th source in the time domain, \( w_{i,l}(n) \) is the convolutive white Gaussian noise, \( L \) is the number of sources, and “\( \ast \)” is the convolution operator. Since the human auditory system analyses the received mixtures in both time and frequency domain [Brown and Cook, 1994], the TF domain of \( y_i(n) \) is investigated here, and it is calculated via applying the STFT to \( y_i(n) \), to obtain:

\[
y_i(m,\omega) = \sum_{l=1}^{L} x_l(m,\omega)I_{i,l}(\omega)w_{i,l}(m,\omega),
\]

(6.4)

where \( m \) is the discrete time bin label, whereas \( \omega \) is the angular frequency variable. It is important to note that \( I_{i,l}(\omega) \) is not time dependent, due to the assumption of time-invariant mixing systems. Focusing on the binaural system problem, the interaural spectrogram can be defined as [Mandel et al., 2010]:

\[
y^{IS}(m,\omega) = \frac{y_1(m,\omega)}{y_2(m,\omega)} = 10^{\alpha^{ILD}(m,\omega)/20} \exp[j\phi^{IPD}(m,\omega)],
\]

(6.5)

where \( \alpha^{ILD}(m,\omega) \) and \( \phi^{IPD}(m,\omega) \) are the ILD and IPD of the observation, respectively, and \( j = \sqrt{-1} \).

6.3 Proposed Interaural Cue Model

The IPD and ILD cues can be modelled to generate probability distributions in order to identify the dominant source, given each TF bin. The novel IPD model that considers both the direct sound and first early reflection is proposed in this section. Furthermore, the ILD model (that was presented in [Mandel et al., 2010]) is described. Finally, it will be shown how to combine the two model parameters, to generate the probability distributions.

6.3.1 Interaural Level and Phase Differences

The IPD mathematical model is defined to match the behaviour of the observed IPD \( \phi^{IPD}(m,\omega) \). Different from what was done in [Mandel et al., 2010], where only the
Interaural Level and Phase Differences

Figure 6.2: The IPD as a function of frequency for a single source convoluted to an ideal BRIR formed by only direct sound and first reflection. The IPD model presented in [Mandel et al., 2010] (MESSL) is the straight line, whereas the here proposed IPD model (ER-MESSL) is the fluctuating curve.

direct sound was included in the model, the model here proposed is based on Figure 6.1.

Considering ideal BRIRs formed by direct sound and first reflection, the two channel frequency responses can be written as:

\[
\hat{I}_{1,l}(\omega) = \exp[-j\omega n^D_l] (P_{0,1,l} + P_{1,1,l} \exp[-j\omega n^{DF}_l]); \\
\hat{I}_{2,l}(\omega) = \exp[-j\omega n^D_l] (P_{0,2,l} \exp[-j\omega n^{DS}_l] + P_{1,2,l} \exp[-j\omega (n^{DF}_l + n^{ST}_l)]).
\]  
(6.6)

Their ratio is the interaural frequency response model, and it is given by:

\[
\hat{I}_l(\omega) = \frac{P_{0,1,l} + P_{1,1,l} \exp[-j\omega n^{DF}_l]}{P_{0,2,l} \exp[-j\omega n^{DS}_l] + P_{1,2,l} \exp[-j\omega (n^{DF}_l + n^{ST}_l)]}.
\]  
(6.7)

The phase of this equation, denoted as \(\hat{I}_l^\text{ang}(\omega)\), corresponds to the proposed IPD model.

Referring to the \(l\)-th source, the difference between the observed IPD \(\phi_{\text{IPD}}(m,\omega)\) and the IPD model can be then formulated as the phase residual:

\[
\hat{\phi}_{l}^{\text{IPD}}(m,\omega; C_l) = \phi_{l}^{\text{IPD}}(m,\omega) - \hat{I}_l^\text{ang}(\omega; C_l),
\]  
(6.8)

that is forced to be wrapped into the interval \([-\pi, \pi]\); and:

\[
C_l = [n^{DS}_l, n^{DF}_l, n^{ST}_l, P_{0,1,l}, P_{1,1,l}, P_{0,2,l}, P_{1,2,l}].
\]  
(6.9)

An example of the IPD model fitting the observed data is shown in Figure 6.2. Also a visual comparison with the model firstly introduced in [Mandel et al., 2010] is shown.

On the other hand, the ILD cue \(a_l^{\text{ILD}}(m,\omega)\) is modelled directly considering the frequency-dependent BRIR from the observation, as in [Mandel et al., 2010]:

\[
a_l^{\text{ILD}}(\omega) = 20 \log_{10} \left| \frac{I_{1,l}(\omega)}{I_{2,l}(\omega)} \right|,
\]  
(6.10)
where “|·|” indicates the absolute value.

### 6.3.2 Probability Distributions

The interaural convolutive noise is defined as \( w_I^l(m, \omega) = w_{1,l}(m, \omega)/w_{2,l}(m, \omega) \). Its absolute value \(|w_I^l(m, \omega)|\) is assumed to have normal distribution. Therefore, regarding the ILD cue, the probability of each TF bin being associated to source \( l \) can be written as a Gaussian distribution [Alinaghi et al., 2014]:

\[
p(\alpha_{ILD}(m, \omega)|l) = \mathcal{N}(\alpha_{ILD}(m, \omega)|\mu_{ILD}(\omega), \sigma_{ILD}^2(\omega)),
\]

where \( \mu_{ILD}(\omega) \) is the mean of the distribution, as it will be better described later in Section 6.4.1, and \( \sigma_{ILD}^2(\omega) \) is the variance of the distribution.

Regarding the IPD cue, a top-down approach is firstly needed, to overtake the ambiguity produced by unwrapped phase [Mandel et al., 2010]. In fact, unwrapped phases cannot be uniquely assigned to the related interaural time difference (ITD). Then, the phase residual \( \hat{\phi}_{IPD}(m, \omega; C_l) \) can be modelled through a Gaussian distribution:

\[
p(\hat{\phi}_{IPD}(m, \omega)|l, C_l) = \mathcal{N}(\hat{\phi}_{IPD}(m, \omega; C_l)|\mu_{IPD}(\omega; C_l), \sigma_{IPD}^2(\omega; C_l)),
\]

where \( \mu_{IPD}(\omega; C_l) \) and \( \sigma_{IPD}^2(\omega; C_l) \) are the IPD model distribution mean and variance, respectively. They will be better described in Section 6.4.1.

To sum up, by assuming the IPD and ILD observations to be conditionally independent given their related parameters, their probability distributions can be combined as:

\[
p(\alpha_{ILD}(m, \omega), \hat{\phi}_{IPD}(m, \omega)|l, C_l) = \mathcal{N}(\alpha_{ILD}(m, \omega), \hat{\phi}_{IPD}(m, \omega; C_l)|\Xi),
\]

where \( \Xi = \{\mu_{ILD}(\omega), \sigma_{ILD}^2(\omega), \mu_{IPD}(\omega; C_l), \sigma_{IPD}^2(\omega; C_l)\} \).

### 6.4 Source Separation Model

#### 6.4.1 Parameter Estimation from Mixtures

The parameters characterising the interaural cue probability models, introduced in Section 6.3, can be estimated for a specific source \( l \). This would have been a trivial problem up on availability of dominant source information for each TF bin. However, the source \( l \) dominating a specific TF slot is not directly observable from the mixtures, thus it is
usually considered as latent variable. On the other hand, \( l \) can be inferred from the interaural cues and observed models. Therefore, the EM algorithm is selected to estimate the parameters, creating a relationship with the expectation of \( l \). The seven variables defined in \( C_l \) are treated in the EM as hidden variables. Specifically, they are modelled as discrete random variables, where the sets of allowable values are specified a priori, as was also done in [Mandel et al., 2010].

In order to define the log-likelihood of the observations, the parameters to estimate can be generally named as \( \Omega = \{ \Xi, \iota_l, C_l \} \), where \( \iota_l, C_l \) is the marginal class membership, described as the joint probability of being each TF bin dominated by source \( l \) and parameters \( C_l \). In other words \( \iota_l, C_l = p(l, C_l) \). The log-likelihood of the observations can be then defined as [Mandel et al., 2010]:

\[
\mathcal{L}(\Omega) = \sum_{m, \omega} \log p(\alpha^{\text{ILD}}(m, \omega), \hat{\phi}^{\text{IPD}}(m, \omega)|\Omega) = \sum_{m, \omega} \log \sum_{l, C_l} \iota_{l, C_l} p(\alpha^{\text{ILD}}(m, \omega)|l)p(\hat{\phi}^{\text{IPD}}(m, \omega)|l, C_l),
\]

following the ILD and IPD distributions defined in Equations 6.11 and 6.12, respectively. This definition assumes that the IPD and ILD cues are independent. As a result, the joint probability is written as the product of individual probabilities. Such an assumption may not hold in practice, but it provides a convenient way for dealing with the issues related to the optimisation of the log-likelihood function, as well as the parameter estimation of the probabilistic model. In addition, the number of sources must be specified a priori [Mandel et al., 2010].

### 6.4.2 Expectation-Maximisation (EM)

As already discussed above, to estimate the parameters and the probability of each TF bin, the EM algorithm is employed. During the expectation step, the occupation likelihood of source \( l \) with delays and amplitudes \( C_l \) is calculated for each TF slot, given the observations (i.e. \( \alpha^{\text{ILD}}(m, \omega) \) and \( \hat{\phi}^{\text{IPD}}(m, \omega) \)), and the estimated parameters \( \Omega_{\text{est}} \):

\[
\nu_l(m, \omega; C_l) = \iota_{l, C_l} p(\alpha^{\text{ILD}}(m, \omega)|l)p(\hat{\phi}^{\text{IPD}}(m, \omega)|l, C_l).
\]

(6.15)
This expectation is then used in the maximisation step, to re-estimate the parameters, and maximise the likelihood. The ILD parameters are updated as [Mandel et al., 2010]:

\[
\mu_{i,\text{ILD}}(\omega) = \frac{\sum_{m, C_l} \alpha_{i,\text{ILD}}(m, \omega) \nu_l(m, \omega; C_l)}{\sum_{m, C_l} \nu_l(m, \omega; C_l)},
\]

\[
\sigma_{i,\text{ILD}}^2(\omega) = \frac{\sum_{m} (\alpha_{i,\text{ILD}}(m, \omega) - \mu_{i,\text{ILD}}(\omega))^2 \sum_{m, C_l} \nu_l(m, \omega; C_l)}{\sum_{m, C_l} \nu_l(m, \omega; C_l)}, \quad (6.16)
\]

whereas the IPD residual parameters are updated as:

\[
\mu_{i,\text{IPD}}(\omega; C_l) = \frac{\sum_{m} \hat{\phi}_l(m, \omega; C_l) \nu_l(m, \omega; C_l)}{\sum_{m} \nu_l(m, \omega; C_l)},
\]

\[
\sigma_{i,\text{IPD}}^2(\omega; C_l) = \frac{\sum_{m} (\hat{\phi}_l(m, \omega; C_l) - \mu_{i,\text{IPD}}(\omega; C_l))^2 \nu_l(m, \omega; C_l)}{\sum_{m} \nu_l(m, \omega; C_l)}.
\]

(6.17)

In addition, also the marginal class membership is updated, as following:

\[
\iota_{i,C_l} = \frac{1}{B} \sum_{m, \omega} \nu_l(m, \omega; C_l),
\]

(6.18)

where \(B\) is the number of TF bins. Since it has been proved in [Mandel et al., 2010] that, the use of frequency-dependent parameters provides better results, also here the frequency dependence is applied to every parameter.

Through the proposed method, the direct sound and the first early reflection interaural cues are modelled. However, the reverberation is still not included in the model. Therefore, a garbage source, is also utilised [Mandel et al., 2010]. In other words, if the number of sources in the mixtures is \(L\), \(L + 1\) Gaussians are generated for the GMM. The garbage source is designed to be considered as the source dominating the TF bins that are not dominated by any of the other sources. This idea come from the consideration that, while direct sound and first reflection have a main directional component, the reverberation is commonly assumed to be diffuse. The garbage source also allows the parameters of the main sources to be estimated more accurately, since poor fitting TF bins are not included in their probability evaluations [Mandel et al., 2010].

Once the EM algorithm has run a certain number of iterations, that were set a priori, the model parameters are estimated. From them, probabilistic soft masks are generated for each source, marginalising over the estimated coefficients in \(C_l\):

\[
M_l(m, \omega) = \sum_{C_l} \nu_l(m, \omega; C_l).
\]

(6.19)
The separated source signal \( l \) can be finally obtained as:

\[
\hat{y}_{i,l}(m, \omega) = y_i(m, \omega)M_l(m, \omega), \quad \forall m, \forall \omega.
\] (6.20)

### 6.4.3 Initialisation Method

The initialisation part plays a crucial role for the EM algorithm performance, since the log likelihood is not convex. A poor initialisation can lead EM to find the local maxima, affecting the source separation results.

In [Mandel et al., 2010], only the direct sound was used to model the source, thus, the estimation of the parameters utilised to initialise \( \eta_{l,C_l} \), was done through a phase transform based algorithm [Aarabi, 2002]. However, in the proposed method, also the first reflection is employed. Therefore, a different algorithm has to be utilised, which would be able to estimate every parameter included in \( C_l \). Assuming to have availability of RIRs recorded through a multichannel array of microphones, placed at the same binaural listener position, source and image source positions can be estimated through the ISDAR method, that was described in Chapter 4. This represents the second novel contribution of this chapter with respect to the literature.

ISDAR is based on spherical coordinates, considering the listener at the centre of the coordinate system. The radial distances of the source and image source are calculated as

\[
\rho_{e,l} = \frac{1}{M} \sum_{i=1}^{M} \hat{n}_{e,i,l} c_0,
\]

where \( c_0 \) is the sound speed, and \( \hat{n}_{e,i,l} \) is either the estimated direct sound \( (e = 0) \) or first reflection \( (e = 1) \) TOA of the \( i \)-th microphone. Through the azimuth directions of arrival (DOAs) \( \Theta_{e,l} \), the source and image source positions in the Cartesian coordinate system are given by:

\[
b_{x,e,l} = \rho_{e,l} \cos \Theta_{e,l}; \quad b_{y,e,l} = \rho_{e,l} \sin \Theta_{e,l}.
\] (6.21)

TOAs are estimated through the novel clustered dynamic programming projected phaseslope algorithm (C-DYPSA) (for further details about this algorithm, please refer to Chapter 4, Section 4.2.5), whereas DOAs are estimated through the delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988]. The parameters estimated by ISDAR are used to initialise the GMMs (their variance is initially set to be one).

Regarding the other probability distributions, they are differently initialised. The exact value of the ILD prior mean is estimated, as in [Mandel et al., 2010], utilising a set of synthetic binaural room impulse responses, using a regression on ITD, frequency, and interaction terms up to the third order. On the other hand, the garbage source is initialised to have uniform \( \eta_{l,C_l} \), a uniform distribution across IPD, and an ILD distribution with 0 mean across the frequencies.
6.5 Experiments and Results

6.5.1 Datasets

BRIRs were recorded in four rooms, being characterised by having different size and reverberation time (RT60) (or, equivalently, average absorption coefficients $\bar{\alpha}$). The four rooms are named as “Vislab”, “Digital World Research Centre” (DWRC), “BBC usability laboratory” (BBC UL), and “Studio1”. Their pictures are shown in Figure 6.3, whereas a schematic representation of their plans, together with the respective recording setup, are reported in Figure 6.4. Their RT60s and the average absorption coefficients $\bar{\alpha}$, calculated through the inverse of the Sabine’s equation [Kuttruff, 2009] (see Equation 4.15), are reported in Figure 6.5. Two different dummy heads were employed (i.e. a Cortex Manikin Mk2 Binaural Head and Torso Simulator and a Neumann KU100 dummy head), depending on the availability during the recording session. In addition, to obtain data useful for the EM algorithm initialisation, a 48-channel bi-circular array with a typical microphone spacing of 21 mm and an aperture of 212 mm was utilised. Dummy head and double circular array were recorded separately, to avoid interference
given by proximity. All the BRIR and RIR recordings were made utilising the sampling frequency of $f_s = 48$ kHz, employing the swept-sine technique [Farina, 2000].

Beyond all of this, other two measures, describing the dataset peculiarities, are defined: the direct to reverberant ratio (DRR) [Zahorik, 2002], and the average target-interferer separation angle (AVG-TISA). These two additional measures are important since they will, later in this chapter, allow a more accurate discussion, over the different separation performance achieved, comparing the datasets. DRR was calculated as the ratio between the energy carried by the direct sound and the rest of the BRIR. Instead, AVG-TISA was calculated as the average angle separating the target source from the interferer, considering all the possible target-interferer combinations. DRR and AVG-TISA characterising the four datasets are reported in Table 6.1, together with the related frequency-dependent RT60s.

**The Characteristics of the Recorded Rooms.** Vislab is an acoustically treated room at the University of Surrey, where the “Surrey Sound Sphere”, having radius of 1.68 m, was assembled Coleman et al. [2014a]. The room has dimensions $7.79 \times 6.10 \times 3.98$ m$^3$, and the sphere was built having centre on the Cartesian coordinate point $(3.99; 3.95; 1.62)$ m, considering a vertex of the room as the centre of coordinates. $L_{TOT}^{\text{TOT}} = 12$ loudspeakers clamped on the sphere equator, at a height of 1.62 m, with
azimuth 0, ±30, ±60, ±90, ±110, ±135, ±180 degrees, were selected for these experiments, considering 0 degree, as the frontal direction with respect to the dummy head. The dummy head employed was the Cortex Manikin Mk2 Binaural Head and Torso Simulator. Both dummy head and double-circular microphone array were placed at the centre of the sound sphere. The RT60 value is almost constant over the one third octave bands between 500 Hz and 5 kHz, and its average over these frequencies, is about 300 ms.

DWRC is another room at the university of Surrey, utilised by the DWRC research group. Furnished as a living room-like area, it reproduces a typical acoustical environment that can be found at homes. It has a shoebox-like shape, with dimensions $5.98 \times 4.27 \times 2.32 \, \text{m}^3$. A Cortex Manikin Mk2 Binaural Head and Torso Simulator stood on a sofa, with position related to the room, in Cartesian coordinates $(3.86; 0.59; 1.00) \, \text{m}$. $L^{\text{TOT}} = 3$ loudspeakers were placed at the front of the dummy head, with angles $0^\circ$ and $\pm 27^\circ$. The double concentric circular array of 48 microphones was positioned right behind the dummy head, i.e. at coordinates $(3.86; 0.33; 1.00) \, \text{m}$. The RT60 is constant over third octave bands between 500 Hz and 5 kHz, and the average among these bands is 267 ms.

BBC UL is a room at the BBC research and development centre, in Salford, UK. Similarly to DWRC, it is furnished to resemble a typical living room environment. Its shape
is approximated as a shoebox of dimensions $5.57 \times 5.44 \times 2.91 \text{m}^3$. Considering one of the room vertices as the centre of the coordinate system, a Neumann KU100 dummy head was positioned on an armchair, at the Cartesian coordinates $(2.52; 2.73; 1.07) \text{m}$. The double-circular array of microphones was placed at the same position. $L^\text{TOT} = 5$ loudspeakers were employed, lying on the head horizontal plane at the azimuth angles of $0^\circ$, $\pm 37^\circ$ and $\pm 110^\circ$. The RT60 is constant over the third octave bands between 500 Hz and 5 kHz, with an average of 274 ms.

Since the RT60s related to the three already introduced rooms were similar, an additional room was chosen to undertake the experiments: Studio1. It is a large recording studio at the University of Surrey, with dimensions $17.08 \times 14.55 \times 6.50 \text{m}^3$, having an RT60 of $940 \text{ms}$, averaged over the third octave bands between 500 Hz and 5 kHz. A Cortex Manikin Mk2 Binaural Head and Torso Simulator was used as dummy head, lying at the point $(7.12; 8.78; 1.34) \text{m}$ of the Cartesian coordinate system, considering one of the room vertices as the centre of the coordinates system. $L^\text{TOT} = 3$ loudspeaker positions were selected, having their height similar to the dummy head one. Their coordinates are $(4.71; 13.65; 1.18) \text{m}$, $(8.65; 12.80; 1.18) \text{m}$, and $(7.12; 13.35; 1.18) \text{m}$. The double-circular microphone array was positioned about $2 \text{m}$ far from the dummy head. Therefore, its recorded data has been manually “tuned”, in order to have the same virtual position of the head, and to be then employed by the initialisation algorithm.

The microphone and loudspeaker positions, as well as the room dimensions, were manually measured through a laser distance meter. In this Chapter, the evaluation metrics are not defined by considering any of the loudspeaker, microphone, or reflector position groundtruth position. Thus, small measurement errors do not highly affect the experimental results. The only dataset that may be affected by these ones is Studio1, where the microphone array RIRs were manually “tuned” for the initialisation algorithm. This can also be seen as one of the causes for the lower Studio1 performance with respect to the other datasets (see later sections).

**The Utterances.** Fifteen utterances having same length were randomly selected from the TIMIT acoustic-phonetic continuous speech corpus [Garofolo et al., 1993]. Due to its low sampling frequency (i.e 16 kHz), TIMIT may be nowadays considered as composed of low quality recordings. However, it is deemed to be good enough for the experiments of this chapter. In general, the IPD observations degrade for frequencies above 5-6 kHz, that is lower than the TIMIT Nyquist frequency of 8 kHz. For each combination of target source and interferer(s), $U = 15$ random combinations of the fifteen utterances were selected and tested. Therefore, the number of mixtures generated and tested for each dataset is given by:

$$\Upsilon = \frac{L^\text{TOT}!}{L!(L^\text{TOT} - L)!} U,$$

(6.22)
where the symbol “!” represents the factorial operation, and \( L \) is the number of sources in the mixture.

These utterances were recorded with a sampling frequency three times lower than the BRIR one (i.e. 16 kHz instead of 48 kHz). Thus, during the generation of the material for the experiments, the BRIRs were decimated by a factor 3, before performing the convolution with the speech signals. Furthermore, the utterances were normalised before the convolution, in order to have the same root mean square energy.

### 6.5.2 Evaluation Metrics

**Source to distortion ratio (SDR).** The source to distortion ratio (SDR) metric is based on energy ratios, thus, is typically reported in dB. Following Equation 6.4, the ideal target signal \( l \), that arrives at each \( i \)-th channel free from any interference and noise, can be defined, for each TF bin, as:

\[
y_{i,l}^{\text{tar}}(m, \omega) = x_l(m, \omega)I_{i,l}(\omega).
\]  

Hence, the source \( \hat{y}_{i,l}(m, \omega) \), separated by a source separation method as defined in Equation 6.20, can be decomposed as [Vincent et al., 2006]:

\[
\hat{y}_{i,l}(m, \omega) = y_{i,l}^{\text{tar}}(m, \omega) + E_{\text{interf}} + E_{\text{noise}} + E_{\text{artif}},
\]  

where \( E_{\text{interf}} \) is the interference error term, \( E_{\text{noise}} \) the noise error term, and \( E_{\text{artif}} \) represents the errors provided by general artefacts. From Equation 6.24, different ratios can be calculated to evaluate the separation accuracy, depending on the error term that one wants to emphasise. SDR is chosen here, since it takes into account all the three error terms. It is calculated as [Vincent et al., 2006]:

\[
SDR = 10 \log_{10} \left( \frac{||y_{i,l}^{\text{tar}}(m, \omega)||}{||E_{\text{interf}} + E_{\text{noise}} + E_{\text{artif}}||^2} \right),
\]  

where \( || \cdot || \) represents the Euclidean norm operator. Once the SDR for each of the \( \Upsilon \) combinations of sources is obtained, the overall result of the dataset is calculated as the mean of these samples:

\[
SDR = \frac{1}{\Upsilon} \sum_{\upsilon=1}^{\Upsilon} SDR_{\upsilon},
\]  

where \( \upsilon \) is the tested mixture index.
Perceptual evaluation of speech quality (PESQ). The perceptual evaluation of speech quality (PESQ) is utilised to evaluate the speech intelligibility, considering distortions that are different from the ones analysed by SDR, for instance, signal delays or losses. It was firstly designed to evaluate quality of speech transmitted through telephonic channels, however, during the last decade, it has been widely employed to evaluate speech enhancement system quality [Loizou, 2013]. This kind of evaluation is related to the Mean Opinion Score (MOS) of human subjective assessments, therefore, the PESQ unit of measure is MOS. Before proceeding with the PESQ value calculation, \(\hat{y}_{i,l}(m, \omega)\) and \(y^\text{tar}_{i,l}(m, \omega)\) are aligned in time, in terms of both amplitudes and delays, by employing Wiener filters [Loizou, 2013]. Through two parameters that model symmetric and asymmetric disturbances, a parametric function is then employed, mapping the differences between the processed version of \(\hat{y}_{i,l}(m, \omega)\) and \(y^\text{tar}_{i,l}(m, \omega)\), to subjective assessment results. As for the SDR, the overall PESQ is calculated as the mean over all the target-interferer combinations:

\[
PESQ = \frac{1}{\Upsilon} \sum_{\upsilon=1}^{\Upsilon} PESQ_{\upsilon}.
\]

(6.27)

### 6.5.3 Control Masks

One of the main requirements that are needed to perform a fair evaluation of source separation systems are the so called performance bounds [Vincent et al., 2012]. Reference signals must be generated from the mixtures, to be compared with the output of the proposed source separation methods. Regarding the lower bound, random masks were generated and utilised to extract signals from the mixture.

On the other hand, regarding the upper bound, it has been chosen to calculate a near-optimal binary mask, through Oracle estimators [Vincent et al., 2007]. The amount of distortion, related to the \(l\)-th, in the mixture is given by:

\[
E_l^\text{dist}(m, \omega) = \sum_i [\hat{y}_{i,l}(m, \omega) - y_{i,l}^\text{tar}(m, \omega)]^2 \\
+ \sum_l \sum_i y_{i,l}^\text{tar}(m, \omega)^2, \quad \forall m, \forall \omega.
\]

(6.28)

Therefore, the near-optimal binary mask can be generated for each source \(l\) and each TF bin. This can be done comparing its distortion with respect to the distortions related to the other sources in the mixture:

\[
M_l^{\text{ora}}(m, \omega) = \begin{cases} 
1, & E_l^\text{dist}(n, \omega) < E_{l'}^\text{dist}(m, \omega), \quad \forall l \neq l' \\
0, & \text{otherwise}.
\end{cases}
\]

(6.29)
where $l'$ is an index referring to a sources in the mixture, that is different from $l$. In the experiments, Equation 6.28 is calculated utilising instead of $y_{m}^{\text{IR}}(t, \omega)$, its version obtained by windowing the related BRIR direct sound, as in [Mandel et al., 2010]. An example of generated Oracle mask is depicted in Figure 6.6, together with the respective soft masks estimated by both MESSL and ER-MESSL.

6.5.4 Source Separation Experiments

The experiments performed were mainly focused on analysing the source separation performance, employing mixtures composed of two sources ($L = 2$), i.e. the target and interfering source. However, in addition to these, mixtures composed of three sources ($L = 3$) were also evaluated, i.e. one target and two interfering sources. The number of maximum iterations for the EM algorithm was set, for all the experiments, to be 16. The BRIRs and the utterances introduced in Section 6.5.1 were utilised to create the reverberant mixtures described in Equation 6.3. Both $L = 2$ and $L = 3$ experiments were designed to compare the novel method, with the one in the literature that considers only direct sound for the IPD model [Mandel et al., 2010]. In addition, also the results obtained by applying the ideal masks are reported as reference. The performance in separating the target source is compared, by calculating, for both the methods and the masks, the related SDR and PESQ.

**One Interfering Source.** The source separation results, considering the one interfering source case, are calculated utilising the SDR and PESQ metrics, described in Section 6.5.2. In Tables 6.2 and 6.3, the source separation performance are evaluated averaging SDR and PESQ over all the $\Upsilon$ possible target-interferer combinations, exploiting Equations 6.26 and 6.27.
Table 6.2: SDRs obtained separating the target source from a mixture with one interferer source.

<table>
<thead>
<tr>
<th></th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>−0.78</td>
<td>−0.61</td>
<td>0.15</td>
<td>0.06</td>
<td>−0.30</td>
</tr>
<tr>
<td>MESSL [Mandel et al., 2010]</td>
<td>4.44</td>
<td>2.54</td>
<td>5.47</td>
<td>0.58</td>
<td>3.26</td>
</tr>
<tr>
<td>ER-MESSL</td>
<td>4.91</td>
<td>2.68</td>
<td>5.67</td>
<td>0.67</td>
<td>3.48</td>
</tr>
<tr>
<td>ORACLE DS</td>
<td>6.22</td>
<td>5.04</td>
<td>6.82</td>
<td>0.68</td>
<td>4.69</td>
</tr>
</tbody>
</table>

Table 6.3: PESQs obtained separating the target source from a mixture with one interferer source.

<table>
<thead>
<tr>
<th></th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.36</td>
<td>1.45</td>
<td>1.32</td>
<td>1.37</td>
<td>1.38</td>
</tr>
<tr>
<td>MESSL [Mandel et al., 2010]</td>
<td>1.96</td>
<td>1.93</td>
<td>2.06</td>
<td>1.82</td>
<td>1.94</td>
</tr>
<tr>
<td>ER-MESSL</td>
<td>2.00</td>
<td>1.93</td>
<td>2.07</td>
<td>1.82</td>
<td>1.96</td>
</tr>
<tr>
<td>ORACLE DS</td>
<td>2.34</td>
<td>2.45</td>
<td>2.45</td>
<td>1.96</td>
<td>2.30</td>
</tr>
</tbody>
</table>

Table 6.4: P-values corresponding to the paired t-test comparing the results given by MESSL and ER-MESSL, with one interferer.

<table>
<thead>
<tr>
<th></th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDR</td>
<td>0.00 %</td>
<td>8.56 %</td>
<td>0.00 %</td>
<td>11.95 %</td>
</tr>
<tr>
<td>PESQ</td>
<td>0.00 %</td>
<td>38.4 %</td>
<td>4.61 %</td>
<td>5.79 %</td>
</tr>
</tbody>
</table>

Table 6.2 shows that ER-MESSL, the source separation method proposed within this chapter, outperforms the baseline (i.e. the MESSL method [Mandel et al., 2010]), when applied to any of the four datasets described in Section 6.5.1. These results also show that, in general, DWRC and Studio1 are problematic datasets, producing low SDR values for both the methods (although the proposed ER-MESSL is still better than MESSL). Observing Table 6.3, although the DWRC and Studio1 datasets provide a PESQ for ER-MESSL that is comparable to the one produced by MESSL, the PESQ result, that was averaged over all the datasets, shows that ER-MESSL performs the better. In fact, in Vislab and BBC UL, ER-MESSL presents higher value of PESQs than MESSL. The reason for the issues faced by DWRC and Studio1 can be found observing Table 6.1: they have low DRRs and narrow AVG-TISAs. Low DRR entails difficulties for both the algorithms, since the IPD curve, that was described in Figure 6.2, is highly distorted by the strong reverberation. At the same time, narrow AVG-TISA affects the overall results, since small angles between target and interferer correspond to small variations between the IPD and ILD cues related to the two signals in the mixture.

Assuming the Υ values of SDR and PESQ being normally distributed, the paired t-test was employed, to determine if the results generated through MESSL and ER-MESSL are significantly different. In Table 6.4, p-values are reported, which corresponds to the null-hypothesis probability to be accepted. It is interesting to note that, considering a significance level of 5 %, results from Vislab and BBC UL reject that hypothesis,
whereas DWRC and Studio1 accept it. These results confirm what was already shown by Tables 6.2 and 6.3, determining the improvement given by ER-MESSL lower in DWRC and Studio1 than in BBC UL and Vislab.

The SDR results, for the four datasets, can also be reported as functions of the target source position. In particular, fixing the target source at every angle available, the SDR was calculated for every corresponding position of the interferer, and then averaged. Results are reported in Figure 6.7. Following the front/back confusion, which is produced by the IPD cue [Wenzel et al., 1993], the loudspeakers, that were positioned behind the listener, are considered as projected to the front. It is clear that the proposed ER-MESSL performs better than MESSL for almost every position of the target source. The only exceptions are $-37^\circ$ for the BBC UL dataset and $27^\circ$ for the DWRC. As also depicted in Figures 6.3, in DWRC, the loudspeaker positioned at $27^\circ$ stood next to a chest of drawers. Therefore, the localisation of the first reflection, using ER-MESSL, introduced errors. Similarly, in BBC UL, the loudspeaker corresponding to the $-37^\circ$ angle was positioned next to a lateral wall.

Results reported through Figure 6.7, in particular the ones referring to Vislab and BBC UL, represent the characteristic behaviour of binaural source separation methods. In fact, separation achieves higher performance when the target signal comes from a frontal direction (i.e. $0^\circ$), rather than a lateral one [Makous and Middlebrooks, 1990]. To provide a better visualisation of this characteristic, the same results are reported in Figure 6.8, nevertheless, without assuming the front/back confusion, thus, placing every loudspeaker at its exact location. Interesting features, that can be seen through this kind of visualisation, are the front/back and left/right asymmetries in the performance.

**Figure 6.7:** SDRs for different target source positions. The blue lines with circular marks refer to the proposed ER-MESSL, whereas the red ones with crossed marks refer to MESSL [Mandel et al., 2010].
6.5.4. Source Separation Experiments

Figure 6.8: Polar plot of the SDRs (in dB) for “Vislab” (a), and “BBC UL” (b), regarding MESSL [Mandel et al., 2010] (red line with crossed marks) and ER-MESSL (blue line with circular marks).

Figure 6.9: SDRs for different interferer positions, having fixed the target source at $0^\circ$. The solid lines with circular marks represent the results related to ER-MESSL, whereas the dashed lines with crossed marks refer to MESSL [Mandel et al., 2010]. The blue lines correspond to Vislab, the green lines to BBC UL, the black ones to DWRC, and the red ones to Studio1.

These are known characteristics for IPD based binaural localisation methods, as it was studied in [Katz and Noisternig, 2014]. These differences in the performance are mainly caused by the employed spherical representation of the human head (see Chapter 2). The head asymmetries, that are present in reality, produce variations in the IPD model that are not taken into account.

Beyond these results, the source separation performance, in terms of SDR, were also calculated by fixing the frontal loudspeaker ($0^\circ$ of azimuth) as target source. The position of the interferer was then varied and the results reported in Figure 6.9. This kind of visualisation is provided to be coherent with the state-of-the-art, since it is the typical way that results are reported in the literature for source separation. Symmetric source
Table 6.5: SDRs obtained separating the target source from a mixture with two interferer sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>-4.97</td>
<td>-5.62</td>
<td>-4.49</td>
<td>-13.28</td>
<td>-7.09</td>
</tr>
<tr>
<td>MESSL [Mandel et al., 2010]</td>
<td>2.01</td>
<td>0.63</td>
<td>3.08</td>
<td>0.52</td>
<td>1.56</td>
</tr>
<tr>
<td>ER-MESSL</td>
<td>1.93</td>
<td>0.45</td>
<td>2.80</td>
<td>0.47</td>
<td>1.41</td>
</tr>
<tr>
<td>ORACLE DS</td>
<td>5.29</td>
<td>4.29</td>
<td>6.48</td>
<td>1.05</td>
<td>4.28</td>
</tr>
</tbody>
</table>

Table 6.6: PESQs obtained separating the target source from a mixture with two interferer sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.35</td>
<td>1.43</td>
<td>1.42</td>
<td>1.44</td>
<td>1.41</td>
</tr>
<tr>
<td>MESSL [Mandel et al., 2010]</td>
<td>1.81</td>
<td>1.80</td>
<td>1.87</td>
<td>1.81</td>
<td>1.82</td>
</tr>
<tr>
<td>ER-MESSL</td>
<td>1.66</td>
<td>1.73</td>
<td>1.83</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>ORACLE DS</td>
<td>2.23</td>
<td>2.27</td>
<td>2.43</td>
<td>2.13</td>
<td>2.27</td>
</tr>
</tbody>
</table>

positions with respect to the sagittal plane were considered as one unique direction. The SDR of such directions was reported as their mean. By observing the results of Vislab (that is the only dataset having loudspeaker positions all around the listener), it is evident that the SDR is directly proportional to the target-interferer separation angle (TISA), until about 60°, where it starts to drop. Similar results were also observed in [Alinaghi et al., 2014]. Regarding the other datasets, the lower number of recorded loudspeaker positions leads in a lack of data to be visualised. However, it is still possible to note that, for both DWRC and BBC UL, the proposed ER-MESSL has higher performance than MESSL, for the available interfer positions (i.e. two angles for DWRC and one for BBC UL). Instead, for Studio 1, MESSL seems to have higher performance than ER-MESSL for the only TISA available. Nevertheless, Studio1 confirms to be a problematic dataset, with both the method SDRs being lower than 1 dB.

Two Interfering Sources. After having analysed the source separation performance, with respect to the proposed ER-MESSL and the baseline MESSL [Mandel et al., 2010], the case with more than one interfering source was also investigated. Utilising the same datasets described in Section 6.5.1, the performance on separating a target source, given a mixture containing other two interferer sources, is evaluated.

In Tables 6.5 and 6.6 the source separation performance is reported as average between SDR and PESQ over all the Υ possible target-interferer combinations, exploiting Equations 6.26 and 6.27. These results confirm the fact the ER-MESSL is able to separate sound sources from mixtures, also considering underdetermined conditions. However, SDRs and PESQs are considerably low for every dataset, considering both the methods. This is due to the fact that, with an additional interferer, the IPD and ILD cues become complicated to model, with high probability of having multiple sources in a single TF
Table 6.7: Ranges around the initialised parameter values, that are tested by the EM algorithm to find the maximum of the log-likelihood function.

<table>
<thead>
<tr>
<th></th>
<th>Vislab</th>
<th>DWRC</th>
<th>BBC UL</th>
<th>Studio1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_l^{DF}$, $n_l^{DS}$, $n_l^{ST}$</td>
<td>$\pm 0.13$ ms</td>
<td>$\pm 0.13$ ms</td>
<td>$\pm 0.19$ ms</td>
<td>$\pm 0.31$ ms</td>
</tr>
</tbody>
</table>

bin. This issue mainly affects ER-MESSL, where it produces lower SDRs and PESQs with respect to MESSL. This is caused by modelling the IPD cue through a non-linear function (see Equation 6.7), instead of a line as it was in MESSL [Mandel et al., 2010]. It entails the model to overfit the observations in the TF space [Hawkins, 2004].

### 6.5.5 Implementation Issues

As discussed in Section 6.4.1, and following what was also done in [Mandel et al., 2010], the parameters in $C_l$ are updated during the EM algorithm, allowing a set of values specified a priori. Since the proposed ER-MESSL is composed of seven parameters, exploring this seven dimensional parameter space as it was done in [Mandel et al., 2010] is computationally more expensive. Therefore, for the experiments, the amplitude values $P_{e,i,l}$ were fixed, without allowing the EM algorithm to update them. At the same time, the size of the interval chosen to search the best time-dependent parameters $n_l^{DF}$, $n_l^{DS}$ and $n_l^{ST}$ was empirically found. The sizes of the sets of allowed values, for each dataset, is reported in Table 6.7.

### 6.6 Conclusion

Through this chapter, ER-MESSL was proposed, a novel source separation method, modelling the IPD cues considering both the information carried by the first early reflection and the direct sound. Furthermore, the image source locator ISDAR was utilised to initialise the employed EM algorithm.

Experiments were performed by applying ER-MESSL to four reverberant environments, and comparing the results with the source separation performance provided by MESSL [Mandel et al., 2010]. Considering the scenario where only one interferer source is present together with the target, the experiments showed the proposed ER-MESSL to outperform MESSL. In fact, the proposed model separates the sources with both higher SDR and PESQ. Typical binaural source separation behaviours were also observed among the results. For instance, better results were achieved when the target source was frontally positioned with respect to the listener. Moreover, narrow angles between target and interferer degrade the results, as well as low DRRs. Additional experiments were also
performed considering a double interferer scenario. Although they demonstrate the capability of ER-MESSL of localising and separating target sources in underdetermined conditions, it did not perform better than MESSL. This is due to the fact that, if too many competing sources are present in the mixture, the high number of parameters modelling the IPD cue in ER-MESSL, lead the model to overfit the observed data.

Following the analysis made, future work may be conducted on testing this method in other audio signal processing areas, exploiting its high performance with few interferers present in the mixture. Alternatively, the model can be extended to work with multichannel arrays of microphones.
Chapter 7

Conclusion

The main research problem that was tackled by this thesis concerned the acoustic reflector localisation. This problem was evaluated under the assumption of having availability of recorded multichannel room impulse responses (RIRs). With this respect, four novel 3D reflector localisation methods were proposed: three belonging to the image-source reversion group (i.e. the image source direction and ranging-loudspeaker image bisection (ISDAR-LIB), the mean-ISDAR-LIB, and the median-ISDAR-LIB); and one that is representative of the direct localisation group (i.e. the ellipsoid tangent sample consensus (ETSAC)).

The three image-source reversion methods rely on localising the image sources. This was done by proposing the combination of a novel time of arrival (TOA) estimator, i.e. the clustered dynamic programming projected phase slope algorithm (C-DYPSA) and the delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988], as direction of arrival (DOA) estimator. In fact, having knowledge of the reflection TOA allows the estimation of the distance (range) between image source and listening position, whereas the DOA provides information about azimuth and elevation of the image source. Thus, the three image source spherical coordinates were obtained. This algorithm was named as ISDAR. The LIB algorithm [Tervo and Tossavainen, 2012] was then employed to revert the image source method [Allen and Berkley, 1979]. In other words, by assuming the source position as known, the plane bisecting the line connecting source and estimated image source at their midpoint was defined as the desired reflector. The method developed by joining ISDAR and LIB was proposed as the ISDAR-LIB method. Mean-ISDAR-LIB and median-ISDAR-LIB were also proposed. They applied post-processing algorithms to ISDAR-LIB, combining multiple loudspeaker information. In particular, the mean and median of the estimated planes were calculated, respectively.
Regarding ETSAC, it was proposed as a direct localisation method, thus, it directly estimates the reflector position, without any preliminary step. It is based on generating 3D ellipsoids, one for each source-sensor combination. These ellipsoids were constructed including the reflection TOA information: the major axis was chosen to have the same length as the reflection path, setting source and sensor as foci. This was done to exploit the geometrical property defining the sum of the distances between a point on the ellipsoid surface and the two foci as constant. Algorithms typically employed by imaging researchers, such as the random sample consensus (RANSAC) [Fischler and Bolles, 1981], were utilised to find the plane that is common tangent to every ellipsoid. This plane corresponds to the reflector position.

Performance of the proposed methods were compared to two baselines, to discover the better approach for reflector localisation, given compact microphone array RIRs. Simulation of recording conditions, with background noise and microphone position displacements, were used to test the methods by varying room size, absorption coefficient and direct to noise ratio (DNR). Results showed ETSAC performing better than the others, in every condition. Furthermore, it was observed that every method produced large errors, given small environments. In addition, the general trend of the errors was directly proportional to the absorption coefficient. Mean-ISDAR-LIB and ETSAC were found to be robust to low DNR conditions. Experiments with recorded RIRs were also performed and divided into three main tasks. Firstly, the image localisation algorithms proposed in [Dokmanić et al., 2013] and [Tervo and Tossavainen, 2012] were compared to the novel ISDAR. Results show that ISDAR provided the better performance. The second part of the experiments compared the novel image-source reversion methods ISDAR-LIB, mean-ISDAR-LIB and median-ISDAR-LIB, with the proposed direct localisation method ETSAC. Results show that ETSAC localised reflectors with an average of 5 cm root mean square error (RMSE), i.e. 42% lower than the best alternative method, here tested. However, as ISDAR-LIB ran 200 times faster than ETSAC, it has an advantage for fast processing applications, such as tracking. In the last experiment, the reflectors estimated through ETSAC were converted into their corresponding image sources, and compared with the image locators in [Dokmanić et al., 2013] and [Tervo and Tossavainen, 2012]. This showed the percentage of gross errors dropping drastically from 38% (multilateration [Dokmanić et al., 2013]) to 13% (ETSAC).

Future work may consider exploring alternative microphone array arrangements over a large set of rooms, optimal beamformer designs for DOA estimation, and robust methods for multiple loudspeaker ISDAR-LIB post-processing.

Having introduced novel methods to identify acoustic reflector positions, two application areas were explored: spatial audio and blind source separation. Regarding spatial audio,
a new method was proposed to parameterise information related to the acoustic of a recorded environment, in order to create reverberant spatial audio objects (RSAOs). On the other hand, considering blind source separation, a novel method was proposed to separate speech signals. This method takes into account the first early reflection during the definition of the interaural phase difference (IPD) model.

Starting from RSAO, different algorithms were proposed to extract the environment acoustical information from recorded RIRs. This information was then represented as parameters, to be sent through the channel to a decoder. This decoder was able to plausibly recover the original RIRs from them. Depending on the part of RIR under investigation, different approaches were undertaken. For instance, direct sound and early reflections were analysed in both time and frequency domain. C-DYPSA was utilised to parameterise TOAs, whereas DSB to parameterise DOAs. The frequency domain was analysed using the linear predictive coding (LPC), and the coefficients of the estimated all-pole filter were then defined as parameters. On the other hand, the late reverberation was decomposed in octave bands. Each of them were analysed in the time domain, and the decaying envelopes fitted by exponential functions. The related parameters were defined as these frequency-dependent exponential coefficients. Having availability of all of these parameters, RSAOs were then generated, at the decoding stage, by recreating the RIRs and convolving them with downmixed audio signals. The direct sound and early reflections parts of the RSAOs were mapped to the available loudspeakers through vector base amplitude panning (VBAP) [Pulkki, 1997]. The late reverberation was sent to all the loudspeakers as decorrelated signals.

A formal experiment was also carried out, by performing subjective assessments. Three different perceptual acoustic properties were evaluated: the source distance, the room size, and the listener envelopment. These characteristics were varied by manually modifying the RSAO parameters, that were extracted from recorded RIRs. Results showed that, editing the RSAO parameters, the perceived room size and source distance were altered. Moreover, the envelopment perceived by the listener mostly depends on the reproduction system available.

Future work may consider refining the parametrisation process, also including additional parameters to better describe the RSAO properties. Furthermore, production tools for parameter editing may be developed, also relating the parameter values to perceptual properties.

Regarding the blind source separation application, a novel method was presented, named as early reflection model-based expectation maximisation source separation and localisation (ER-MESSL), which takes into account both the first early reflection and the direct sound of the target signal. It is based on the work that was firstly presented in [Mandel
et al., 2010], i.e. the model-based expectation maximisation source separation and localisation (MESSL) algorithm. There, binaural features related to the direct sound IPD and interaural level difference (ILD) were modelled observing mixtures. Utilising these models, a Gaussian mixture model (GMM) was employed to determine the probability of a source to be dominant, considering each mixture time-frequency (TF) bin. Based on these probabilities, a soft TF mask was then created to separate the signals. In the novel method that is proposed within this thesis, also features related to the first reflection IPD were extracted, and utilised for the model creation.

Experiments were then carried out, employing recordings made within four rooms, having different reverberation time (RT60) and size. The main experiment involved the use of speech mixtures composed of the target source and one interferer. Results showed ER-MESSL performing better than MESSL in terms of separation quality. In addition, it was proved that, as for MESSL, using ER-MESSL the best performance was achieved when the target source was placed frontally with respect to the listener position. Moreover, results showed that low direct to reverberant ratios (DRRs) and narrow angles between target and interferer degraded the performance. Another experiment was also performed, by adding an additional interferer within the mixture. Although proving that ER-MESSL can be also employed for underdetermined scenarios, in this case, MESSL showed higher performance. This is due to the large number of parameters utilised in ER-MESSL to model the IPD, that, in scenarios where more than two speakers are involved, resulted in overfitting the data.

As future work, it may be interesting to apply ER-MESSL in other audio signal processing areas. In fact, the good performance provided with one interferer suggests that ER-MESSL may perform well in conditions with only target source and background noise. Another possible extension of this work may be the creation of a similar method, that can be used with multichannel arrays of microphones.
Appendix A

Multichannel Room Impulse Response Visualisation

A.1 Introduction

A room impulse response (RIR) is a signal characterising the acoustic of a recorded environment. It underpins many audio signal processing research areas (e.g. spatial audio, source separation, source tracking, audio reverberation and dereverberation), since it provides information about microphone and loudspeaker positions, room geometry and room size. Multichannel RIRs are usually utilised to apply algorithms estimating parameters such as directions or times of arrival (DOAs or TOAs, respectively).

Researchers typically utilise databases of recorded RIRs to test their methods. One of the first publicly-available datasets was the one presented in [Wen et al., 2006], where, like in [Hadad et al., 2014], a large number of loudspeaker positions were measured. However, the employment of a uniform linear array (ULA) of microphones limited its applicability, since methods to analyse room acoustics in 3D must exploit at least a 2D configuration of microphones. In [Kayser et al., 2009] and [Erbes et al., 2015], one of the main contributions was given by datasets of binaural recordings. B-format datasets were provided by [Stewart and Sandler, 2010], with microphones being placed in a grid which covered almost the entire plan of the rooms. However, the microphones were spatially too sparse, and algorithms assuming the far-field could not be applied.

Recently, several algorithms for multichannel RIR visualisation have been proposed. They were mainly designed to show different acoustic properties of the recorded rooms. In [Melchior et al., 2010], plane wave decomposition (PWD) was utilised to visualise the amount of energy arriving, over time, to a uniform circular array (UCA) from any DOA.
In [Farina et al., 2011] reflections were visualised using directional multichannel RIRs recorded with an spherical array of microphones. Reflector positions can also be estimated and graphically shown as planes [Dokmanić et al., 2013, Tervo and Tossavainen, 2012]. In [Pätynen et al., 2013], spatio-temporal RIR visualisations were made to analyse concert hall acoustics. Image source location visualisations, related to RIRs, were presented in [Tervo et al., 2013].

In this appendix, new datasets are presented, together with visualisation techniques for multichannel RIRs recorded using a compact UCA within two rooms: the first one at the University of Surrey; the second one at the Emmanuel Church in Guildford. Other datasets recorded using a compact uniform rectangular array (URA) were also made available. The array’s compactness allowed to use algorithms by assuming the far-field, and offered a listener’s perspective of the recorded rooms. To make explicit the information contained in the multichannel RIRs and provide a graphical representation of the rooms involved, room visualisation methods were applied. Section A.2 introduces the visualisation algorithms; Section A.3 presents the datasets and shows the related visualisations; and Section A.4 concludes the discussion.

### A.2 Room Visualisation Techniques

In this section, the proposed visualisation techniques are described, together with the room characteristics that they demonstrate.

**Raw Data and DOA-Time Energy Analysis.** One useful technique to understand the room acoustics is to visualise the DOA of acoustic energy over time. Here, a visualisation similar to [Melchior et al., 2010] was achieved by steering a superdirective beamformer (the superdirective array, (SDA) [Bai and Chen, 2013]) towards every azimuth direction, with a resolution of one degree. The energy arriving from each direction is visualised after calculating the short-term power average by sliding a 0.37 ms Hann window along the beamformed RIRs. The purpose of this appendix is to propose new methods to visualise RIRs. Therefore, the length of this window was empirically derived to enhance the visualisation output quality. In order to find a window length able to enhance the sound quality, other window sizes have to be found. This representation can be considered as evolution of [Hulsebos, 2004], where the author presented a visualisation of multichannel RIRs, generated plotting the raw RIR signals, in their time domain, adjacently one to each others.

**Reflection and Reflector Localisation.** The techniques presented here aim to visualise reflections and reflecting surfaces. A first model was based on image source
localisation. To do so, two parameters were utilised: TOAs and DOAs. The dynamic programming projected phase-slope algorithm (DYPSA) [Naylor et al., 2007], modified to be used with RIRs, is used to extract the TOAs. Based on them, the RIRs were segmented. The segmented signals were then used to calculate the DOA for the early reflections using a 3D delay-and-sum beamformer (DSB) [VanVeen and Buckley, 1988]. Finally, the reflector is drawn as the plane perpendicular to the line intersecting the image source and the loudspeaker and passing through their mid-point. This process was described in Chapter 4, under the name of image source direction and ranging loudspeaker image bisection (ISDAR-LIB).

A second model used ellipsoids to estimate the reflector positions. A set of ellipsoids were generated, having foci on the microphone-source combinations and major axis of the reflection’s path length. A random sample consensus (RANSAC)-based technique [Fischler and Bolles, 1981] was used to find the estimated reflector location, i.e. the common tangent plane to all the ellipsoids generated, as shown in Chapter 4, under the name of ellipsoid tangent sample consensus (ETSAC).

### A.3 Recorded Dataset Visualisation

In this section, the multichannel RIR datasets recorded using a compact microphone array in three different rooms at the University of Surrey are presented. Then, the visualisation technique outputs are shown, when applied to these measurements, and comment on the room acoustic features highlighted.

#### A.3.1 Recorded Multichannel RIR Datasets

Several sets of RIRs are available. Here, the three datasets, summarised in Table A.1, that were used to apply the visualisation methods presented in Section A.2, are described. Further sets are available online\(^1\). Countryman B3 omni lavalier microphones were used for each dataset.

**AudioBooth.** The “AudioBooth” is an acoustically treated room at the University of Surrey. A \( L = 17 \) channel loudspeaker array was mounted on a truncated geodesic

\(^{1}\text{http://cvssp.org/data/s3a/}\)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dimension (m)</th>
<th>RT60 (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Church II</td>
<td>19.68 × 24.32 × 5.97</td>
<td>1500–1200</td>
</tr>
<tr>
<td>Studio1</td>
<td>14.55 × 17.08 × 6.50</td>
<td>1400–1100</td>
</tr>
<tr>
<td>AudioBooth</td>
<td>4.12 × 4.98 × 2.10</td>
<td>413–115</td>
</tr>
</tbody>
</table>
sphere with the equator at 1.02 m elevation. The array comprised nine Genelec 8020B loudspeakers around the equator at 1.68 m radius and $0^\circ, \pm 30^\circ, \pm 70^\circ, \pm 110^\circ$ and $\pm 155^\circ$ in azimuth relative to the centre channel. At $\pm 30^\circ$ and $\pm 110^\circ$, further loudspeakers were placed at $\pm 30^\circ$ elevation. The microphone array, positioned at the centre of the loudspeaker array, was a 48 channel double concentric UCA having 24 microphones evenly spaced around radii of 0.085 m and 0.106 m. A sound-field microphone was also positioned at the centre of the double UCA to avoid up-down ambiguities. RIRs were recorded at a sampling frequency of $f_s = 48$ kHz, by using the log sine sweep method [Farina, 2000].

**Studio1.** RIRs were also recorded in “Studio1”, a large recording studio at the University of Surrey. A total of $L = 15$ loudspeaker positions were placed around the microphone array with distances between 2.0 m and 3.0 m, with four at a height of 1.50 m, eight at 1.18 m and three at 0.30 m. As before, RIRs were recorded at $f_s = 48$ kHz by the log sine sweep method [Farina, 2000]. The loudspeakers were Genelec 1032B and the same 48 channel double UCA was used as for the AudioBooth.

**Church II.** RIRs were recorded in two rooms at Emmanuel Church: the “OldChurch” and “MainChurch”, that will be named as “Church I” and “Church II”, respectively. Visualisations of Church II are given here, whereas the Church I data and documentation are available online. The Church II multichannel RIRs were recorded using Genelec 8030A loudspeakers positioned at $0^\circ$, $\pm 30^\circ$ and $\pm 110^\circ$ in azimuth and $0^\circ$ and $30^\circ$ elevation at a radius of 5 m, giving a total of ten positions. The 48 channel dual UCA and Soundfield microphones were used as for the AudioBooth.

**Further datasets.** Further datasets recorded in “Studio2” and the “Vislab” are available, in each case having $L = 60$ loudspeakers equally spaced around a radius of 1.68 m, and with various positions of a 48 channel uniform rectangular microphone array combined to make a grid of measurement positions. In Studio2, 864 different microphone positions were measured, and in Vislab 384 positions were measured. In each case the maximum length sequence (MLS) technique was used at sampling frequency $f_s = 48$ kHz$^2$.

A.3.2 Multichannel RIRs Visualisation

The multichannel RIR visualisation techniques presented in Sec. A.2 were applied to the recorded data. As shown in Figure A.1, the multichannel RIR raw data representation allows visualisation of the sound waves arriving to the microphones. In addition, the DOA-time energy analysis emphasises the DOAs of each reflection captured. It can be observed a huge difference among the different datasets. In particular, the amount of

$^2$These measurements are available from DOI: http://dx.doi.org/10.15126/surreydata.00808179.
A. Multichannel Room Impulse Response Visualisation

Figure A.1: Raw multichannel RIRs data visualisation and DOA-time energy analysis (here titled beamformed data), relative to the three datasets: AudioBooth (A), Studio1 (B) and Church II (C).

Figure A.2: AudioBooth reflection and reflector estimation, showing the first six reflections (blue circles) due to a single loudspeaker (green star) (A), first reflection (blue circles) of multiple loudspeakers (green stars) simultaneously (B), and reflector estimation through ETSAC (C). The red circles represent the UCA position.

Reflections clearly visible in AudioBooth and MainChurch distinguish them from Studio1. This Studio1 characteristic is due to the cluttering introduced by the measurement setup [Francombe et al., 2015]. Later early reflections in AudioBooth are diffuse, implying the capacity of this room to propagate low frequency modes. From Figure A.2, reflection positions are observable as blue spots over the shoebox geometry recreated from groundtruth. Here, the dataset employed is the AudioBooth. The three subfigures show: how it is possible to extract first-order (represented inside the shoebox) and higher-order reflections using one loudspeaker (Figure A.2 (A)); the possibility of selecting just the first reflection from each loudspeaker (Figure A.2 (B)); performance on localising the reflectors (Figure A.2 (C)).
A.4 Conclusion

A new database of RIR measurements was recorded using a compact microphone array. Visualisation methods were applied to them, highlighting the detail inherent in compact array perspective of the room acoustics. The presented datasets, formatted following the Spatially Oriented Format for Acoustics (SOFA) [AES69, 2015], are available for download.\(^3\)

\(^3\)Measurements available at DOI: http://dx.doi.org/10.15126/surreydata.00808465.
Appendix B

The CISPE Algorithm

B.1 Introduction

An acoustic signal is affected by the environmental characteristics. To identify the signal-environment interaction, the sound received at the listener is defined as the convolution between the reproduced sound and a room impulse response (RIR). The room geometry and the related microphone and source positions define a specific RIR. Knowledge of the room shape can improve algorithms used in various applications such as source separation, speech recognition, media production and music transcription. This also offers a potential in different research areas, including localisation mapping, spatial audio or audio forensics.

To determine the position of each microphone utilised, different algorithms are available. Under the assumption of knowing the position of more than two loudspeakers in the 3D space, in [Sachar et al., 2005] the authors presented a method based on a cost function, implementing the triangulation technique. This was possible having calculated the distances between each sensor and source from the times of arrival (TOAs) obtained producing chirp signals from every loudspeaker. The same technique to extract TOAs was used in [Raykar et al., 2005] where two different methods exploiting TOAs and time differences of arrival (TDOAs) were presented. The maximum likelihood (ML) technique was exploited to estimate microphone and speaker locations. Assuming TOAs known a-priori, microphone and source positions were estimated in [Crocco et al., 2012] for both far- and near-field cases. Receivers and transmitters were localised in [Kuang et al., 2013] applying minima solvers to a matrix containing the distances between the microphones and sources. A different approach was proposed in [Birchfield and Subramanya, 2005] where the classical multidimensional scaling (MDS) was extended into the so-called basis-point classical MDS (BCMDS). By physically measuring (through a
tape) only the distances between each microphone and a small number of basis points, the entire squared-distance matrix, containing the squares of all the interpoint distances, was constructed and decomposed to find the wanted positions. In [Chen et al., 2007] an energy based approach was presented. Assuming microphones and loudspeakers as lying on the same 2D plane, the energy of the audio segments was exploited to use the ML estimation for sensor and source positions. Making the assumption of having knowledge of the source position and the six reflectors position of a shoebox-like room, using the image sources for the first and second order reflection, the microphone position was estimated in [Parhizkar et al., 2014].

Regarding the reflector position, it can be estimated by exploiting knowledge of a single RIR [Dokmanić et al., 2011, Moore et al., 2013] or multiple channel systems [Antonacci et al., 2012, Filos et al., 2012]. In [Moore et al., 2013], the authors estimated the geometry of the room by calculating the positions of the image sources, based on TOAs and TDOAs between high-order reflections. However, TOAs of second-order reflections were necessary, and with real RIRs it is not always possible to reliably detect them. In [Tervo and Korhonen, 2010], the reflector position was estimated exploiting the inverse mapping of the acoustic multipath propagation problem in 2D. The strength was that they did not assume knowledge of RIRs, however, localisation failed at low signal-to-noise ratios (SNRs). Another way to find the reflector positions was proposed in [Canclini et al., 2011a], where the authors generated constraints from direct sound and first reflection DOAs in a 2D geometry. Exploiting image source theory, it was also possible to define the shape of a room considering the uniqueness between it and a single RIR, in case of polygonal geometries [Dokmanić et al., 2011] and L-shaped rooms [Marković et al., 2013a]. However, these algorithms were not robust to noise. The method in [Tervo and Tossavainen, 2012] iteratively searches the planes exploiting the image-source locations estimated through a maximum likelihood based algorithm. Another approach is to estimate DOAs related to all the reflections, direct sound and interference, employing a spherical harmonic domain minimum variance distortionless response (MVDR) beamformer. Then the TDOAs of the direct sound and reflections were obtained through a cross-correlation method [Sun et al., 2011]. In [Antonacci et al., 2012], the authors presented a method to estimate the position of the walls using TOAs to generate ellipses, that were tangent to them. This algorithm related distances calculated directly from RIRs with the ellipse’s property that the sum of the distances from the two foci to any point on the ellipse is a constant (see Chapter 3. However, the 2D scenario they have considered assumed that a perfectly absorbent floor and ceiling existed.

In this appendix, a new method to estimate the microphone positions is presented. It is an iterative algorithm based on a rough initial position estimate. This is then applied to a source and reflector estimation model to create a complete room geometry
estimation model. The reflector estimation method is the ellipsoid tangent sample consensus (ETSAC), whereas the source locator is similar to image source direction and ranging (ISDAR). Both ETSAC and ISDAR are proposed in Chapter 4. In Section B.2 the estimation of sensors, sources and reflector position is presented. Section B.3 shows simulations performed and results. Finally Section B.4 draws the conclusion.

B.2 Recording Setup Estimation

In Chapter 4, a method to estimate a reflector position having RIRs and microphone positions available was presented. Despite the good performance, the approximations made during each microphone position measurements were identified as cause of errors. In the following subsections, a new iterative approach to refine the microphone positions using a uniform circular array (UCA) or uniform rectangular array (URA) will be presented. Then, the source position and reflector localisation models will be reported.

B.2.1 Sensor Positions - The CISPE Algorithm

Algorithms that assume the microphone position knowledge are affected by a non-precise measurement. The cross-correlation based iterative sensor position estimation (CISPE) algorithm is here proposed. It is based on the cross-correlation between the recorded RIRs and the beamformed signals. Providing performance good enough for the purposes of this appendix and being simple, the delay-and-sum beamformer (DSB) [Van Trees, 2002] was used. The position of each microphone is updated every cycle. This procedure is applied to all the $M$ microphones used.

Pre-processing and Initialisation. Naming the RIR recorded between the $i$-th sensor and $l$-th source as $I_{i,l}(n)$, (as it was defined in Equation 2.47), where $n$ is the discrete time variable, a pre-processing step is performed over the data. Most of a RIR energy is concentrated on the direct sound, therefore, to avoid noise during the calculations, a Hamming window $H_{\text{seg}}$ of length $T_{\text{seg}} = 2.1$ ms is applied to the RIRs, to select the direct sounds only. A method for selecting the peaks of signals has been developed based on the dynamic programming projected phase slope algorithm (DYPSA) [Naylor et al., 2007]. This was initially designed to estimate glottal closure instances from speech signals, and has been modified to make it applicable to RIRs, as described in Chapter 4. At this point, segmenting the RIR using DYPSA, it is possible to obtain the signal $I_{i,l}^{S}(n)$. This signal was then filtered, using bandpass filters, obtained by observing where the most of the RIRs frequency content was centred. The subband which results to have the flatter frequency content is selected for each RIR. The filtered RIR is defined as $I_{i,l}^{SF}(n)$. 
Algorithm B.1 The CISPE algorithm

1: **procedure** *Position estimation*
2: \(I_{i,l}^S \leftarrow \text{Direct sound segmented } I_{i,l}\)
3: \(I_{i,l}^{SF}(n) \leftarrow \text{Bandpass filtering } I_{i,l}^S\)
4: \(A_{\text{init}}^l \leftarrow \text{Microphone position initialisation}\)
5: for \(i \leftarrow 1, M\) do
6: \(\text{for } l \leftarrow 1 : L\) do
7: \(I_{i,l}^B(n) \leftarrow \text{Beamformed } I_{i,l}^{SF}(n)\)
8: \(\text{while } \overline{Q_{CC}^{l,n_i,l}} \geq 0\) do
9: \(n_{i,l}^C \leftarrow -0.007 : 0.007\) do
10: \(I_{l}^{BD}(n) \leftarrow \text{Beamformed delayed } I_{i,l}^{SF}(n)\)
11: \(\overline{Q_{CC}^{l,n_i,l}} \leftarrow \text{Max of the cross-correlation using } I_{i,l}^{BD}(n)\)
12: \(I_{i,l}^B(n) \leftarrow \text{Update the beamformed signal variable with } I_{i,l}^{BD}(n)\)
13: \(E_{i,l}^C \leftarrow (c_0 \cdot \arg \max_{n_{i,l}^C} Q_{CC}^{l,n_i,l})/f_s\)
14: \(A_{\text{part}}^l \leftarrow \text{Estimated microphone position given the } l\text{-th source information}\)
15: \(A_i \leftarrow \text{Mean of the estimated microphone positions over every source}\)

As the majority of the iterative algorithms, also CISPE needed to be initialised. A rough estimation of the \(M\) microphones position was used for this aim. In the Cartesian coordinate system these positions defined as \(A_{\text{init}}^i = (A_{\text{init}}^x,i, A_{\text{init}}^y,i)\).

**Iterative Core.** The filtered RIRs \(I_{i,l}^{SF}(n)\) were used as input of the DSB. Considering each single loudspeaker, the beamformed signals \(I_{i,l}^B(n)\) were obtained, where \(l\) is the loudspeaker index. The cross-correlation between the beamformed signal and the \(M\) recorded RIRs was calculated, and the maximum values averaged over the \(M\) microphones:

\[
Q_{CC}^{l,i} = \sum_{u=0}^{T_{\text{seg}} - n - 1} I_{i,l}^{SF}(i + n)I_{i}^B(u); \quad \overline{Q_{CC}^l} = \frac{1}{M} \sum_{i=1}^{M} \max Q_{CC}^{l,i}, \quad (B.1)
\]

where \(u\) is a variable used to calculate the cross-correlation, and \(T_{\text{seg}}\) the length of both \(I_{i,l}^{SF}\) and \(I_{i,l}^B\). The second step was to apply a time sample delay \(n_{i,l}^C \in \mathbb{N}\), to one microphone at a time checking the new \(\overline{Q_{CC}^l}\). The set of allowed delays was chosen as the range associated to a spatial displacement of \(\pm 7\) mm. By defining the delayed RIR as \(I_{i,l}^{BD}(n) = I_{i,l}^{SF}(n - n_{i,l}^C)\) the new beamformed signal \(I_{i,l}^{BD}(n)\) was calculated exploiting it, and the new cross-correlation average \(\overline{Q_{CC}^{l,n_i,l}}\) was given by substituting \(I_{i,l}^{SF}(u + n)\) with \(I_{i,l}^{BD}(u + n)\) and \(I_{i,l}^B(u)\) with \(I_{i,l}^{BD}(u)\) into Equation B.1. The \(n_{i,l}^C\) value that gave the highest \(\overline{Q_{CC}^{l,n_i,l}}\) provided the error:

\[
E_{i,l}^C = \frac{c_0 \cdot \arg \max_{n_{i,l}^C} Q_{CC}^{l,n_i,l}}{f_s}, \quad (B.2)
\]
B. The CISPE Algorithm

Figure B.1: The microphone positions in a $M = 36$ microphones URA (left) and a $M = 24$ microphones UCA (right), initialised (blue) and estimated (red). Displacements are 3 times magnified, for a more comprehensive representation.

where $c_0$ is the sound speed and $f_s$ the sampling frequency. In the Cartesian coordinate system the point estimated through this error was $A_{i,l}^\text{part} = (A_{x,i,l}^\text{part}, A_{y,i,l}^\text{part})$, where:

$$A_{x,i,l}^\text{part} = E_{l,i}^C \cos(\Theta_l) + A_{x,i}^\text{init} \quad \text{and} \quad A_{y,i,l}^\text{part} = E_{l,i}^C \sin(\Theta_l) + A_{y,i}^\text{init}$$ (B.3)

and $\Theta_l$ is the azimuth DOA related to the $l$-th loudspeaker. Calculating $A_{i,l}^\text{part}$ for every $L$ source, the final estimated position of the $i$-th microphone $A_{i,l}^\text{est} = (A_{x,i}^\text{est}; A_{y,i}^\text{est})$ was given by:

$$A_{x,i} = \frac{1}{L} \sum_{l=1}^{L} A_{x,i,l}^\text{part} \quad \text{and} \quad A_{y,i} = \frac{1}{L} \sum_{l=1}^{L} A_{y,i,l}^\text{part}.$$ (B.4)

CISPE is recursive since it is repeated until the calculated $Q_l^\text{CC}$ does not increase from the previous one, and it is applied to every microphone. The pseudo-code is reported in the Algorithm B.1, whereas an example of the CISPE output is shown in Figure B.1

B.2.2 Source and Reflector Positions

Source Localisation. Assuming the microphone positions as known, the following method needs the distances from the microphones to the source, and the DOA, to estimate the source position. Distances are obtained using the TOAs of the direct sounds, extracted from RIRs using the DYPSA algorithm introduced in Section B.2.1. Since the output is a sequence of non-zero values placed on the time samples corresponding to the RIR peaks, TOAs (in samples) for direct sound and first order reflections are defined as
where \( n_{e,i} \), where \( e \) is the reflector index (i.e. \( e = 0 \) defines the direct sound). Distances from the source are then obtained \( \rho_{0,i} = n_{0,i}f sc_0 \).

To calculate DOAs for signals received by a microphone array composed by \( M \) elements, several classical methods can be adopted such as Bartlett, Capon, or the estimation of signal parameters via rotational invariance techniques (ESPRIT) [Van Trees, 2002]. The multiple signal classification (MUSIC) algorithm [Schmidt, 1986] was chosen for the present study, since it can estimate DOAs related to sources and image sources with high accuracy and stability. Beyond this, the fundamental requirement for using MUSIC, i.e. knowledge of the steering vector, is observed, since the microphone positions were estimated through the algorithm shown in Section B.2.1. The microphone array shapes enable the estimation of the azimuth DOA \( \Theta_0 \). The radial distance \( \rho_{0,i} \) is calculated by DYPSA. For this reason, given \( \rho_{0,i} \) and \( \Theta_0 \), and lying the \( i \)-th microphone on the point with coordinates given by Equation B.4, the source position coordinates are found

\[
\begin{align*}
    b_{x,0} &= A_{x,i} + \rho_{0,i}\cos(\Theta_0), \\
    b_{y,0} &= A_{y,i} + \rho_{0,i}\sin(\Theta_0).
\end{align*}
\]

This method, although utilising MUSIC instead of DSB, is similar to the image source direction and ranging (ISDAR), the algorithms proposed in Chapter 4.

**Ellipsoid Generation.** Having knowledge of both microphone and source positions, firstly, ellipsoids are generated, then the reflector was searched using a random sample consensus (RANSAC)-based algorithm [Fischler and Bolles, 1981].

The idea is to construct an ellipsoid having its major axis equal to the first order reflection path and foci on the microphone-source positions, creating an ellipsoidal set of possible points where the reflector is tangent. The general equation characterising a quadratic surface in the 3D continuous space included 10 parameters: \( h_{11}, h_{22}, h_{33}, h_{44}, h_{12}, h_{13}, h_{14}, h_{23}, h_{24} \) and \( h_{34} \). They can be placed in a \( 4 \times 4 \) symmetric matrix \( E \) to create a model in homogeneous coordinates. A unitary sphere centred on the origin of the system is defined as

\[
E_I = \begin{bmatrix} I & 0 \\ 0 & -1 \end{bmatrix},
\]

where \( I \) is a \( 3 \times 3 \) identity matrix. Transformations of translation, rotation and scaling are applied to model the ellipsoid with the required centre position, axes directions and lengths [Akenine-Moller et al., 2008]. Therefore, the matrix defining the ellipsoid relative to the \( i \)-th microphone and the \( e \)-th reflector is:

\[
E_{e,i,l} = T_{i,l}^T R_{i,l}^T S_{e,i,l}^T E_I S_{e,i,l} R_{i,l} T_{i,l}.
\] (B.5)

Considering the source position \( B_0 = (b_{x,0,l}; b_{y,0,l}; b_{z,0,l}) \) and the \( i \)-th microphone lying on the point \( A_i(A_{x,i}; A_{y,i}; A_{z,i}) \), the sphere centre position is translated to the midpoint between the two foci, through matrix \( T_{i,l} \). The scaling matrix \( S_{e,i} \) enlarges (or shrinks) the sphere to have the major axis defined as \( Q_{e,i}^{maj} = \rho_{e,i} \), whereas the two minor axes
are identical and coincide with $Q_{e,i}^{\min} \equiv \sqrt{\rho_{e,i}^2 - \rho_{0,i}^2}$. Finally, a rotation transformation is applied to each axis, and the three rotation matrices are combined as $R_i = R_{x,i}R_{y,i}R_{z,i}$.

**Reflector Search.** The required plane is the one which is tangent to every ellipsoid. A plane can be defined in homogeneous coordinates, and written as an array $p = [p_1 \ p_2 \ p_3 \ p_4]^T$, which is tangent to $E$ if it satisfies the equation $p^T E^* p = 0$, where $E^*$ is the adjoint matrix of $E$. A reflector position search method, based on RANSAC, was used. This was the method proposed in Chapter 4 under the name of ellipsoid tangent sample consensus (ETSAC). The idea was to randomly select a certain number of points on the ellipsoid indexed by $l = 1$ and $i = 1$, and verify, by setting a threshold, which subset generated the plane closest to the required one. $C$ points were randomly selected on $E_{e,1,1}$ and the normal vectors $u_p$ calculated to obtain the $p$-th plane tried during the algorithm $p_p$. To verify if the plane was tangent to all the $N = ML$ ellipsoids, where $M$ is the number of microphones and $L$ the number of sources, $|p_p^T E_{e,l,i}^* p_p| = \zeta_{e,l,p}$ was calculated for each of them. Since the plane was perfectly tangent if $\zeta_{e,l,p} = 0$, a threshold $\tau_1$ was set and, when $\zeta_{e,l,p} > \tau_1$, the ellipsoid was considered non-tangent. The plane that had the most ellipsoid support was selected.

**B.3 Experiments**

The algorithms described above have been implemented in Matlab and several experiments were performed. Exploiting measured RIRs, the performance of ETSAC were observed applying the CISPE algorithm as preprocessing. RIR measurements from three different laboratories at the University of Surrey have been used. One dataset was recorded exploiting a double concentric UCA (see Figure B.2b), whereas the other two used a URA.
B.3.1 Recording Setup

**UCA Recordings.** RIRs were recorded in a large recording studio called “Studio1” with dimensions 17.08 × 14.55 × 6.50 m³ and a RT60 of 894-945 ms, between 500 Hz and 2 kHz. \( L^{TOT} = 15 \) different loudspeaker positions were used and \( L = 3 \) were selected for the purposes of this article, named from “A” to “C”. These 3 loudspeakers were positioned at a height of 1.5 m, lying on a circle around the UCA (\( M = 24 \) microphones) with radius of 1.5 m. Defining the loudspeaker B as the one at 0°, A was positioned at −45° and C at 45°. The UCA is formed by a double concentric set of microphones with radius 8.5 cm and 10.6 cm respectively. For the aim of this appendix, the inner UCA only was employed. The sample frequency used was 48 kHz and the swept-sine technique was used to measure RIRs [Farina, 2000].

**URA Recordings.** A reproduction and measurement system was mounted on a spherical structure, the “Surrey Sound Sphere” [Coleman et al., 2014b] (see Figure B.2a). It was placed in two acoustically treated rooms. The first one is called “Studio2”, with dimensions 6.55 × 8.78 × 4.02 m³ and RT60 235 ms averaged over the 0.5 kHz, 1 kHz and 2 kHz octave bands. The second room is called “Vislab”, with dimensions 7.90 × 6.00 × 3.98 m³, and RT60 of 326 ms averaged as for “Studio2”. \( L^{TOT} = 60 \) loudspeakers (Genelec 8020b) were clamped to the equator to form a circular array (radius of 1.68 m). \( M = 48 \) microphones (Countryman B3 omni) were attached to a grid mounted on a microphone stand. The height of the equator and the microphones, is 1.62 m. The sample frequency used was \( f_s = 48 \) kHz. For this article, 8 sources lying on the equator with azimuth 0°, 90°, 120°, 150°, 180°, 270°, 300° and 330°, and \( M = 36 \) microphones having a 6 × 6 squared configuration with an inter-element spacing of 5 cm, were used. Considering the centre of the sphere as the origin if a coordinate system, the central microphone of the URA is placed in (0.0; 0.0; 1.62) m for “Studio2”, whereas for the “Vislab” dataset in (0.675; 0.000; 1.620) m.

B.3.2 Reflector Estimation

To test the improvements introduced by CISPE algorithm, the root mean square error (RMSE), generated by ETSAC, was calculated considering the z-axis value at \( x^{ev} = 5 \) points, lying on the estimated plane, equally spaced between the sources and microphones. From these values, the expected ones were subtracted to obtain the errors \( \epsilon_{i,l,q}^{ref} \).

---

1 Available at: http://cvssp.org/data/s3a/
2 Available at: http://cvssp.org/soundzone/resource/
where \( q \) is the index related to the analysed points. Hence, considering \( N \) ellipsoids:

\[
\text{RMSE} = \sqrt{\frac{1}{x^{\text{ev}} N} \sum_{i=1}^{M} \sum_{l=1}^{L} \sum_{q=1}^{x^{\text{ev}}} \epsilon_{i,l,q}^2}.
\]  

(B.6)

The model was tested using different numbers of microphones, \( M \in \{5, 7, 9, 16, 25, 36\} \) for the URA and \( M \in \{7, 13, 19, 24\} \) for the UCA. 100 combinations of \( L = 3 \) loudspeakers (randomly taken over the 8 selected for “Studio2” and “Vislab” and 3 selected for “Studio1” were used for each different number of microphones. Given that CISPE is working, for now, with 2D space, three sources lying on the same plane of the UCA were selected.

The RMSE for each number of microphones used, averaged over 100 trials, is reported in Figure B.3. These results show the improvement given by the introduction of CISPE in every dataset. Although the reflector localisation performance improves, with a low number of microphones (in general, between 5 and 9) the improvement given by CISPE is lower than for a higher number of microphones. CISPE’s aim is to reduce the errors that are given by microphone misplacements. Whenever few microphones are used, all the other sources of error (e.g. SNR level affecting TOA estimation, etc) are more effective, since less information can be used to attenuate the single channel errors. This is the cause limiting the improvement given by CISPE. On the other hand, by using too many microphones, a too wide URA may be placed inside the Surrey Sound Sphere. Therefore, the far field assumption is not respected any more. Following the Fraunhofer rule [Balanis, 2005], with the sphere radius 1.68 m, using 36 microphones, the signals are considered in the near field for every frequency over 3 kHz. For 25 microphones the break frequency increases to around 10 kHz. For this reasons, it is possible to state that the maximum number of microphones selectable to have a URA inside the sphere is 25.
CISPE, a new algorithm to estimate the microphone positions in a URA or UCA, has been presented, together with an already available source and reflector position estimator. Experiments for real RIRs, recorded using the two different microphone array configurations, were performed observing the reflector estimation model. RMSEs showed improvements on the model with the introduction of CISPE as preprocessing. Future work will investigate the behaviour of the CISPE-ETSAC combination selecting different subsets of microphones in the double UCA.
Bibliography


Bibliography


Bibliography


Dokmanić, I., Daudet, L., and Vetterli, M. (Shanghai, China, 2016). From acoustic room reconstruction to SLAM. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.


ECMA-404 (2013). The JSON data interchange format.


Bibliography


Bibliography


Bibliography


Bibliography

Proc. of the IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP).


