UNIVERSITY OF CRETE
DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF SCIENCES AND ENGINEERING

DRACOSS: a framework for direction of arrival estimation and counting of multiple sound sources with microphone arrays

by

Despoina Pavlidi

PhD Dissertation

Presented

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

Heraklion, February 2018

UNIVERSITY OF CRETE

DEPARTMENT OF COMPUTER SCIENCE

DRACOSS: a framework for direction of arrival estimation and counting of multiple sound sources with microphone arrays

PhD Dissertation Presented

by **Despoina Pavlidi**

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

APPROVED BY:

Author: Despoina Pavlidi

Supervisor: Athanasios Mouchtaris, Associate Professor, University of Crete

---

Committee Member: Panagiotis Tsakalides, Professor, University of Crete

---

Committee Member: Ville Pulkki, Associate Professor, Aalto University

---

Committee Member: Boaz Rafaely, Professor, Ben-Gurion University of the Negev

---

Committee Member: Yiannis Stylianou, Professor, University of Crete

---

Committee Member: Petros Maragos, Professor, National Technical University of Athens

---

Committee Member: Gerasimos Potamianos, Associate Professor, University of Thessaly

---

Department Chairman: Angelos Bilas, Professor, University of Crete

Heraklion, February 2018

This dissertation is dedicated to the memory of Ioannis Emm. Pountourakis, Professor, NTUA.

For believing in me when I did not believe in myself, for building and supporting my “research scientific character", for teaching me and sharing beautiful life values in unexpected spontaneous conversations, for still inspiring me.

Κι ένα σκαλοπάτι φτάνει να φανεί κι άλλο μπροστά

Της Γνώσης σκαλοπάτια όσο ανεβαίνεις
κι άλλο ν΄ ανεβαίνεις θέλεις.
Τα σκαλοπάτια αρχίζουν της ζωής
από το πρώτο "κλάμα".
Είναι το άλφα, κεφαλαίο, για τη Γνώση·
και το ωμέγα, και τούτο κεφαλαίο,
εσύ και μόνον εσύ
θα βρεις να συναντήσεις.
Κι ακόμη, αν το συναντήσεις,
να ξέρεις υπάρχουν κι άλλα σκαλοπάτια
και μη σε νοιάζει αν δεν τα πατήσεις.
Είναι ωραίο π΄ ανέβηκες το πρώτο
και κάμποσα ακόμη.
Και όπως περιγράφει ο Καβάφης:
ο Θεόκριτος λέει στον ποιητή Ευμένη
που μόνο στο πρώτο της ποίησης το σκαλοπάτι είχε ανέβει
"Εδώ που έφτασες λίγο δεν είναι,
τόσο που έκαμες, μεγάλη δόξα."

Αποστόλης Παυλίδης
"στο τριανταφυλλένιο μου τριαντάφυλλο"
24/02/2018

Acknowledgments

I would like to thank my supervisor, Associate Professor Athanasios Mouchtaris, for giving me the chance, the training and the freedom to evolve, question, and create under his guidance. A grateful “thank you" to my advisory committee members, Professor Panagiotis Tsakalides and Associate Professor Ville Pulkki for the inspiration and support. It was a great honor to have as examination committee members, Professor Boaz Rafaely, Professor Yiannis Stylianou, Professor Petros Maragos and Associate Professor Gerasimos Potamianos.

During the years of my PhD studies I met and collaborated with very good researchers and interesting people, two of which I feel the need to specially thank. Dr. Anthony Griffin coached me during my MSc studies and the first two years of my PhD journey. He is the greatest mentor I have ever met, passionate about his work, inspirational, patient and kind to his students. Anthony, thank you from the deepest parts of my heart for the support, help and inspiration throughout the years of our collaboration, but also for later, you were always there... When I first met Dr. Symeon Delikaris-Manias we were both at the beginning of our PhD studies. I was lucky enough to collaborate with him for the biggest part of my PhD. His motivation, his perspective and his passion dragged me out of my plateau and inspired me to continue and complete my studies. Apart from a great colleague, Symeon is now also a great friend to share the passion for Crete and its magnificent landscape.

As a scholar of FORTH-ICS and student of CSD-UOC, I cannot but thank the greatest secretaries each institute would be grateful to have, Ms. Maria Prevelianaki, Ms. Gelly Kosma and Ms. Rena Kalaitzaki. Ms. Prevelianaki, secretary of FORTH-ICS, was a savior whenever I had to arrange a conference trip, or any other administration matter. I made her life difficult quite a few times, but she would always come up with a solution. Ms. Gelly Kosma, postgraduate studies secretary of CSD-UOC, is always available and of service to any student in need, and so was to me. I gratefully thank her for sorting problems out, for answering anxious questions, for helping with all the paper work. What can I say for Ms. Rena Kalaitzaki, “mama Rena", the head of secretariat of CSD-UOC....Ms. Kalaitzaki was officially the first person I met when I joined the department as a MSc student and I immediately felt her warmhearted welcome that would accompany me throughout the years of my studies. I am so grateful for meeting this great woman...the “heart of CSD"! I cannot but thank the administrators of both hosting institutes. A special thanks to Mr. Vaggelis Karayiannis of FORTH-ICS and Mr. Yiannis Surlatzis of CSD-UOC for their help.

During the scholarship years at FORTH-ICS I was a member of the TNL and SPL labs. Being a member of such active and buzzing research groups gave a significant boost to my research activity. Special thanks to colleagues Yiannis, Tasos, Nikos, Manolis, Grigoris, and Maria. Collaborating and interacting with such interesting people made the lab life far more interesting!

My friends had always a central role in my life, which became even more significant through the PhD studies. Charoula, Nandia, Nansu, Katerina, Petros, Maria, Liz, Irini, Panagiotis, your support, your available shoulder to cry to, your eager ears to listen to my problems, your open arms to hug my worries meant the world to me! Thank you for being there whenever I needed you!

Being surrounded by such a wonderful family made me who I am, and always gave the strength to pursue more! Pavlos, Manolis, Alexandra, Konstantinos, baby Apostolis and young Manolis, Toma, Lina, thank you for being the family everybody would be jealous of! I love you with all the strength of my heart! Thanking my parents, Apostolis and Koula, always brings tears to my eyes, and this time is not an exception...I couldn' t do any of the things I have achieved in my life if it wasn't for them...Mom, dad, I love you and need you so much...

Stefanos, I wouldn' t have made it without your caring love... I am so, so lucky to meet you and share my life with you! “Thank you" sounds so poor and little to describe everything I owe to you....

Abstract

Technological advances have infiltrated our everyday life more than ever before. High intelligence devices and gadgets, equipped with cutting-edge technology algorithms, facilitate and empower our lifestyle. Smart-home automation, next generation hearing aids, robots with autonomous navigation systems have brought to the foreground of the research community audio signal processing problems. One such problem is the estimation of the number of sources and the directions from which sound originates, what we most frequently call *direction of arrival (DOA)* estimation.

The problem of DOA estimation is active for more than thirty years, consequently a plethora of algorithms have been proposed in the literature. Some of them can be considered classic and frequently come from the telecommunications research area. Beamforming techniques belong in this category, where an appropriately weighted sum of the signals of a microphone array is used to form a receiving beam, which scans the space and detects areas of activity. Subspace approaches, such as the well-known MUSIC algorithm, formulate a spatial function that gets maximized when activity is detected, relying on the decomposition of the array sample covariance matrix. Other algorithms stemmed from research activity on blindly separating mixtures of audio signals, i.e., the blind source separation (BSS) problem. Independent component analysis methods, where the goal is to estimate a demixing matrix, which reveals DOA information, and sparse component analysis methods, which exploit the sparsity of activity of the sources in some appropriately chosen domain, both fall into the BSS category. A recently emerging category is that of estimating the intensity vector, which points towards the net flow of sound energy, hence, revealing the corresponding DOA of the generating sound source.

The aforementioned methods fail at either estimating accurately DOAs when multiple sources are simultaneously active, e.g., beamforming techniques, or they are computationally heavy and significantly affected by the amount of available data, e.g., ICA and subspace approaches, while some are restricted by specific array geometries. We, thus, observe the lack of a methodology than can address the problem of DOA estimation holistically, aiming at tackling all aforementioned aspects of the problem.

In this thesis we aim at filling this gap with our proposed DRACOSS framework, i.e., an integrated framework for tackling the problem of DOA estimation and counting of multiple, simultaneously active, sound sources utilizing microphone arrays. DRACOSS is developed in two-dimensional (2D) and three-dimensional (3D) spaces, using a uniform circular array and a spherical microphone array respectively. DRACOSS constitutes a proce-

dure of four distinct steps: (a) exploitation of the sparsity of sound signals, (b) local single-source DOA estimation, (c) histogram formation, and (d) post-processing of the histogram. We detect the sparsity of involved sound signals in the time-frequency domain by utilizing a relaxed sparsity assumption, which relies on the estimation of a mean correlation coefficient between pairs of microphones. We proceed with the collection of local DOA estimates in detected single-activity areas, which will then be used to form histograms. For the 2D case we employ a local DOA estimator, designed specifically for circular arrays and form one-dimensional histograms. For the 3D case we use an intensity vector estimator and then form two-dimensional histograms. In both cases, by post-processing the histograms we provide counting and DOA estimation results for all active sound sources. DRACOSS performs robustly under a wide collection of simulated and real scenarios in terms of noise and reverberation conditions, in terms of the number of simultaneously active sources and in comparison with state-of-the-art methods. We also propose the formulation of two classic DOA methods, i.e., beamforming and MUSIC, through the DRACOSS framework, which manages to significantly improve their performance. Aiming at constantly improving our approach and following the vivid technological stream, we show recent, very promising results on counting by utilizing deep neural networks.

Keywords: direction of arrival, counting, microphone arrays, time-frequency domain, spherical microphone arrays, spherical harmonic domain, sound intensity vector, histogram processing, sparsity

Supervisor: Athanasios Mouchtaris
Associate Professor
Computer Science Department
University of Crete

Περίληψη

Οι τεχνολογικές εξελίξεις έχουν διεισδύσει στην καθημερινότητά μας όπως ποτέ άλλοτε. Συσκευές υψηλής νοημοσύνης, εξοπλισμένες με αλγορίθμους τελευταίας τεχνολογίας, διευκολύνουν και ενισχύουν τον τρόπο ζωής μας. Συστήματα αυτοματισμού για έξυπνα σπίτια, ακουστικά βαρηκοΐας επόμενης γενιάς, ρομπότ με αυτόνομα συστήματα πλοήγησης φέρνουν στο προσκήνιο της επιστημονικής κοινότητας προβλήματα επεξεργασίας σημάτων ήχου. Ένα από αυτά τα προβλήματα είναι η εκτίμηση του πλήθους και των κατευθύνσεων από τις οποίες προέρχεται ο ήχος, αυτό που συνήθως ονομάζουμε εκτίμηση κατεύθυνσης άφιξης.

Το πρόβλημα της εκτίμησης κατεύθυνσης άφιξης είναι ενεργό για πάνω από τριάντα χρόνια, συνεπώς πληθώρα αλγορίθμων έχει προταθεί στη σχετική βιβλιογραφία. Μερικοί από αυτούς τους αλγορίθμους προέρχονται από την περιοχή των τηλεπικοινωνιών. Σε αυτήν την κατηγορία ανήκουν οι τεχνικές διαμόρφωσης δέσμης, στις οποίες χρησιμοποιώντας κατάλληλα βάρη, ένα άθροισμα των σημάτων μίας συστοιχίας μικροφώνων σχηματίζει μία δέσμη δέκτη που σαρώνει το χώρο και ανιχνεύει περιοχές ακουστικής δραστηριότητας. Οι προσεγγίσεις υποχώρων, όπως ο διάσημος αλγόριθμος MUSIC, διαμορφώνουν μία χωρική συνάρτηση η οποία μεγιστοποιείται στα σημεία που ανιχνεύεται δραστηριότητα χρησιμοποιώντας την αποσύνθεση της δειγματικής μήτρας ετεροδιακύμανσης της συστοιχίας. Άλλοι αλγόριθμοι έχουν αναδυθεί μέσα από τις προσπάθειες διαχωρισμού μειγμάτων ηχητικών σημάτων. Σε αυτήν την κατηγορία ανήκουν οι μέθοδοι που στηρίζονται στην ανάλυση ανεξάρτητων στοιχείων, στοχεύοντας στην εκτίμηση μίας μήτρας διαχωρισμού που εμπεριέχει πληροφορία για την κατεύθυνση άφιξης, καθώς και οι μέθοδοι ανάλυσης αραιών στοιχείων που εκμεταλλεύονται την αραιότητα της δραστηριότητας των πηγών σε κάποιον κατάλληλα επιλεγμένο χώρο. Μία προσέγγιση που έχει τραβήξει πρόσφατα το ενδιαφέρον είναι αυτή της εκτίμησης του διανύσματος έντασης του ηχητικού πεδίου, το οποίο έχει κατεύθυνση προς την καθαρή ροή ηχητικής ενέργειας, συνεπώς μπορεί να παρέχει την κατεύθυνση άφιξης της γεννήτριας ηχητικής πηγής.

Οι προαναφερθείσες μεθοδολογίες αποτυγχάνουν είτε στην ακριβή εκτίμηση της κατεύθυνσης άφιξης όταν πολλαπλές πήγες είναι ταυτόχρονα ενεργές, όπως οι τεχνικές διαμόρφωσης δέσμης, είτε ενέχουν υψηλό υπολογιστικό κόστος και εξαρτώνται σημαντικά από τον όγκο των διαθέσιμων δεδομένων, όπως οι τεχνικές ανεξαρτήτων στοιχείων και υποχώρων, ενώ κάποιοι αλγόριθμοι απευθύνονται σε συγκεκριμένες τοπολογίες συστοιχιών μικροφώνων. Διαφαίνεται, συνεπώς, η έλλειψη κάποιας μεθοδολογίας που να αντιμετωπίζει το πρόβλημα της εκτίμησης κατεύθυνσης άφιξης ολιστικά και να μπορεί να ανταπεξέλθει σε όλες τις διαφορετικές πτυχές του προβλήματος.

Σε αυτήν τη διατριβή αποσκοπούμε να καλύψουμε αυτό το κενό και προτείνουμε ένα ολοκληρωμένο πλαίσιο για την επίλυση του προβλήματος εκτίμησης πλήθους και κατεύθυνσης άφιξης

πολλαπλών, ταυτόχρονα ενεργών πηγών με τη χρήση συστοιχιών μικροφώνων. Το πλαίσιο, το οποίο ονομάζουμε εφεξής DRACOSS, αναπτύσσεται σε διδιάστατους και τριδιάστατους χώρους, χρησιμοποιώντας μία ομοιόμορφη, κυκλική συστοιχία και μία σφαιρική συστοιχία μικροφώνων αντίστοιχα. Το DRACOSS αποτελεί ουσιαστικά μία διαδικασία τεσσάρων ευκρινών βημάτων: (α) εκμετάλλευση της αραιότητας των ηχητικών σημάτων, (β) τοπική εκτίμηση κατεύθυνσης άφιξης μίας πηγής, (γ) σχηματισμός ιστογράμματος, και (δ) επεξεργασία του ιστογράμματος. Ανιχνεύουμε την αραιότητα των ηχητικών σημάτων στο πεδίο των χρονο-συχνοτήτων χρησιμοποιώντας μία χαλαρή υπόθεση αραιότητας που στηρίζεται στην εκτίμηση ενός συντελεστή μέσης συσχέτισης μεταξύ σημάτων ζευγών μικροφώνων. Σε επόμενο βήμα συλλέγουμε τοπικές εκτιμήσεις κατεύθυνσης άφιξης από όλες τις περιοχές μοναδιαίας δραστηριότητας, τις οποίες και χρησιμοποιούμε για να σχηματίσουμε ιστογράμματα. Σε διδιάστατους χώρους, ως τοπικό εκτιμητή κατεύθυνσης χρησιμοποιούμε έναν αλγόριθμο ειδικά σχεδιασμένο για κυκλικές συστοιχίες και σχηματίζουμε μονοδιάστατα ιστογράμματα, ενώ για τους τριδιάστατους χώρους χρησιμοποιούμε εκτιμήσεις του διανύσματος ηχητικής έντασης και σχηματίζουμε διδιάστατα ιστογράμματα. Και στις δύο περιπτώσεις με περαιτέρω επεξεργασία των ιστογραμμάτων παρέχουμε εκτιμήσεις του πλήθους και των κατευθύνσεων άφιξης όλων των ενεργών ηχητικών πηγών. Το DRACOSS παρουσιάζει εύρωστη απόδοση τόσο σε προσομοιωμένα, όσο και σε πραγματικά σενάρια, για διάφορες συνθήκες θορύβου και ανακλάσεων και για διάφορα πλήθη εμπλεκόμενων πηγών. Επίσης, το προτεινόμενο πλαίσιο υπερέχει πολλών, γενικώς αναγνωρισμένων μεθόδων της βιβλιογραφίας. Επιπροσθέτως προτείνουμε την ανάπτυξη δύο κλασικών μεθόδων εκτίμησης άφιξης, της μεθόδου σχηματισμού δέσμης και του αλγορίθμου MUSIC, υπό το προτεινόμενο πλαίσιο DRACOSS, βελτιώνοντας έτσι σημαντικά την απόδοσή τους. Αποσκοπώντας στη συνεχή βελτίωση της προσέγγισής μας, ακολουθώντας, δε, τις τελευταίες τεχνολογικές τάσεις, παρουσιάζουμε πρόσφατα και πολλά υποσχόμενα αποτελέσματα αναφορικά με την εκτίμηση πλήθους ενεργών πηγών, χρησιμοποιώντας βαθιά νευρωνικά δίκτυα.

Λέξεις κλειδιά: κατεύθυνση άφιξης, εκτίμηση πλήθους, συστοιχίες μικροφώνων, χώρος χρονο-συχνοτήτων, σφαιρικές συστοιχίες μικροφώνων, χώρος σφαιρικών αρμονικών συνιστωσών, διάνυσμα ηχητικής έντασης, επεξεργασία ιστογραμμάτων, αραιότητα

Επόπτης: Αθανάσιος Μουχτάρης
Αναπληρωτής Καθηγητής
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Contents

List of Figures

List of Tables

Acronyms

2D two-dimensional 1, 4, 23, 32, 49, 53, 56, 61, 63, 64, 66, 68–71, 75–77, 81, 83, 84, 100–102

3D three-dimensional 1, 5, 16, 23, 49, 54, 55, 59, 83, 84, 101, 102, 107

AIC Akaike information criterion 18

AIV active intensity vector 81

BSS blind source separation 2, 13, 14, 17

CICS circular integrated cross spectrum 28, 81

CNN convolutional neural network 24, 71, 72, 84

DirAC directional audio coding 15, 81

DOA direction of arrival 1–4, 9–18, 21–26, 28, 31, 32, 34, 36, 38, 41, 44–47, 49, 53, 56–63, 65, 66, 68–70, 75–77, 79–84, 107

DPD direct path dominance 13, 16, 17, 22, 23, 61, 67–69, 84

DRACOSS direction of arrival and counting of sound sources 4, 5, 21–27, 29, 32, 34, 36–38, 40–45, 49, 54, 59–65, 67, 68, 72, 75–79, 81, 83, 84, 100, 107

ESPRIT estimation of signal parameters via rotational invariance techniques 9, 11, 13, 14

GCC-PHAT generalized cross-correlation phase transform 2, 10

GSCT generalized state coherence transform 14, 15, 37–42

GUI graphical user interface 77

ICA independent component analysis 2, 10, 14, 21, 37–42

ImmACS immersive audio communication system 77, 79, 84

IV intensity vector 61, 62, 67

JADE joint approximate diagonalization of eigenmatrices 14, 39

LPC linear predictive coding 18, 43, 44, 122, 123

LSTM long short-term memory 72

MAEE mean absolute estimation error 34, 37, 38, 45, 81, 82

MASS minimum angular source separation 76, 77

MCC mean correlation coefficient 22, 23, 37, 49, 51, 57, 58, 84

MDL minimum description length 18, 43, 44

MEE Mean estimation error 57–59, 62, 63, 65–67, 69, 70

MEMS Microelectromechanical systems 77, 84

MUSIC multiple signal classification 2, 9, 11–13, 16, 23, 37, 40–42, 49, 61, 67–69, 83, 84

NMP networked music performance 79

NN neural network 71–73

PCSF perpendicular cross-spectra fusion 81, 82

PIV pseudointensity vector 16, 17

PS peak search 18, 43, 44

RIR room impulse response 63, 66, 69

RR recursively regularized 14, 39, 40

SCA sparse component analysis 2, 3, 10, 11, 17, 18, 21

SCAIV spatially constrained active intensity vector 81

SCB spatially constrained beamforming 59, 61, 62, 67

SFT spherical Fourier transform 108, 109, 112

SHD spherical harmonic domain 9, 12, 13, 15, 23, 50

SHF spherical harmonic function 109

SHT spherical harmonic transform 107, 109

SIR signal to interference ratio 76, 77

SNR signal to noise ratio 25, 29, 31, 33, 38–40, 44, 57, 58, 61, 62, 66, 67, 69, 75, 76

SRP steered response power 62–64

SSZ single source zone 17, 22, 23, 25, 27, 28, 33, 49, 51, 53, 57, 58, 60, 104

std standard deviation 53, 54, 68

STFT short-time Fourier transform 10, 22, 44

SVD singular value decomposition 13, 16

TDOA time difference of arrival 2, 9, 14

TF time-frequency 13, 16, 17, 22, 23, 25–27, 49, 51–53, 57, 58, 63, 68, 81

UCA uniform circular array 1, 4, 75–77, 81

WASN wireless acoustic sensor network 2, 75

WDO W-disjoint orthogonality 3, 13, 17, 23, 27, 37, 40–42, 57, 63, 84

Chapter 1
Introduction

Approaching the end of the second decade of the 21st century, we all have noticed tremendous changes in our everyday communications. As humans, we still primarily communicate using our senses and our ability to speak, yet we more and more interact with smart devices that assist our everyday living. These devices are able to sense the surrounding acoustic scene and to extract important acoustic features to enable their further acting. Of the fundamental acoustic features is the direction of arrival (DOA) of the active sound sources creating the acoustic scene.

1.1 General Objective

Direction of arrival estimation of audio sources is a natural area of research for array signal processing, and one that has had a lot of interest over recent decades [63]. The scope of this thesis is to present a complete framework for the estimation of the direction of arrival when multiple sound sources are simultaneously active. It provides also solutions for the estimation of the number of active sources in cases where this information is not known. The presented framework was originally developed for two-dimensional (2D) spaces, i.e., covering cases where sound sources exist on the same plane. It was later expanded to the three-dimensional (3D) space. The devices used were a uniform circular array (UCA) for the 2D case and a spherical microphone array for the 3D space, however the generic nature of the proposed framework does not restrict the use of other topologies, e.g., linear or planar arrays.

1.2 Motivation and Vision

Accurate estimation of the DOA of an audio source is a key element in many applications. One of the oldest and most common is in teleconferencing, where the knowledge of the location of a speaker can be used to steer a camera, or to enhance the capture of the desired source with beamforming, thus avoiding the need for lapel microphones. Nowadays the information of the DOA of the sound sources gets even more important as smart home au-

tomation becomes more prominent and smart devices equipped with microphone arrays invade our homes and everyday lives. Other audio signal processing applications that use the information of the DOA are those that deal with speech enhancement and separation and those related to wireless acoustic sensor networks (WASN) for exact location estimation of the sound sources. Other applications include event detection and tracking, robot movement in an unknown environment, high quality audio scenes recordings and next generation hearing aids [7, 10, 83, 95, 117].

Apart from the importance of the DOA information on the aforementioned application areas, the problem of DOA estimation is of great interest. It gets even more challenging when multiple sources are active, in adverse environments and using devices of common size, average cost and capabilities.

1.3 Research Questions

The focus in the early years of research in the field of DOA estimation was mainly on scenarios where a single audio source was active. Most of the proposed methods were based on the time difference of arrival (TDOA) at different microphone pairs, with the generalized cross-correlation phase transform (GCC-PHAT) being the most popular [62]. Improvements to the TDOA estimation problem—where both the multipath and the information among multiple microphone pairs were taken into account—were proposed in [12]. An overview of TDOA estimation techniques can be found in [19].

Localizing multiple, simultaneously active sources is a more difficult problem. Indeed, even the smallest overlap of sources—caused by a brief interjection, for example—can disrupt the localization of the original source. A system that is designed to handle the localization of multiple sources sees the interjection as another source that can be simultaneously captured or rejected as desired. An extension to the GCC-PHAT algorithm was proposed in [9] that considers the second peak as an indicator of the DOA of a possible second source. One of the first methods capable of estimating DOAs of multiple sources is the well-known MUSIC algorithm and its wideband variations [7, 11, 34, 54, 104, 123]. MUSIC belongs to the classic family of subspace approaches, which depend on the eigendecomposition of the covariance matrix of the observation vectors.

Derived as a solution to the blind source separation (BSS) problem, independent component analysis (ICA) methods achieve source separation, enabling in parallel multiple source localization, by minimizing some dependency measure between the estimated source signals [69, 72, 103]. The work of [86] proposed performing ICA in regions of the time-frequency representation of the observation signals under the assumption that the number of dominant sources did not exceed the number of microphones in each time-frequency region. This last approach is similar in philosophy to sparse component analysis (SCA) methods [22, ch. 10]. These methods assume that one source is dominant over

the others in some time-frequency windows or "zones". Using this assumption, the multiple source propagation estimation problem may be rewritten as a single-source one in these windows or zones, and the above methods estimate a mixing/propagation matrix, and then try to recover the sources. By estimating this mixing matrix and knowing the geometry of the microphone array, we may localize the sources, as proposed in [16, 92, 112], for example. Most of the SCA approaches require the sources to be W-disjoint orthogonal (WDO) [122]—meaning that in each time-frequency component, at most one source is active—which is approximately satisfied by speech in anechoic environments, but not in reverberant conditions. On the contrary, other methods assume that the sources may overlap in the time-frequency domain, except in some tiny "time-frequency analysis zones" where only one of them is active (e.g., [22, p. 395], [94]). Unfortunately, most of the SCA methods and their DOA extensions are computationally intensive and therefore off-line methods (e.g., [16] and the references within).

Other than accurate and efficient DOA estimation, an important issue in sound source localization is estimating the number of active sources at each time instant, known as source counting. Many methods in the literature propose estimating the intrinsic dimension of the recorded data, i.e., for an acoustic problem, they perform source counting at each time instant. Most of them are based on information theoretic criteria (see [38] and the references within). In other methods, the estimation of the number of sources is derived from a large set of DOA estimates that need to be clustered. In classification, some approaches to estimating both the clusters and their number have been proposed (e.g. [48]), while several solutions specially dedicated to DOAs have been tackled in [22, p. 388], [70] and [5].

Therefore the research questions that this thesis aims at tackling are: the problem when multiple audio sources are simultaneously active and the problem of overlapping of the audio sources at the time-frequency domain. We aim to investigate how the number of sources affects the performance of the DOA estimation as well as how the environmental conditions, i.e., noise and reverberation deteriorate the accuracy of the estimates. We also aim to study the computational load of the proposed framework.

We propose and envision a framework to solve the problem of DOA estimation and counting of multiple sound sources in a complete fashion. Our framework embraces already proposed techniques and proposes novel ones which contribute significantly to the problem under investigation.

1.4 The Approach

Our proposed framework is based on the following steps:

- detecting areas in the time-frequency domain where one source is dominantly active over the others.

- single-source DOA estimation algorithms can be applied over these zones, providing local DOA estimates
- collecting these DOA estimations into a histogram to enable the localization of multiple sources
- effective post-processing of histograms of local DOA estimates for accurate final DOA estimation of multiple sources and counting.

1.5 Contributions of the Dissertation

The main contributions of the thesis are:

- the presentation of a generic framework for DOA estimation which is not restricted by a specific array topology. On the contrary, the framework, called from now on DRACOSS, could be used with any microphone array and has been so far applied with UCAs and spherical microphone arrays.
- accurate DOA estimation when multiple sources are active. Our proposed framework does not pose any inherent restriction on the number of simultaneously active sources.
- robust performance under adverse conditions. DRACOSS manages to achieve accurate DOA estimation under high reverberation and noise, even though the signal model adopted by the framework is simple and does not take into account the effects of reverberation or noise.
- low computational load and real-time performance.
- the proposed framework shows high modularity and scalability. Since the core building blocks of DRACOSS are independent with each other, one could easily modify and extend the framework incorporating advanced techniques and algorithms at each of the fundamental blocks. In this context, an important enhancement to known state-of-the-art DOA estimation algorithms was proposed under the DRACOSS framework.

To the best of the author's knowledge, DRACOSS is the first framework that addresses the problem of DOA estimation and counting in such a holistic way. DRACOSS collects techniques and algorithms, previously proposed to solve other signal processing problems and combines the benefits of such techniques, providing superior DOA estimation accuracy compared to state-of-the art methods. Specifically:

1. DRACOSS utilizes efficiently, in the problem of DOA estimation, a sound source sparsity criterion previously proposed and used for blind source separation [94], namely the mean correlation coefficient (MCC). The efficiency of the criterion in DOA estimation was compared against the more simplistic W-disjoint assumption. The results of this comparison were shown in App. A, publication no.1 and App. A publication no.6. It is important to note here that DRACOSS related publications, such as A, publication no.1, were of the first that paid such attention to the selection of single source areas of activity for improving the DOA estimation accuracy of multiple sources and this is a great novelty element of the framework.

2. DRACOSS for 2D spaces is one of the first frameworks that achieves *real-time* multiple sources DOA estimation, also shown in App. A, publication no.1. The framework has been demonstrated to operate in real time in an international conference demo session and was shown to perform robustly with analog and digital devices in App. A, publication no.8.

3. The novel use of the MCC for DOA estimation, apart from making the proposed DOA estimation approach more accurate, provided also the advantage of reduced computational complexity as it was also shown in App. A, publication no.1. DRACOSS is more accurate and computationally efficient compared also against the classical MUSIC algorithm and compared against an ICA-based algorithm.

4. Taking advantage of sound sources sparsity, DRACOSS manages to efficiently use single source DOA algorithms for multiple sources DOA estimation. In App. A, publication no.1 we have used a single source DOA algorithm designed for circular arrays and in App. A, publication no.6 we have used a single source sound intensity-based one. In both cases, these algorithms could not have been used for multiple sources DOA estimation unless incorporated into a framework such as DRACOSS.

5. Even though histograms have been previously used for DOA estimation, DRACOSS introduces practical, intuitive, simple, yet efficient post-processing of single dimension and 2D histograms. Avoiding to use algorithms which assume a priori knowledge of the number of simultaneously sources, we achieve simultaneous accurate DOA estimation as shown in App. A, publication no.1, 6, 7, 9. In App. A, publication no.1 we also show simultaneous and accurate counting of the number of sources.

6. In App. A, publication no.7 we introduced a novel hybrid DOA estimation solution in areas of single source activity by efficiently combining the benefits of sound intensity estimation and beamforming. In the same publication our approach was compared against the MUSIC algorithm and was shown to be superior in terms of accuracy.

7. By introducing the idea of local DOA estimates and the formation of histograms into the state-of-the art methods, beamforming and the MUSIC algorithm we managed to significantly improve their performance as it was shown in App. A, publication no.9.

8. In DRACOSS we also present the novel idea of utilizing our histograms and the information they bear in order to train neural networks which accurately estimate the number of sources.

9. DRACOSS and its ideas were efficiently used in numerous other signal processing problems as it was shown in App. A, publications no. 2, 3, 4, 5, 10, 11, 12.

1.6 Outline of the Dissertation

The rest of this dissertation is organized in the following way: In Chapter 2 we present the state-of-the art of DOA estimation methods with microphone arrays. Part II is the main part of the dissertation where the DRACOSS framework is presented. In Chapter 4 we present DRACOSS in the 2D space and in Chapter 5 we show how DRACOSS was deployed to the 3D space. Finally, in Chapter 6 we present applications and uses of the proposed framework to other neighboring problems. This dissertation is concluded in Chapter 7. Basic signal processing theory applied in the DRACOSS framework as well as theorems that govern its functionality can be found in Part III.

Part I

State of the Art

Chapter 2
Direction of arrival estimation: an overview of state-of-the-art methods

The sound source direction of arrival estimation using microphone arrays has interested many signal processing researchers for more than 40 years. Thus, it is natural to have a vast collection of DOA estimation algorithms. Some of them can be considered classic and frequently come from the telecommunications research area (e.g., beamforming) and some others were developed later in time and were more dedicated to the broadband nature of the audio signals (e.g., sparse component analysis-based approaches), hence they can be characterized as modern DOA algorithms.

An attempt to categorize the large collection of DOA estimation algorithms could be as such:

- beamforming techniques: DOA estimation using beamforming is one of the first approaches. The basic concept in all proposed algorithms lies on the “scanning" of the space of interest with the beamformer of the engineer’s choice. The output power of the beamformer is estimated for every steered direction and the one that maximizes the power is considered as the DOA of the source. Several beamformers have been used in the literature, from the well known Capon’s beamformer (a.k.a. MVDR) [17], to more advanced ones like the superdirective beamformer proposed by Cox et al [24] or beamformers formulated in the spherical harmonic domain (SHD) [97] which facilitate the manipulation of spherical topologies.

- subspace approaches: Subspace approaches can be categorized along with beamforming techniques as spectral-based [63], since in both cases a spectrum-function of the DOA is estimated, the peaks of which reveal the DOAs in quest. In this category belong the well-known MUSIC and ESPRIT algorithms [102, 104] and their numerous variations [7, 11, 34, 123, 124]

- time difference of arrival approaches: Among all the approaches proposed in the literature, numerous ones are based on the time difference of arrival (TDOA) [19] at

different microphone pairs to estimate the direction of arrival. Many of them use the generalized cross-correlation phase transform (GCC-PHAT) [62], which has significant limitations in the case of multiple sources, reverberant environments and/or when microphones are placed around rigid bodies. Such limitations have been partially solved by considering ratios of the GCC-PHAT peaks [9] and by using the redundant information contained in more than two microphones [12].

- independent component analysis: During the first decade of the 21st century various methods appeared based on independent component analysis (ICA) as a by-product of research on the blind-source separation problem. The goal of these methods is to estimate the matrix that mixes the original source signals into the mixture signals received at the array microphones. The mixing matrix contains information of the DOA of the sources, thus its accurate estimation can lead to accurate DOA estimation of the sources. However, one of the basic drawbacks of ICA methods is the need of a high observation length in order to accurately estimate the mixing matrices, which, in turn, reduces the responsiveness of a potentially adopting scheme. Apart from that, estimating the mixing matrices is a computationally costly operation which hinders the real-time functionality of such algorithms. Traditional ICA methods are limited to overdetermined cases, i.e., when the number of sources is lower than the number of sensors, however recent methods have been proposed to overcome this limitation as the one proposed in [87]. Some other proposed works on ICA for DOA estimation can be found in [69, 71, 72, 74, 103].

- sound intensity estimation based approaches: Methods that rely on the estimation of the net flow of the sound energy, i.e., the sound intensity, appeared quite recently in the literature. Even though the estimation of the sound intensity can be a demanding task requiring dedicated hardware [25], approximations to the sound intensity value have been proposed which require simple signal processing procedures, such as in [57, 95]. DOA estimation methods that rely primarily on the estimation of the sound intensity vector can also be found in [36, 45–47, 79].

- sparse component analysis approaches: Sparse component analysis (SCA) methods [41] may be seen as natural extensions of multiple sensor single source localization methods to multiple source localization. They basically assume that sources are sparse in an analysis domain obtained after a sparsifying transform (usually a short-time Fourier transform (STFT)) and that, as a consequence, one source is dominant over the others in some time-frequency windows or "zones". Using this assumption, the multiple source propagation estimation problem may be rewritten as a single-source one in these windows or zones. Their main advantage is their flexibility to deal with not only overdetermined configurations, i.e., the cases where the number

of sources is lower than the number of sensors, but also with underdetermined ones, where the number of sources is higher than that of the sensors. Since in these methods the multiple source localization problem is converted to a single-source one in the detected single-source areas at the transformation domain, this category contains methods that could be characterized as hybrids that combine the benefits of a method from another category with the ones of the SCA category. Such examples can be found in [112].

In the following sections we will refer in details to some of the aforementioned algorithms, based on their popularity and whether they were used in our proposed work for comparative purposes.

2.1 Subspace approaches

The interest in the decomposition of the covariance matrix of a sensor array in order to derive the signal parameters goes back in 1930 [63]. However subspace approaches were in fact established with the introduction of the multiple signal classification (MUSIC) algorithm in 1986 by Ralph O. Schmidt [104]. Three years later another, equally popular, subspace algorithm was proposed by Roy and Kailath, known as ESPRIT, i.e., estimation of signal parameters via rotational invariance techniques [102]. In the years that followed several extensions, enhancements, and modifications have been proposed for both MUSIC and ESPRIT. In the following sections we describe the original structure of these estimators as well as the modifications that are of interest in the context of DOA estimation of audio sources.

2.1.1 The multiple signal classification (MUSIC) algorithm

MUSIC was proposed as a "high resolution" parametric spectral analysis algorithm suitable for narrowband signals. Thus, assuming that N_S independent sources exist in the environment and create a wavefield perceived by a Q-microphones sensor array, the $Q \times Q$ covariance matrix of the array, C_X, can be expressed and decomposed as :

$$C_X = VC_SV^H + \sigma_n^2 I = U_S \Lambda_S U_S^H + \sigma_n^s U_n U_n^H, \tag{2.1}$$

where V is the array's steering matrix, C_S is the source signals covariance matrix, I is the identity matrix, U_s is the $Q \times N_S$ signal subspace matrix of N_S eigenvectors and Λ_S is a $N_S \times N_S$ diagonal matrix with the N_S corresponding eigenvalues on the main diagonal. The matrix U_n contains the eigenvectors that span the noise subspace, assuming noise to be zero-mean, stationary, temporally and spatially white of variance σ_n and uncorrelated with all sound source signals [7, 63, 104].

The narrowband MUSIC pseudospectrum, the N_S local maxima of which reveal the DOAs of interest, is estimated for every direction of arrival, φ, as:

$$h_{\mathrm{M}}(\varphi) = \frac{1}{V U_{\mathrm{n}} U_{\mathrm{n}}^{H} V^{H}} \tag{2.2}$$

Broadband MUSIC extensions

MUSIC is an algorithm that was originally developed for narrowband signals. Thus, for wideband signal localization—which is the case when working with sound—it needs to be extended in a wideband suited fashion. In the literature one can find various extensions of the MUSIC algorithm from narrowband to wideband cases [7, 11, 34, 54, 55, 68, 123, 124]. We will first refer to the one that comes as a natural extension, while remains simple and straightforward.

Extending the MUSIC algorithm to involve broadband signals requires the application of the algorithm for every frequency component of interest. Assuming our signals contain frequency domain components indexed as $k = 1, \ldots, N_K$, this straight strategy implies the estimation of N_K covariance matrices and therefore the estimation of N_K MUSIC pseudospectra, one for each frequency bin. A common approach, then, is to estimate the average, e.g., the arithmetic mean [7] over all the frequency bins:

$$h_{\mathrm{MB}}(\varphi) = \frac{1}{N_K} \sum_{k=1}^{N_K} h_{\mathrm{M}}(k; \varphi) \tag{2.3}$$

or the geometric mean [121] over the frequency range of interest as:

$$h_{\mathrm{MB}}(\varphi) = \left[\prod_{k=1}^{N_K} h_{\mathrm{M}}(k; \varphi) \right]^{1/N_K} \tag{2.4}$$

In [7] Argentieri and Danes developed one more broadband version of the MUSIC algorithm with involves the use of frequency invariant beamformers in order to estimate a focalized array and noise covariance matrices to a reference frequency component. This way one escapes from the computationally costly operation of singular value decomposing the array covariance matrix at each frequency of interest. This variation is called beamspace broadband MUSIC and for the focalization process it uses the approach presented in [119] which relies on modal analysis and beamforming.

A broadband variation of MUSIC with the formulation being developed entirely in the spherical harmonic domain (SHD) was recently presented by Nadiri and Rafaely in [82]. In their algorithm the covariance matrix, C_a, of the received soundfield is estimated in the

SHD as:

$$C_a = \frac{1}{TK}\sum_{j_t=0}^{T-1}\sum_{j_k=0}^{K-1} a_{lm}(\tau - j_t, k - j_k), \tag{2.5}$$

where a_{lm} are the spherical harmonic coefficients of order l and degree m, τ and k denote the time and frequency indexes and N_T and N_K denote the interval of time and frequency smoothing respectively. The algorithm proceeds with the selection of time-frequency (TF) bins, appropriate for DOA estimation by applying the direct-path dominance (DPD) test, i.e., the selection of those TF bins where the ratio between the first and second highest eigenvalues of the covariance matrix, C_a, is greater than a user defined threshold. According to the method, at those selected bins the received signal contains information of only direct sound, thus it is free, with high probability, of any contaminating reflections. The next step involves the decomposition of the covariance matrix into the signal and noise subspaces. The MUSIC pseudospectrum is estimated either by summing together all spatial spectral from the selected bins or by clustering the signal subspaces to N_S clusters, as many as the known number of sources, and estimating one MUSIC pseudospectrum for each speaker.

2.1.2 Estimation of signal parameters via rotational invariance techniques-The ESPRIT algorithm

The second most well known subspace algorithm for DOA estimation is the estimation of signal parameters via rotational invariance techniques (a.k.a. the ESPRIT algorithm), originally presented in [102]. The algorithm depends on a displacement invariance property of the array. The microphones have to appear in pairs with identical displacement vectors, that is the array can be described as being comprised by two identical subarrays, being psychically displaced from each other by a constant distance . ESPRIT is computationally more efficient in comparison to MUSIC, since it does not require knowledge and scanning of the array's manifold, however it requires double the size of sensors unless the array geometry is that of a linear array. The method relies on the singular value decomposition (SVD) of the array's covariance matrix and on the estimation of the rotation operator, i.e., a diagonal matrix that relates the measurements from one subarray to those of the displaced subarray, and contains the information of the DOA.

ESPRIT has been formulated in the cylindrical harmonics domain by Teutsch and Kellerman in [115], where the subarrays of individual microphones have been replaced by "sub-modal arrays" of individual harmonics. The method was named EB-ESPRIT and for the evaluation a circular array mounted on a rigid cylindrical baffle was used. ESPRIT was further combined with a well known method for blind source separation (BSS), DUET, [59] in order to deal with limitations of the DUET algorithm, such as the strong WDO assumption

and the constraint of its functionality on two mixtures. The combination of ESPRIT and DUET gave birth to the DESPRIT algorithm for the BSS problem presented in [101].

2.2 Independent component analysis for DOA estimation

DOA estimation methods based on independent component analysis (ICA) emerged through research on blind source separation (BSS) for convolutive mixtures. In ICA—as evident from its naming—the input source signals are assumed mutually independent and, thus, separation and localization is achieved by minimizing some dependency measure between the estimated output signals. In other words, the goal is to find a separating matrix (demixing matrix) such that the output signal vectors are as independent as possible. ICA approaches for BSS have been developed both for the time and frequency domains. In this section we will refer to a method proposed by Nesta et al [86] in 2012 which was also used in comparisons with our proposed framework. For a more in depth discussion on ICA and convolutive BSS with ICA, the interested reader is referred to [22, 53].

2.2.1 Generalized state coherence transform

The generalized state coherence transform (GSCT) defines a multivariate likelihood function of the time-delay associated to each source [86] and extends the authors' previous work on single source TDOA estimation to multiple sources. The GSCT method can be divided into two main parts, the estimation of the demixing matrices at each frequency component (frequency domain ICA) and the extraction of the DOAs from the estimated demixing matrices. Estimating the demixing matrices in the frequency domain overcomes an inherent problem of traditional ICA methods, enabling the algorithm to estimate more sources than sensors, given that the sources' number does not exceed the sensors' number at each frequency of interest.

For the first step of the GSCT method any ICA estimator can be used, such as the joint approximate diagonalization of eigenmatrices (JADE) method [18] which exploits the fourth-order cumulants relying on the statistical independence of the sources, or the recursively regularized (RR-ICA) algorithm [88] which exploits the consistency of the mixing matrices across frequencies and the continuity of the time activity of the sources. Given the demixing matrices, the GSCT function is then estimated, which is a multivariate likelihood measure between the acoustic propagation model and the observed propagation vectors, obtained by row-wise ratios between the elements of each inverted demixing matrix. The GSCT function is given by:

$$G(\mathbf{D}) = \sum \mathbf{g}(E(\mathbf{D})), \tag{2.6}$$

where $\mathbf{D}$ is the model vector of time differences of arrival between adjacent microphones,

$E(\mathbf{D})$ is the error measure between the model and the observation vectors and $\mathbf{g}(E(\mathbf{T}))$ is a non-linear monotonic function which decreases as the error measure increases. The summation in Eq. (2.6) takes place over all frequency components and ratios in all the columns of the inverted demixing matrices. As a non-linear function, $\mathbf{g}(E(\mathbf{T}))$, the authors propose a kernel-based one [86]:

$$\mathbf{g}(E(\mathbf{T})) = \frac{1}{2\pi k \frac{F_s}{N_K}} e^{-E^2(\mathbf{D})/(2a_k^2)}, \tag{2.7}$$

where a_k is a spatial selectivity controlling parameter, F_s denotes the sampling frequency and N_K is the total number of frequency sampling points. By associating each time delay vector, $\mathbf{D}$, of the propagation model to its corresponding DOA, one can estimate the DOAs as the local maxima of the GSCT function.

2.3 Sound intensity estimation methods

Sound intensity is a measure of the flow of sound energy through a surface per unit area, in a direction perpendicular to this surface. Thus, knowledge of the sound intensity vector reveals the DOA of the source generating the sound energy as the direction opposite to the direction of the sound intensity. Intensity-based methods utilize a pressure and a particle velocity component to analyze the sound field. In practice, the pressure and particle velocity are estimated with an omnidirectional and three dipole microphones respectively [113]. Due to its tolerable latency, the intensity vector is an ideal candidate for real-time DOA estimation and has been previously employed in time-frequency domain spatial sound processing [95]. Its performance has been examined in reverberant environments [67] and a pseudo intensity vector has been formulated in the SHD [36, 57].

The intensity vector is defined as [37]:

$$\mathbf{I} = \frac{1}{2}\mathrm{Re}\left\{p^* \mathbf{v}\right\}, \tag{2.8}$$

where p is the sound pressure and $\mathbf{v}$ is the particle velocity vector.

2.3.1 Intensity vector approximation with B-format signals for DOA estimation

In [95] the sound intensity vector—and consequently the DOA component of the DirAC system—is estimated by using the components of a B-format signal [80], that is the omnidirectional signal of the B-format, $w(\tau, k)$ is used for the estimation of the sound pressure and the three figure-of-eight signals, $x(\tau, k)$, $y(\tau, k)$ and $z(\tau, k)$ are used for the estimation

of the velocity, i.e.,:

$$\begin{aligned} p &= w(\tau,k) \\ \mathbf{v} &= \tfrac{1}{Z_0\sqrt{2}}\left[x(\tau,k), \quad y(\tau,k), \quad z(\tau,k)\right]^T, \end{aligned} \tag{2.9}$$

where Z_0 is the characteristic acoustic impedance of air. The B-format channels are represented here at the TF domain, with τ denoting the time index and k denoting the frequency index. Since the dipole components of the B-format signal are scaled by a factor of $\sqrt{2}$ [64], in Eq. (2.11) we have the term $\frac{1}{\sqrt{2}}$. More on the B-format signals can be found in [39].

In [61] the DOA estimation is performed in the same spirit as in [95].

2.3.2 Pseudointensity vectors for DOA estimation

In [57] the authors proposed to obtain an estimate of the sound intensity vector by means of the zero and first-order spherical harmonic signals (or eigenbeams). They called this estimate a pseudointensity vector (PIV), formulated as:

$$\mathbf{I}(\tau,k) = \frac{1}{2}\mathrm{Re}\left\{\left[\frac{s_{00}^*(\tau,k)}{b_0(k)}\right]\begin{bmatrix} s_x(\tau,k) \\ s_y(\tau,k) \\ s_z(\tau,k) \end{bmatrix}\right\}, \tag{2.10}$$

where the sound pressure is approximated by the zero-th order equalized eigenbeam (see also Section C.5.1) and the particle velocity is approximated by a vector whose each element corresponds to averages of 1^{st} order steered eigenbeams with the negative phase towards the x, y and z-axis respectively. Thus, in relation with Eq.(2.8),

$$\begin{aligned} p &= \tfrac{s_{00}(\tau,k)}{b_0(k)} \\ \mathbf{v} &= \left[s_x(\tau,k), \quad s_y(\tau,k), \quad s_z(\tau,k)\right]^T. \end{aligned} \tag{2.11}$$

Each eigenbeam average is estimated as

$$s_a(\tau,k) = \sum_{m=-1}^{1} Y_{1m}(\Omega_a)s_{1m}(\tau,k), \quad a = \{x,y,z\}, \tag{2.12}$$

where Ω_a is $(0,\pi)$, $(0,-\pi/2)$, and $(-\pi/2,0)$ for each axis. The eigenbeams can be estimated using Eqs. (C.22), (C.25) and Eq.(C.29).

The PIVs are estimated at each TF point and the final DOA estimates are obtained utilizing K-means clustering, assuming the number of sources to be known.

Combining PIVs and the DPD test for enhanced DOA performance

The authors of [36] and [82] combined their approaches in order to improve the DOA estimation of multiple audio sources in the 3D space. In their recent publication in [79] they proposed to apply the DPD test in order to identify TF bins that are dominated by a single direct path signal, and then, instead of performing the MUSIC algorithm over the selected bins, to evaluate the PIVs at those bins. They proposed two approaches at the aforementioned concept. For the first one, the PIVs are estimated in neighborhoods of the selected TF points and finally averaged. The second approach exploits the SVD of the spatial correlation matrix (Eq. (2.1.1)) and estimates the PIVs using the eigenbeams of the signal subspace.

Augmented intensity vectors for DOA estimation

In a series of publications [45–47] Hafezi et al enhance the DOA estimation through the PIVs by exploiting possibly available higher order spherical harmonics. The authors create a grid around the vicinity of the DOA indicated by the PIV and compute a direction-dependent error function for each point of the grid, thus the point with the smallest error provides a refined DOA estimation.

2.4 Sparse component analysis for DOA estimation

Sparse component analysis (SCA) approaches have also originated from the BSS research area. As revealed from their naming, these methods rely on some kind of sparsity of the signals of interest in some processing domain (usually the TF domain). In order for the sparsity to be detected, quantified and exploited, a sparsity (or - on the contrary- activity) measure is estimated which indicates in which areas of the processing domain there is activity of one or more of the signals of interest.

A very popular sparsity assumption, named W-disjoint orthogonality (WDO) was presented in [122] and assumes that in each time-frequency window, at most one source is active. From a signal processing point of view, WDO is a nice assumption which is almost fulfilled by speech signals in anechoic environments. However, this assumption does not hold in reverberant conditions [106] and/or when source signals are musical. On the contrary, other methods assume that the sources may overlap in the time-frequency domain, except in some tiny time-frequency "analysis zones" where only one of them is active (e.g., [94] and the references within). In [94] Puigt proposed the use of "constant-time single-source analysis zones", i.e., a set of frequency-adjacent time-frequency windows over which a cross-correlation coefficient is estimated which characterizes the analysis zone as single source (i.e., single source zone (SSZ)) or not. Another measure for deciding if only one source is active in a TF point is the estimation of coherence between the signals

received by the microphones as presented in [78]. In [95] the authors decide on the selection of a TF area according to the intensity vector estimates in the area. If the estimates agree the area is used for DOA estimation, otherwise the estimates are discarded. The DPD test proposed by Nadiri and Rafaely in [82] is also an activity detection measure (see also Section 2.1.1).

Methods that belong in the SCA family employ one of the aforementioned criteria in order to detect areas where only one source is active and proceed with the DOA estimation over the detected areas by employing (novel or already proposed) single source DOA methods. In this family we can categorize methods described in the preceding sections, e.g., [47, 79, 82, 95]. The framework proposed in this thesis also belongs in the family of SCA methods.

2.5 Counting the number of simultaneously active sources

A problem closely related to that of DOA estimation is the estimation of the number of simultaneously active source, what we often refer to as counting. In general most DOA estimation methods proposed in the literature assume that the number of sources is somehow known a priori. However, in real life conditions this number is not usually known and we have to either arbitrarily set it or estimate it from the received data.

The most well known criteria for estimating the number of sources from the received mixtures of signals were suggested Wax and Kailath in [120], where the authors faced the problem of counting as a model selection one and applied the information theoretic criteria introduced by Akaike (Akaike infromation criterion, a.k.a, AIC) and by Schwarz and Rissanen (minimum description length criterion, a.k.a, MDL).

Instead of minimizing a criterion such as the aforementioned, other approaches approach the problem of counting by trying to estimate the number of significant eigenvalues of the array's covariance matrix. However such approached require setting a threshold for distinguishing between significant and non-significant values, which in practice gives controversial results [54].

Another category in the problem of counting the number of sources is that of estimating the number of clusters in clustering-based methodss. Such an example is the one proposed by Arberet et al in [6], where the number of clusters is decided through a sequential procedure utilizing a confidence and a dissimilarity measure.

In histogram-based algorithms the direct approach is to estimate the number of significant peaks in the DOA histogram, such as in [70]. In previous work [91], we have proposed two ways of tackling the problem, either by peak-searching (PS) and dynamically thresholding the cardinality of a peak and its neighborhood, or by estimating the envelope of the histogram through linear predictive coding (LPC), and then the number of all its local maxima (see also Appendix D).

Part II

DRACOSS: Direction of Arrival and Counting of Sound Sources

Chapter 3
Description of the DRACOSS framework, principles and concepts

In this chapter we describe the DRACOSS framework, the main building blocks and the principles and concepts that govern the functionality of our proposed framework for DOA estimation and counting of sound sources.

The beginning of the development of the DRACOSS framework took place together with our work on DOA estimation utilizing a circular microphone array. The fundamental concepts were established, i.e., the exploitation of the sparsity of audio sources in an appropriately chosen transformation domain and the post-processing of local DOA estimates in the form of histograms, following our early works in DOA estimation [44, 91, 92].

At that time, the sparsity of audio signals was mostly of interest to the audio separation community which was dealing with the demixing of audio mixtures and the ever-popular cocktail party problem [2, 5, 30, 31, 59, 100, 101, 109, 118, 122]. From this community the domain of SCA emerged ([41] in [22]), along with very interesting methods from the ICA domain ([5, 22, 53, 85, 88, 105]).

3.1 DRACOSS building blocks

The DRACOSS framework is comprised by four fundamental building blocks, i.e., :

1. exploitation of the sources sparsity
2. local DOA estimation
3. histogram formation
4. histogram post-processing for final DOA estimation and counting

Each of the aforementioned blocks will be further discussed in the proceedings sections.

3.1.1 Exploitation of the sources sparsity

As previously mentioned, one of the basic ideas in DRACOSS is the exploitation of the sparsity of audio signals in the TF domain. Thus, we can find areas in the TF domain that a source occurs alone or is at least dominant in comparison with other simultaneously present sources.

We developed our DOA estimation method relying on the relaxed sparsity assumption proposed by Puigt and Deville in [94]. According to it, the audio sources may overlap in the transformation domain except in some areas where each source of the mixture occurs alone.The detection of those areas is based on the mean value of a correlation coefficient between pairs of microphones comprising the recording device, thus we call and refer to it as the mean correlation coefficient (MCC). Areas that exhibit an MCC higher that a user-defined threshold are considered as dominated by a single source, thus they are referred to as single-source zones (SSZs). A graphical representation of detected SSZs in the STFT transformation of a microphone signal can be observed in Fig. 3.1a, where hypothetically two sources (the red and the blue) are active.

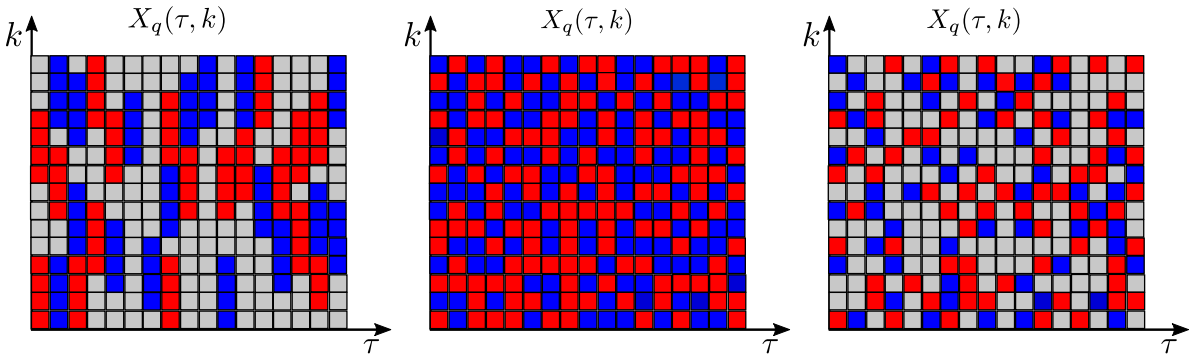


Figure 3.1: Visualization of detected single-source activity areas in the STFT of a microphone signal, $X(\tau,k)$ where hypothetically two sources are active, one represented with the red and the other with the blue color:
3.1a Detected SSZs using the MCC criterion : the red and blue zones correspond to SSZs of hypothetically two different sources, while the grey area indicates overlapping of the sources activity.
3.1b Each TF point is dominated by a single source.
3.1c According to the DPD test, some TF points will be detected as single source activity points. The gray-colored ones represent those that failed the DPD test.

The description of the formulas for estimating the MCC will be described in Section 4.1.1.

The theorem supporting the detection of SSZs through the estimation of the MCC and all necessary assumptions proposed by Puigt and Deville are described in Appendix B.0.3 for completeness of the text.

We have to note here that there are other proposed ways for detecting areas of single-source activity, as also mentioned in Section 2.4. The most widely known is by adopting the WDO assumption, presented by Yilmaz and Rickard in [122], which states that at each TF point of the mixtures' spectrum only one source is active. This is again graphically represented in Fig. 3.1b. In our proposed work we have used the WDO assumption in comparative results in Sections 4.2.1 and 5.2. It is also the sparsity measure we used in our beamforming-based method described in Section 5.4.

Another recently proposed criterion for selecting single source TF points is that of the direct path dominance (DPD) test proposed by Nadiri and Rafaely in [82] in the context of the MUSIC algorithm in the SHD. According to the DPD test a TF point is single source if the effective rank of the corresponding array covariance matrix is equal to one. We observe the graphical representation of the DPD test in Fig. 3.1c. We have used the test in our proposed MUSIC approach in the context of the DRACOSS framework in Section 5.5.

3.1.2 Local DOA estimation

Having detected all available SSZs, the multiple source signal localization problem has been nicely transformed to a single source one, since now one can employ any single source DOA estimation method to the SSZs.

In DRACOSS development for 2D spaces we have used a single source DOA estimation method specifically designed for uniform circular arrays, proposed by Karbasi and Sugiyama in [60]. The method is described in more details in Section 4.1.2.

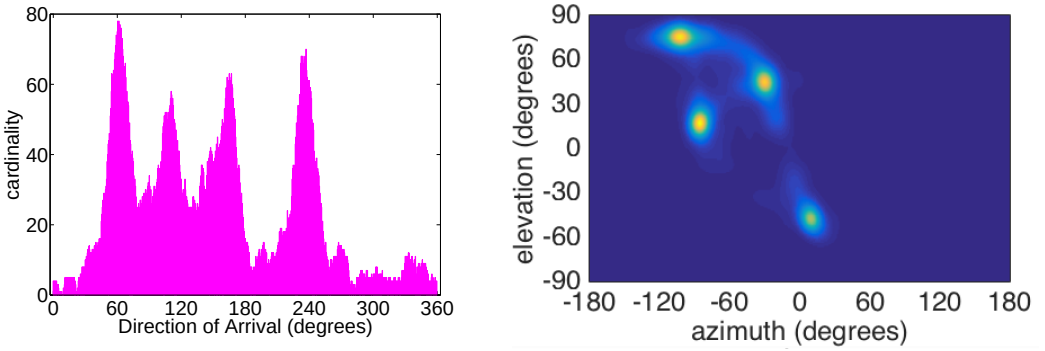
In DRACOSS development for 3D spaces we have used an estimator of the intensity vector for the DOA estimation formulated in the spherical harmonic domain as it was proposed in [57].

Other DOA estimators that we have used are based on beamforming in the SHD (see Section 5.4) or by utilizing locally the MUSIC speudospectra (in Section 5.5).

3.1.3 Histogram formation

The third step of the DRACOSS framework is the formation of a histogram of all local DOA estimates. Histograms are a very easy and straightforward way of visualizing the gathered information of the local DOA estimates. They are of single-dimensionality or 2D depending on the dimensionality of the local DOA estimates of the previous step. An example 1D histogram and 2D histogram of a scenario with four involved sources can be seen in Figure 3.2. Moreover, histograms allow to control the underlining complexity by choosing an appropriate bin width while they can be manipulated with various different approaches in

order to conclude with a final DOA and counting estimation which is also the last building block of DRACOSS.



(a) (b)

Figure 3.2: Histograms of local DOA estimates:
3.2a A 1D histogram of azimuthal DOA estimates, where one can detect four distinct peaks corresponding to the four active sources.
3.2b A 2D histogram of pairs of elevation and azimuth DOA estimates, where one can detect four highlighted areas corresponding to the four active sources.

3.1.4 Histogram post-processing for final DOA estimation and counting

The last step of DRACOSS entails the manipulation of the formed histograms in order to extract the information of the number and corresponding DOAs of all active sources. We apply an iterative procedure, which includes the correlation of the histograms with pulses in order to detect highlighted areas. For one-dimensional histograms we apply Blackman windows. Narrower windows are used to accurately detect peaks in the histogram, while wider windows estimate the contribution of a peak, and consequently of a source to the histogram. Following a thresholds logic which dictates that active sources will have a significant contribution to the histogram, we achieve very good counting results as described in Section 4.2.1.

For the processing of 2D histograms we use Gaussian windows. We follow the same iterative procedure in order to detect the presence of a source and its corresponding DOA, assuming the total number of active sources is a priori known (see also Section 5.1.4). Aiming at avoiding the use of thresholds, we show very recent, yet really promising results on counting by training a convolutional neural network (CNN) in Section 5.6.

Chapter 4
Direction of arrival estimation in the two-dimensional space

In this chapter we present the DRACOSS framework for multiple sound source localization and counting in the two-dimensional space, where a uniform circular microphone array is used to overcome the ambiguities of linear arrays. The proposed framework imposes relaxed sparsity constraints on the source signals. Our method is based on detecting time-frequency (TF) zones, where one source is dominant over the others. Using appropriately selected TF components in these "single-source" zones (SSZs), the proposed method jointly estimates the number of active sources and their corresponding directions of arrival (DOAs) by applying a matching pursuit-based approach to the histogram of DOA estimates. DRACOSS is shown to have excellent performance for DOA estimation and source counting, and to be highly suitable for real-time applications due to its low complexity. Through simulations in various signal-to-noise ratio conditions (SNR) and reverberant environments and real environment experiments, we indicate that our method outperforms other state-of-the-art DOA and source counting methods in terms of accuracy, while being significantly more efficient in terms of computational complexity. We note that the underlying concepts are applicable to any microphone array topology, highlighting the flexibility of the proposed DRACOSS framework.

4.1 DRACOSS in two dimensions

We consider a uniform circular array of Q microphones, with N_S active sound sources located in the far–field of the microphone array (see Appendix B.0.2). Assuming the free-field model, the signal received at each microphone, q_i, is

$$x_i(t) = \sum_{j=1}^{N_S} a_{ij} s_j \left(t - t_i(\varphi_j)\right) + n_i(t), \qquad i = 1, \cdots, Q, \tag{4.1}$$

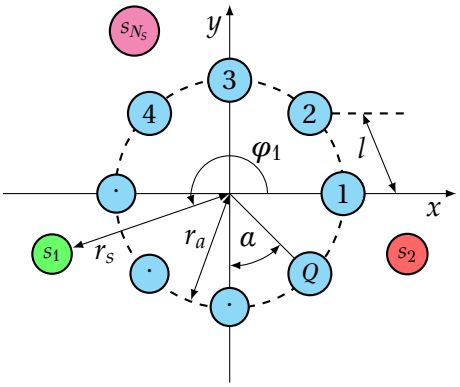


Figure 4.1: Circular sensor array configuration. The microphones are numbered 1 to Q and the sound sources are s_1 to s_{N_S}.

where s_j is one of the N_S sound sources at distance r_s from the center of the microphone array, a_{ij} is the attenuation factor and $t_i(\varphi_j)$ is the propagation delay from the j^{th} source to the i^{th} microphone. φ_j is the DOA of the source s_j observed with respect to the x-axis (Figure 4.1), and $n_i(t)$ is an additive white Gaussian noise signal at microphone q_i that is uncorrelated with the source signals, $s_j(t)$, and all other noise signals.

For one given source, the relative delay between signals received at adjacent microphones, hereafter referred to as microphone pair $\{q_i q_{i+1}\}$, with the last pair being $\{q_Q q_1\}$, is given by [60]

$$\begin{aligned}\tau_{q_i q_{i+1}}(\varphi_j) &\triangleq t_i(\varphi_j) - t_{i+1}(\varphi_j) \\ &= l\sin(\pi - \varphi_j + (i - \frac{1}{2})a)/c,\end{aligned} \tag{4.2}$$

where a and l are the angle and distance between $\{q_i q_{i+1}\}$ respectively and c is the speed of sound. Since the microphone array is uniform, a and l are given by:

$$a = \frac{2\pi}{Q}, \qquad l = 2r_a \sin\frac{a}{2}, \tag{4.3}$$

where r_a is the array radius. We note here that in Eq. (4.2) the DOA φ_j is observed with respect to the x-axis, while in [60] it is observed with respect to a line perpendicular to the chord defined by the microphone pair $\{q_1 q_2\}$. We also note that all angles in Eqs. (4.2) and (4.3) are in radians.

We aim to estimate the number of the active sound sources, N_S, and corresponding DOAs, φ_j, employing the proposed DRACOSS framework. We will thus exploit and detect the sparsity of the involved source signals in the TF domain, in order to estimate local DOAs. We will use the local DOA estimates to form a histogram, which we will process in order to acquire the final DOAs and number of sources. It should be noted that even

though we assume the free-field model, DRACOSS is shown to work robustly in both simulated and real reverberant environments.

4.1.1 Step 1: sound sources sparsity

We adopt the sparsity criterion proposed in [94]. We partition the incoming data in overlapping time frames, on which we compute a Fourier transform, providing a time-frequency (TF) representation of observations. We then define a "constant-time analysis zone", (τ, K), as a series of frequency-adjacent TF points (τ, k). A "constant-time analysis zone", (τ, K) is thus referred to a specific time frame τ and is comprised by K adjacent frequency components. In the remainder of the section, we omit τ in the (τ, K) for simplicity.

We assume the existence, for each source, of (at least) one constant-time analysis zone—said to be "single-source"—where one source is "isolated", i.e., it is dominant over the others. This assumption is much weaker than the WDO assumption [122], since sources can overlap in the TF domain except in these few single-source analysis zones. We note that the WDO assumption could be consider as a sparsity measure and we will show comparative performance results in Section 4.2

For any pair of signals (x_i, x_j), we define the cross-correlation of the magnitude of their TF transform $(X_i(k), X_j(k))$ over an analysis zone as:

$$R'_{i,j}(K) = \sum_{k \in K} \left| X_i(k) \cdot X_j(k) \right|. \tag{4.4}$$

We then derive the correlation coefficient, associated with the pair $\left\{q_i q_j\right\}$, as:

$$r'_{i,j}(K) = \frac{R'_{i,j}(K)}{\sqrt{R'_{i,i}(K) \cdot R'_{j,j}(K)}}. \tag{4.5}$$

Our approach for detecting SSZs is based on the following theorem [94] (see also Section B.0.3):

Theorem 1 *A necessary and sufficient condition for a source to be isolated in an analysis zone* (K) *is*

$$r'_{i,j}(K) = 1, \qquad \forall i, j \in \{1, \ldots, Q\}. \tag{4.6}$$

We detect and characterize as single source zones all constant-time analysis zones that satisfy the following inequality:

$$\overline{r'}(K) \geq 1 - \epsilon, \tag{4.7}$$

where $\overline{r'}(K)$ is the average correlation coefficient between pairs of observations of adjacent microphones and ϵ is a small user-defined threshold.

In our proposed framework we use analysis zones which are constant in time, thus the detected SSZs are also defined in a constant time frame and over consecutive frequency bins. We have the option of utilizing analysis zones constant in frequency and over consecutive time frames as in [93], however we have not explored this option yet.

4.1.2 Step 2: single-source local DOA estimator

Since we have detected all SSZs, we can apply any known single source DOA algorithm over these zones. We propose a modified version of the algorithm in [60]. We have chosen this algorithm because it is computationally efficient and robust in noisy and reverberant environments [60, 92].

We consider the circular array geometry (Figure 4.1) introduced in Section 4.1. The phase of the cross-power spectrum of a microphone pair is evaluated over the frequency range of a single-source zone as:

$$G_{q_i q_{i+1}}(k) = \angle R_{i,i+1}(k) = \frac{R_{i,i+1}(k)}{\left|R_{i,i+1}(k)\right|}, \qquad k \in K, \tag{4.8}$$

where the cross-power spectrum is

$$R_{i,i+1}(k) = X_i(k) \cdot X_{i+1}(k)^* \tag{4.9}$$

and * stands for complex conjugate.

We then calculate the phase rotation factors [60] as:

$$G^{(k)}_{q_i \to q_1}(\varphi) \triangleq e^{-j\frac{2\pi k F_s}{N_k}\tau_{q_i \to q_1}(\varphi)}, \tag{4.10}$$

where $\tau_{q_i \to q_1}(\varphi) \triangleq \tau_{q_1 q_2}(\varphi) - \tau_{q_i q_{i+1}}(\varphi)$ is the difference in the relative delay between the signals received at pairs $\{q_1 q_2\}$ and $\{q_i q_{i+1}\}$, $\tau_{q_i q_{i+1}}(\varphi)$ is evaluated according to (4.2), $\varphi \in [0, 2\pi)$ in radians, F_s is the sampling frequency, N_k is the total number of frequency sampling points and $k \in K$.

We proceed with the estimation of the circular integrated cross spectrum (CICS), defined in [60] as

$$\mathrm{CICS}^{(k)}(\varphi) \triangleq \sum_{i=1}^{Q} G^{(k)}_{q_i \to q_1}(\varphi) G_{q_i q_{i+1}}(k). \tag{4.11}$$

The DOA associated with the frequency component k in the single-source zone with frequency range K is estimated as,

$$\hat{\varphi}_k = \arg \max_{0 \le \varphi < 2\pi} |\mathrm{CICS}^{(k)}(\varphi)|. \tag{4.12}$$

In each SSZ we focus only on "strong" frequency components in order to improve the accuracy of the DOA estimation. In previous works [44, 91, 92], we used only one frequency, corresponding to the strongest component of the cross-power spectrum of the microphone pair $\{q_i q_{i+1}\}$ in a SSZ, giving us a single DOA for each single-source zone. We now propose the use of d frequency components in each single-source zone, i.e., the use of those frequencies that correspond to the indices of the d highest peaks of the magnitude of the cross-power spectrum over all microphone pairs. This way we get d estimated DOAs from each single-source zone, improving the accuracy of the overall system.

This is illustrated in Figure 4.2, where we plot the DOA estimation error versus the SNR for various choices of d. It is clear that using more frequency bins (the terms frequency bin and frequency component are used interchangeably) leads in general to a lower estimation error. We have to keep in mind, though, that increasing d increases the computational complexity, which should be taken into account for a real-time system.

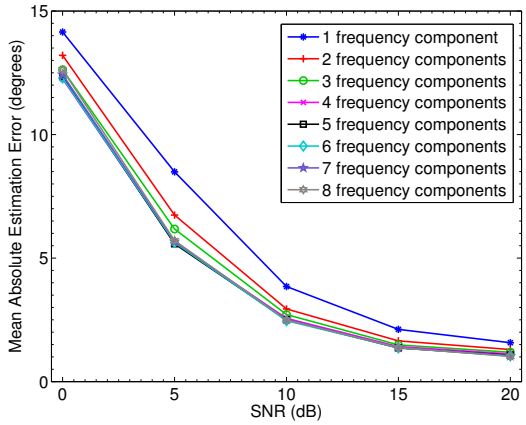


Figure 4.2: DOA estimation error vs SNR in a simulated environment. Each curve corresponds to a different number of frequency components used in a single-source zone.

4.1.3 Step 3: histogram formation

In the previous sections we described the first two steps of the DRACOSS framework, i.e., how we determine whether a constant time analysis zone is single-source and how we estimate the DOAs associated with the d strongest frequency components in a single-source zone. Once we have estimated all the local DOAs in the single-source zones (Sections 4.1.1 & 4.1.2), we form a histogram from the set of estimations in a block of N_T consecutive time frames. We smooth the histogram by applying an averaging filter with a window of length

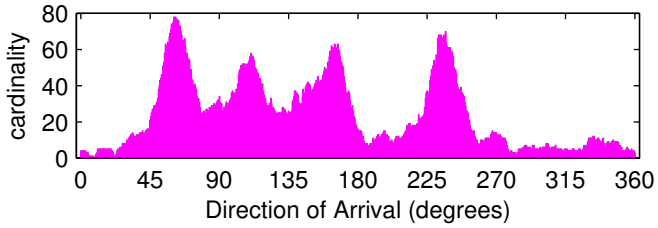


Figure 4.3: Example of a smoothed histogram of four sources (speakers) in a simulated reverberant environment at 20 dB SNR.

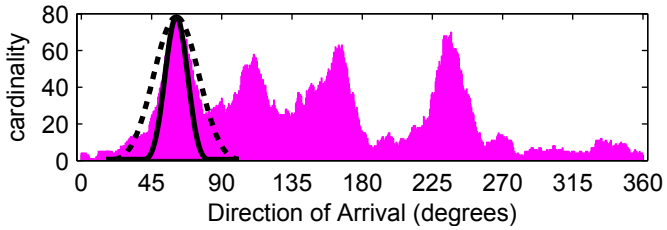


Figure 4.4: A wide source atom (dashed line) and a narrow source atom (solid line) applied on the smoothed histogram of four sources (speakers).

N_L. If we denote each bin of the smoothed histogram as b, its cardinality, $h(b)$, is given by:

$$h(b) = \sum_{i=1}^{N_B} w\left(\frac{b \times 360^\circ/B - \zeta_i}{N_L}\right), \qquad 1 \le b \le B, \tag{4.13}$$

where B is the number of bins in the histogram, ζ_i is the i^{th} estimate (in degrees) out of N_B estimates in a block, and $w(\cdot)$ is the rectangular window of length N_L. An example of a smoothed histogram of four sources at 60°, 105°, 165°, and 240° at 20 dB SNR of additive white Gaussian noise is shown in Figure 4.3.

4.1.4 Step 4: histogram post-processing

In each time frame we form a smoothed histogram from the estimates of the current frame and the $N_T - 1$ previous frames. Once we have the histogram (the length-B vector, h), our goal is to count the number of active sources and to estimate their DOAs.

If we observe the example histogram of four active sources at 20 dB SNR, shown in Figure 4.3, the four sources are clearly visible and similarly shaped, which inspired us to approach the source counting and DOA estimation problem as one of sparse approx-

imation using source atoms. Thus the idea—proceeding along similar lines to matching pursuit—is to find the DOA of a possible source by correlation with a source atom, estimate its contribution and remove it. The process is then repeated until the contribution of a source is insignificant, according to some criteria. This way we can jointly estimate the number of sources and their DOAs (in contrast with [44, 91] where these tasks were performed separately). Having a collection of DOA estimates, it seems natural to apply clustering methods, such as K-means or a Gaussian mixture model [15], however such approaches assume a known number of sources, while in our approach this information is not known and we intend to derive it.

We chose to model each source atom as a smooth pulse, such as that of a Blackman window, although the choice of the window did not prove to be critical. The choice of the width is key, and reasoning and experiments showed that a high accuracy of the method requires wide source atoms at lower SNRs and narrow source atoms at higher SNRs. Furthermore, the resolution of the method—the ability to discriminate between two closely spaced sources—is adversely affected as the width of the source atom increases. This suggests making the width a parameter in the estimation process, however this would come at the cost of an increase in computational complexity—something we wish to avoid—so we chose to use fixed-width source atoms.

Further investigation revealed that a two-width method provided a good compromise between these constraints, where a narrower width is used to accurately pick the location of each peak, but a wider width is used to account for its contribution to the overall histogram and provide better performance at lower SNRs. This dual-width approach is illustrated in Figure 4.4. Note that the wider width source pulse is centered on the same index as the narrow one.

The correlation of the source pulse with the histogram must be done in a circular manner, as the histogram "wraps" from 359° to 0°. An efficient way to do this is to form a matrix whose rows (or columns) contain wrapped and shifted versions of the source pulse, as we now describe.

Let v_J be a length-J row vector containing a length-J Blackman window, then let u be a length-B row vector whose first J values are populated with v_J and then padded with $B-J$ zeros. Let $u^{(o)}$ denote a version of u that has been "circularly" shifted to the right by o elements, the circular shift means that the elements at either end wrap around, and a negative value of o implies a circular shift to the left.

Choose $J = 2J_0 + 1$ where J_0 is a positive integer. The maximum value of v_J (or equivalently u) will occur at $(J_0 + 1)$-th position. Define $z = u^{(-J_0)}$. The maximum value of the length-B row vector z occurs at its first element. Let the elements of z be denoted z_j, and its energy be given by $E_z = \sum z_j^2$. Now form the matrix Z, which consists of circularly shifted versions of z. Specifically, the o-th row of Z is given by $z^{(o-1)}$.

As previously discussed, we need two widths of source atoms, so let Z_N and Z_W be matrices for the peak detection (denoted by "N" for narrow) and the masking operation (denoted by "W" for wide), respectively, with corresponding source atom widths J_N and J_W.

In order to estimate the number of active sources, N_S, we create γ, a length-$N_{S_{MAX}}$ vector whose elements, γ_i, are some predetermined thresholds, representing the relative energy of the i-th source. Our joint source counting and DOA estimation algorithm then proceeds as follows:

1. Set the loop index $i = 1$
2. Form the product $a = Z_N h_i$
3. Let the elements of a be given by a_j,
   find $j^* = \arg\max_j a_i$ such that j^* is further than $n_w \times B/360^\circ$ from all formerly located maximum indices, where n_w denotes a minimum offset between neighboring sources
4. The DOA of this source is given by $(j^* - 1) \times 360^\circ/B$
5. Calculate the contribution of this source as
$$\delta_i = (z_W^{(i^*-1)})^T \frac{a_{j^*}}{E_{z_N}}$$
6. If $\sum \delta_i < \gamma_i$ go to step 10
7. Remove the contribution of this source as
$$h_{i+1} = h_i - \delta_i$$
8. Increment i
9. If $i \leq N_{S_{MAX}}$ go to step 2
10. $\hat{N}_S = i - 1$ and the corresponding DOAs are those estimated in step 4

It should be noted that this method was developed with the goal of being computationally efficient so that the source counting and DOA estimation could be done in real-time. By real-time we refer to the response of our system within the strict time constraint defined by the duration of a time frame. It should be clear that Z_N and Z_W are circulant matrices and will contain $B - J_N$ and $B - J_W$ zeros on each row, respectively, and both of these properties were exploited to provide a reduced computational load.

4.2 Evaluation

We investigated the performance of DRACOSS for 2D spaces in simulated and real environments. In both cases we used a uniform circular array placed in the center of each environment. All the parameters and their corresponding values can be found in Table 4.1, unless otherwise stated.

Since the radius of the circular array is $r_a = 0.05$ m, the highest frequency of interest is set to 4000 Hz in order to avoid spatial aliasing [16, 33]. Note that the final values chosen for the source atom widths (i.e., $J_{\mathrm{N}} = 81$ and $J_{\mathrm{W}} = 161$) correspond to 40° and 80° respectively. However, due to the shape of the Blackman window, the effective widths are closer to 20° and 40°.

Table 4.1: Experimental parameters

parameter	notation	value
number of microphones	Q	8
sampling frequency	F_s	44100 Hz
array radius	r_a	0.05 m
speaker distance	r_s	1.5 m
frame size		2048 samples
overlapping in time		50%
FFT size		2048 samples
TF zones width	K	344 Hz
overlapping in frequency		50%
highest frequency of interest		4000 Hz
single-source zones threshold	ϵ	0.2
frequency bins/SSZ	d	2
number of bins in the histogram	B	720
histogram bin size		0.5°
averaging filter window length	N_L	5°
history length (block size)	N_T	43 frames (1 second)
narrow source atom width	J_{N}	81
wide source atom width	J_{W}	161
noise type	additive white Gaussian noise	

4.2.1 Simulated Environment

We conducted various simulations in a reverberant room using speech recordings. We used the fast image-source method (ISM) [65, 66] to simulate a room of $6 \times 4 \times 3$ meters, characterized by reverberation time $\mathrm{RT}_{60} = 0.25$ s. The uniform circular array was placed in the center of the room, coinciding with the origin of the x and y-axis. The speed of

sound was $c = 343$ m/s. In each simulation the sound sources had equal power and the signal-to-noise ratio at each microphone was estimated as the ratio of the power of each source signal to the power of the noise signal. In real-life situations we do not expect all sources to experience the same SNR, since some speakers may be further from the array and/or more quiet than others.

In order to more accurately measure the performance all around the array, we simulated each orientation of sources in 10° steps around the array, that is, for each sources' set-up, we moved the set-up by 10° for each next-simulation, leading to a total number of 36 different positionings of the same set-up around the array. This is shown more clearly in Figure 4.8.

The performance of our system was measured by the mean absolute estimated error (MAEE) which measures the difference between the true DOA and the estimated DOA over all speakers, all orientations and all the frames of the source signals, unless otherwise stated.

$$\text{MAEE} = \frac{1}{N_O N_F N_S} \sum_{o,f,s} |\varphi_{(o,f,s)} - \hat{\varphi}_{(o,f,s)}|, \tag{4.14}$$

where $\varphi_{(o,f,s)}$ is the true DOA of the s^{th} speaker in the o^{th} orientation around the array in the f^{th} frame and $\hat{\varphi}_{(o,f,s)}$ is the estimated DOA. N_O is the total number of different orientations of the speakers around the array, i.e., the speakers move in steps of 10° in each simulation, which leads to $N_O = 36$ different runs. N_F is the total number of frames after subtracting $N_T - 1$ frames of the initialization period. We remind the reader that N_S is the number of active speakers in the f^{th} frame.

DOA estimation

We present and discuss our results for DOA estimation assuming known number of active sources. In our first set of simulations we investigated the spatial resolution of our proposed method, i.e., how close two sources can be in terms of angular distance while accurately estimating their DOA. Figure 4.5 shows the MAEE against SNR of additive white Gaussian noise, for pairs of static, continuously active speakers for angular separations from 180° down to 20°. The duration of the speech signals was approximately three seconds. DRACOSS performs well for most separations, but the effective resolution with the chosen parameters is apparently around 30°.

In Figure 4.6 we plot an example DOA estimation of four intermittent speakers across time with the speakers at $60^\circ, 105^\circ, 165^\circ$, and 240°. Note that the estimation of each source is prolonged for some period of time after he/she stops talking or respectively is delayed when he/she starts talking. This is due to the fact that the DOA estimation at each time instant is based on a block of estimates of length N_T frames ($N_T = 43$ frames, which

corresponds to 1 second in this example). We refer to these periods as "transition periods", which we define as the time interval starting when a new or existing speaker starts or stops talking and ending N_T seconds later. An example of a transition period is also shown in Figure 4.6 as the grey-shaded area.

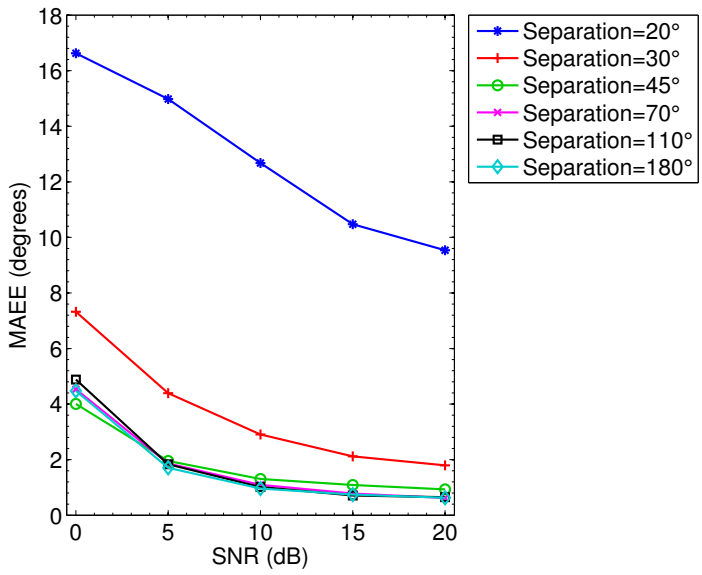


Figure 4.5: DOA estimation error vs SNR for pairs of simultaneously active speakers in a simulated reverberant environment.

We demonstrate how the size of a block of estimates affects the DOA estimation in Figure 4.7. We plot the MAEE versus SNR for the four intermittent speakers scenario for block sizes—also referred to as history lengths—equal to 0.25s, 0.5s and 1s. The speakers were originally located at 0°, 45°, 105° and 180° and even though they were intermittent, there was a significant part of the signals where all four speakers were active simultaneously. There is an obvious performance improvement as the history length increases, as the algorithm has more data to work with in the histogram. However increasing the history also increases the latency of the system, in turn decreasing responsiveness.

Aiming to highlight the consistent behavior of our proposed framework no matter where the sources are located around the array, in Figure 4.8 we plot the absolute error as an average over time, separately for each of six static, simultaneously active speakers and each of 36 different orientations around the array. For the first simulation the sources were located at 0°, 60°, 105°, 180°, 250°, and 315° in a simulated reverberant environment with 20 dB SNR and a one-second history. They were shifted by 10° for each next simulation preserving their angular separations. The duration of the speech signals was

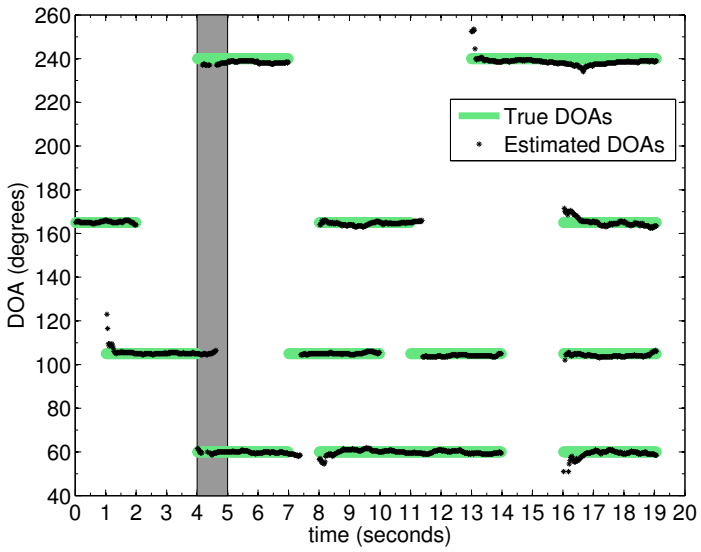


Figure 4.6: Estimation of DOA of four intermittent speakers at $60°, 105°, 165°$, and $240°$ in a simulated reverberant environment with 20 dB SNR and a one-second block size. The gray-shaded area denotes an example “transition period".

approximately 10 seconds and, as already stated, the MAEE was evaluated as the average absolute error in the estimation over time. The MAEE is always below $3°$ for any positioning of the sources around the array for all the sources.

We investigate the robustness to reverberation in Figure 4.9, which shows the MAEE versus SNR for three static, continuously active speakers originally located at $0°$, $160°$, and $240°$ for reverberation time $\mathrm{RT}_{60} = \{0.25, 0.4, 0.6\}$ s. For low reverberation conditions—$\mathrm{RT}_{60} = 0.25$ s—the proposed method performs very well for all SNR conditions as was expected and shown in the preceding results. For medium reverberation with $\mathrm{RT}_{60} = 0.4$ s and source atom widths $J_{\mathrm{W}} = 161(80°)$ and $J_{\mathrm{N}} = 81(40°)$ the MAEE is low for high SNR but increases rapidly for lower signal-to-noise ratios. However, by using wider pulses—i.e., $J_{\mathrm{W}} = 241(120°)$ and $J_{\mathrm{N}} = 141(70°)$—we can mitigate erroneous estimates due to reverberation and keep the error lower than $10°$ for all SNR values. For $\mathrm{RT}_{60} = 0.6$ s—which could characterize a highly reverberant environment—the DOA estimation is effective for SNR values above 5 dB, exhibiting an MAEE lower than $7°$, when using $J_{\mathrm{W}} = 241(120°)$ and $J_{\mathrm{N}} = 181(90°)$. Note that increasing the source atom widths improves the DOA estimation accuracy, but also decreases the resolution of the method.

DRACOSS is a DOA estimation framework, not a tracking one. However motivated by its good performance, we aimed at investigating its tracking potential and we ran simu-

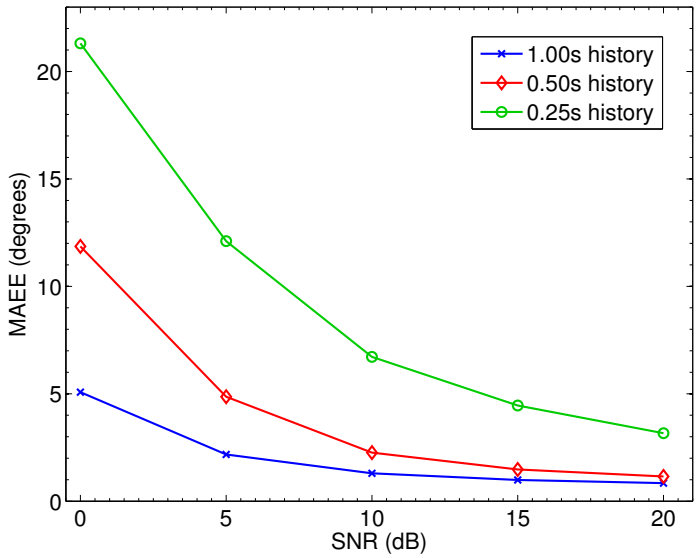


Figure 4.7: DOA estimation error vs SNR for four intermittant speakers in a simulated reverberant environment.

lations that included moving sources. In Figure 4.10 one speaker is static at 90° and the other is moving clockwise. Both speakers were males. In Figure 4.11 two male speakers are moving in a circular fashion around the array. One of them is moving anticlockwise while the other is moving clockwise. We observe a consistent DOA estimation in both scenarios, even though we do not use any source labeling techniques. This preliminary simulation results, along with their real-environment experiments counterparts (Figs. 4.17 and 4.18), indicate that the proposed framework could be extended to include tracking capabilities. The slight shift of the estimations to the right of the true DOA is due to the one-second history length. Anomalies in the DOA estimation are mainly present around the crossing points, which was expected, since the effective resolution of the proposed method is around 30° (see also Figure 4.5).

Comparison with alternative methods

We compared the performance of the DRACOSS framework against MUSIC (Section 2.1.1), and ICA-GSCT (Section 2.2.1). We also compared against the WDO assumption, i.e., we replaced the MCC at step 2 of DRACOSS with the WDO assumption (Section 2.4). The other three steps of the framework stay the same. The performance of the methods was evaluated by using the MAEE over those estimates where the absolute error was found to be lower than 10°—where an estimate is considered to be successful. Along with the MAEE,

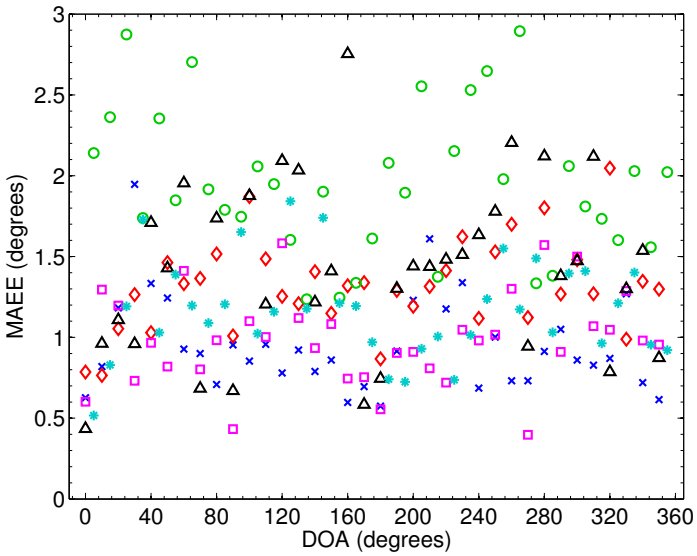


Figure 4.8: DOA estimation error of six static sources versus the true DOA. Different markers correspond to different speakers.

we provide "success scores", i.e., percentages of estimates where the absolute error was lower than 10° (Table 4.2 to be discussed later). Since the error was very high for plenty of estimates especially at lower SNR values for some of the methods, the MAEE over all estimates was considerably affected, not allowing us to have a clear image of the performance. Additionally for some of the methods, e.g., MUSIC, there were cases where the number of detected peaks, and consequenlty DOAs was lower than the number of sources, thus such cases had to be excluded from the evaluation. Furthermore, in a real system, a stable consistent behaviour—which is reflected in the "success scores"— is equally important as accuracy and computational complexity. We could extract similar observations on the consistency of the bahavior of the methods by providing the variances of the estimates. Hoewever, the "success scores" cover also the case where the DOA of a source could not be estimated, thus we believe it is a more appropriate measure in our case. We note that a similar method of performance evaluation was adopted in [16].

In Figure 4.12 we plot the MAEE versus the SNR for six static, continuously active speakers, originally located at 0°, 60°, 105°, 180°, 250°, and 315° in a simulated reverberant environment with a one-second block size. The simulation was performed for each orientation of sources in 10° steps around the array. All four methods exhibit very good results, with an increasing performance from lower to higher SNR values. Even though the differences are small between the methods, we note that DRACOSS exhibits the lowest MAEE for SNR values below 15 dB (and the highest success scores, shown in Table 4.2 to be dis-

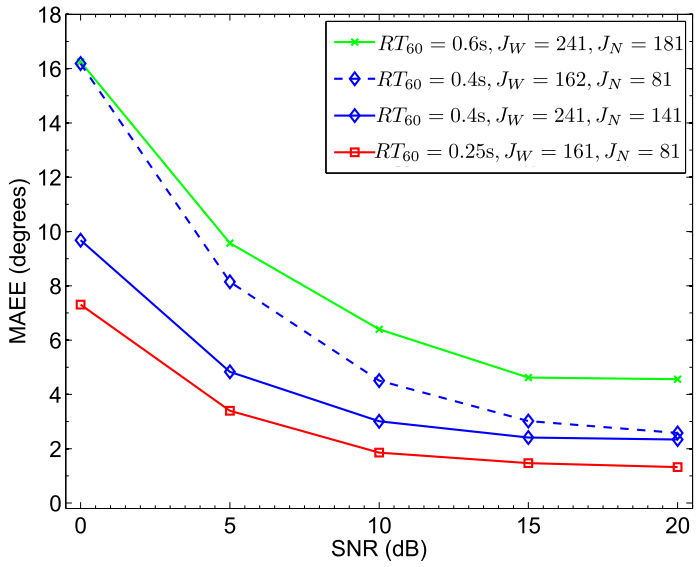


Figure 4.9: DOA estimation error vs SNR for three static, continuously active speakers in a simulated environment for $\mathrm{RT}_{60} = \{0.25, 0.4, 0.6\}$ s.

cussed later). We remind here that the performance of the methods was evaluated by using the MAEE over those estimates where the absolute error was found to be lower than 10°. Thus, the results of Figure 4.12 should be observed and evaluated together with the results in Table 4.2.

For the estimation of the demixing matrices in ICA-GSCT we have used the Joint Approximate Diagonalization of Eigenmatrices (JADE) method [18] which exploits the fourth-order cumulants relying on the statistical independence of the sources. However since the accuracy of the estimation of the demixing matrices (and consequently of the corresponding mixing matrices) for ICA-GSCT at each frequency bin depends on the sufficiency of the observed data—i.e., the block size—we ran the preceding simulation scenario using mixing matrices obtained also with the recursively regularized ICA (RR-ICA) algorithm [88]. The RR-ICA algorithm exploits the consistency of demixing matrices across frequencies and the continuity of the time activity of the sources and recursively regularizes ICA. In this way, it provides improved estimates of the demixing matrices even when a short amount of data is used. We note that the code for RR-ICA is provided by the authors of [88] and can be found in [84]. The maximum number of ICA iterations was set to 20 and the natural gradient step-size to 0.1. The maximum order of the least mean square filter was set to 10 and the corresponding step size to 0.01. These values gave the best results among various parametrizations and are in the range of values recommended in [88].

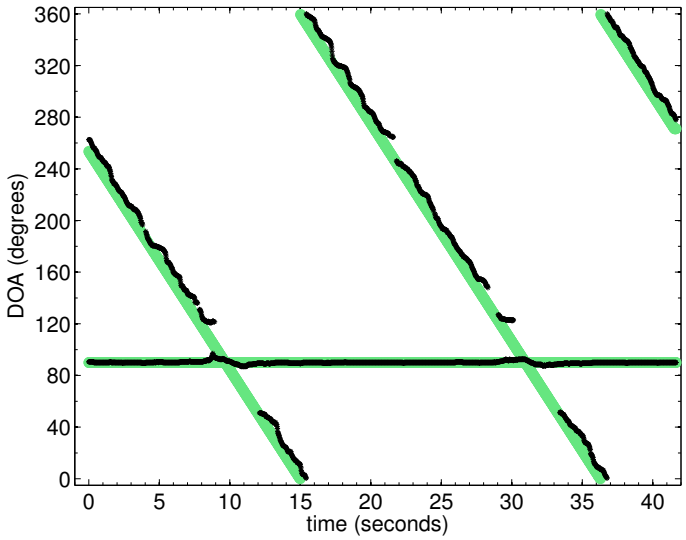


Figure 4.10: Estimated DOA of one static and one moving speaker around the circular array in a simulated reverberant environment at 20 dB SNR.

In Figure 4.13 we compare the performance of ICA-GSCT using these two different methods for the estimation of the mixing matrices, i.e., the JADE algorithm and RR-ICA method. We observe that both methods exhibit good and similar results for all SNR values. We note that RR-ICA performs slightly better for SNR higher than 5 dB as was expected but did not provide a significant improvement compared to JADE for our particular simulation scenario.

Table 4.2: DOA estimation success scores

	SNR(dB)				
Method	0	5	10	15	20
DRACOSS	61.62%	84.07%	95.45%	99.16%	99.69%
WDO	54.96%	80.38%	95.40%	99.57%	99.94%
MUSIC	47.89%	64.82%	77.34%	92.58%	99.89%
JADE ICA-GSCT	55.44%	68.66%	80.38%	89.17%	93.90%
RR-ICA GSCT	40.66%	57.69%	73.70%	88.04%	96.48%

In Table 4.2 we provide success scores (percentages of frames with absolute error $< 10^{\circ}$) for the proposed and all aforementioned methods. We observe that for an SNR of 20 dB, all methods successfully estimate the DOAs for more than 90% out of a total amount

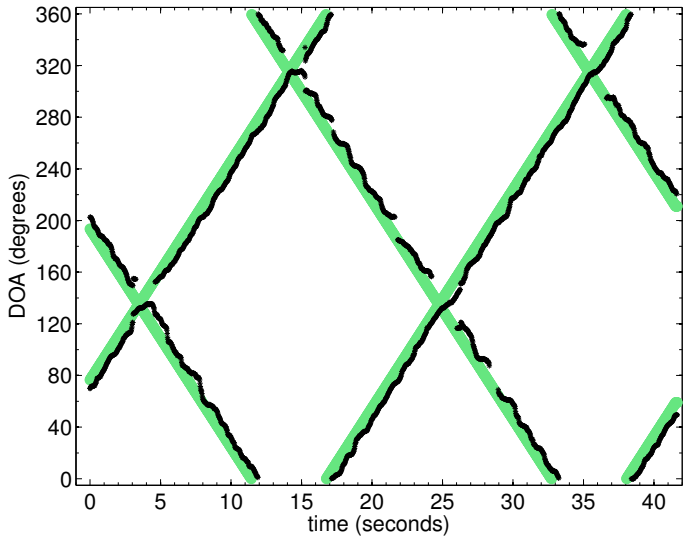


Figure 4.11: Estimated DOA of two moving speakers around the circular array in a simulated reverberant environment at 20 dB SNR.

of approximately 83,000 estimates. Specifically, DRACOSS along with WDO and MUSIC almost achieve score of 100%, with the proposed framework being much more efficient in terms of complexity (see also Table 4.3). When the SNR gets lower, the performance of the methods deteriorates, which can also be observed in Figures 4.12 and 4.13. However, our proposed method's score is higher than the other methods for SNR values below 15 dB.

We present complexity estimation results for the preceding scenario with six sources in Table 4.3. We estimated the total number of operations that each method performs to derive a curve whose local maxima act as DOA indicators. More specifically, we estimated the total number of the following operations: for DRACOSS and WDO, to obtain the smoothed version of the histogram of the estimates; for MUSIC, to estimate the average pseudospectrum; and for ICA-GST, to estimate the GSCT-kernel density function at each time instant. By the term "operation", we refer to any multiplication, addition or comparison, as many dedicated processors—such as DSPs—only take one cycle for each of these operations.

Our framework has the lowest computational complexity. MUSIC requires almost one and a half times as many operations, while WDO needs almost three times as many operations. The complexity of ICA-GSCT is much higher than all the other methods. These results were expected, since WDO follows the same procedure as the proposed method, but for all the frequency components whereas we work with d components in single-source zones only. On the other hand, MUSIC performs eigenvalue decomposition for each frequency component and averages the information from all frequency components, con-

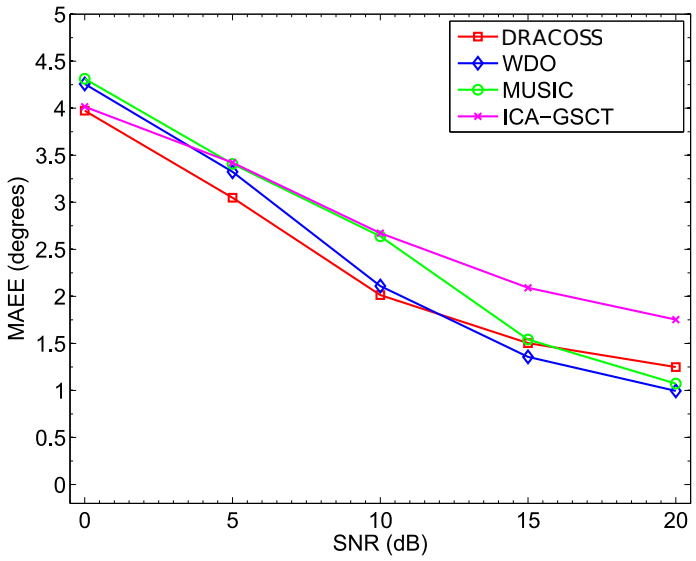


Figure 4.12: DOA estimation error vs SNR for six static speakers in a simulated reverberant environment.

Table 4.3: Computational complexity

Method	number of operations
DRACOSS	2,638,424
WDO	10,235,565
MUSIC	3,903,280
ICA-GSCT	35,254,348

tributing significantly to its high complexity.

Source counting results

In order to evaluate the performance of DRACOSS in counting the number of sources (see Section 4.1.4), we provide source counting results for simulation scenarios ranging from one to six static, simultaneously active sound sources in a reverberant environment with an SNR of 20 dB. In these six simulation scenarios, the smallest angular distance between sound sources was 45° and the highest was 180° while the sources were active for approximately 10 seconds, leading to roughly 14,000 source number estimations for each scenario. The thresholds vector was set to $\gamma = [0.15, 0.14, 0.12, 0.1, 0.065, 0.065, 0.065]$ and the

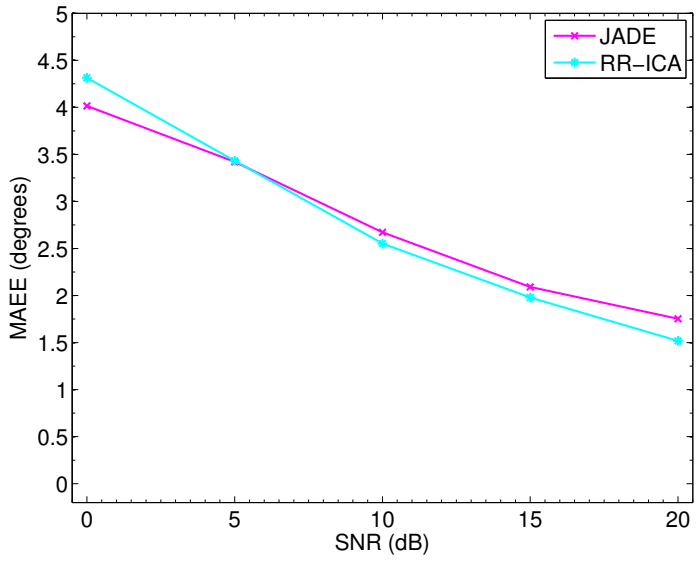


Figure 4.13: DOA estimation error vs SNR for six static speakers in a simulated reverberant environment.

minimum offset between neighboring located sources was set to $n_w = 10°$. We present these results in terms of a confusion matrix in Table 4.4 where the rows correspond to true numbers of sources and the columns correspond to the estimated ones. The method correctly estimates the number of sources more than 87% of the time for all the cases. Overall the method presents very good performance with a mean percentage of success equal to 93.52%.

Table 4.4: Confusion matrix for counting success scores

		$\hat{N}_S$						
		1	2	3	4	5	6	7
N_S	1	100%	0%	0%	0%	0%	0%	0%
	2	0%	100%	0%	0%	0%	0%	0%
	3	0%	3.76%	96.16%	0.08%	0%	0%	0%
	4	0%	0.42%	8.50%	88.84%	2.20%	0.04%	0%
	5	0.01%	2.23%	2.99%	0.55%	88.28%	5.76%	0.18%
	6	0.87%	2.91%	1.42%	0.17%	5.91%	87.84%	0.88%

We compared our DRACOSS framework with additional proposed source counting methods (see Appendix D and [91]) and the minimum description length (MDL) information

criterion [120] under the four intermittent speakers scenario, an example of which can be seen in Figure 4.6. For the Peak Search method (PS), $\gamma_{\text{static}} = 0.05 \sum_b h(b)$ and the LPC order used was 16. The thresholds for DRACOSS were $\gamma = [0.15, 0.14, 0.12, 0.1]$. The minimum offset between neighboring located sources was set to $n_w = 10^\circ$ and was common for all these histogram-based methods. The MDL was estimated in the frequency domain from the STFT of the observations in blocks of N_T frames. In Table 4.5 we give success rates of the source counting (percentage of frames correctly counting the number of sources) for the four methods under consideration with various history lengths and differing values of SNR. The success rates were again calculated over all orientations of the sources in 10° steps around the array (preserving the angular separations) while the transition periods were not taken into account.

We can observe similar behavior as in Figure 4.7. Longer history length leads to increased success rates for all four methods, affecting however, the responsiveness of the system. The MDL method is severely affected by noise and the amount of available data. While it achieves a high percentage of success for one-second history length and 20 dB SNR, this percentage falls dramatically as the history length is reduced and most obviously as the SNR becomes lower. For SNRs equal to 0 and 5 dB the criterion fails completely since it always responds as if there are no active sources. The DRACOSS framework is clearly the best performing source counting method. Moreover, in DRACOSS the DOA estimation and the source counting are performed in a single step (as explained in Section 4.1.4), resulting in computational efficiency.

Table 4.5: Source counting success rates excluding transition periods

Method	History Length	SNR (dB) 0	5	10	15	20
MDL	0.25s	0%	0%	2.3%	15.7%	21.6%
PS	0.25s	34.7%	44.8%	60.2%	71.5%	79.1%
LPC	0.25s	25.7%	40.5%	57.0%	63.0%	64.6%
DRACOSS	0.25s	42.9%	61.5%	77.8%	84.7%	86.7%
MDL	0.5s	0%	0%	6.8%	38.8%	74.8%
PS	0.5s	44.5%	60.1%	77.5%	84.9%	88.2%
LPC	0.5s	35.5%	59.5%	73.8%	75.6%	74.2%
DRACOSS	0.5s	64.3%	84.8%	95.7%	96.7%	96.7%
MDL	1s	0%	0%	21.2%	70.8%	87.7%
PS	1s	47.3%	68.7%	83.6%	90.5%	92.7%
LPC	1s	45.4%	81.9%	85.4%	82.5%	80.1%
DRACOSS	1s	82.1%	99.2%	100%	100.0%	100.0%

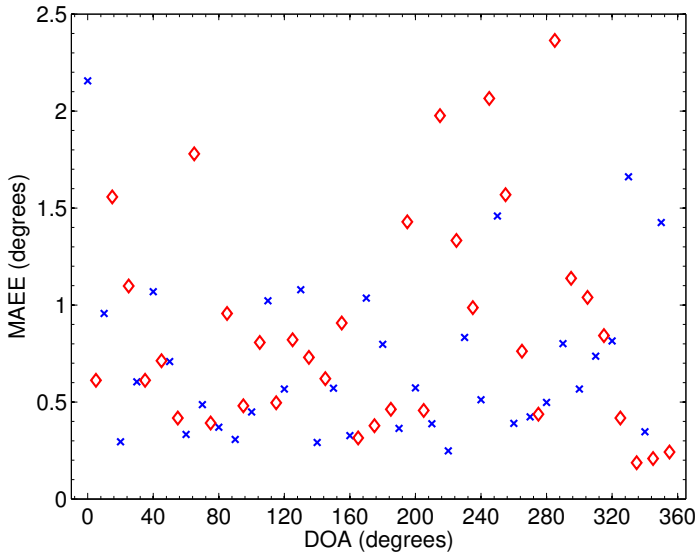


Figure 4.14: DOA estimation error for two speakers separated by 45° versus the true DOA in a real environment. Each different marker corresponds to a different speaker.

4.2.2 Real Environment

We conducted experiments in a typical office room with approximately the same dimensions and placement of the microphone array as in the simulations and with reverberation time approximately equal to 0.4 s. The algorithm was implemented in software executed on a standard PC (Intel 2.40 GHz Core 2 CPU, 2GB RAM). We used eight Shure SM93 microphones (omnidirectional) with a TASCAM US2000 8-channel USB soundcard. We measured the execution time and found it to be 55% real time (i.e., 55% of the available processing time). In the following results, some percentage of the estimated error can be attributed to the inaccuracy of the source positions.

We demonstrate the performance of our DRACOSS system for two simultaneously active male speakers in Figure 4.14. The speakers were separated by 45° and they moved 10° in each experiment in order to test the performance all around the array. The duration of each experiment was approximately six seconds. The signal to noise ratio in the room was, on average, 15 dB. We plot the MAEE versus each different DOA, where the MAEE is evaluated as the mean absolute error in the estimation over time. The mean absolute error is lower that 2.5° for every positioning of the speakers around the array (among 36 different orientations) while for about half of the orientations, the MAEE is below 1° for both speakers.

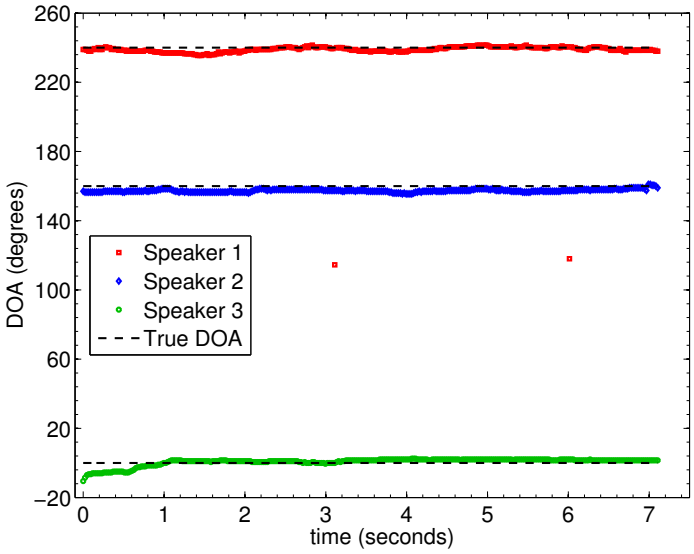


Figure 4.15: Estimated DOA of three static speakers in a real environment.

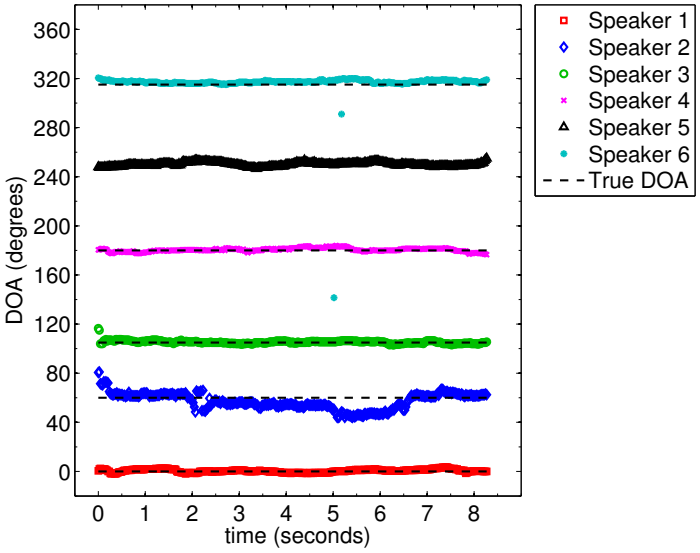


Figure 4.16: Estimated DOA of six static speakers in a real environment.

The next experiment involved three speakers sitting around the microphone array at 0°, 160°, and 240°. The speakers at 0° and 240° were males, while the speaker at 160° was female. The signal to noise ratio in the room was also around 15 dB. In Figure 4.15 we plot the estimated DOA in time. All three speakers are accurately located through the whole duration of the experiment.

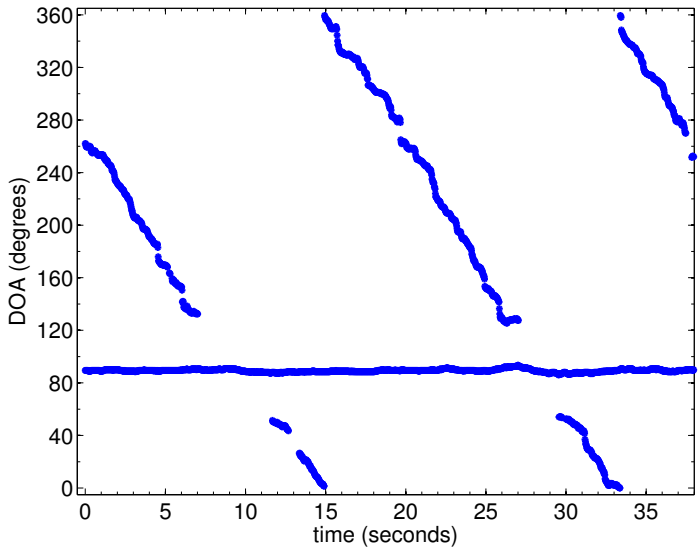


Figure 4.17: Estimated DOA of one static speaker and one moving speaker around the circular array in a real environment.

In Figure 4.16 we plot the estimated DOAs of six static speakers versus time. This experiment is the only one that involved loudspeakers instead of actual speakers. We used six Genelec 8050 loudspeakers that reproduced pre-recorded audio files of six continuously active, actual speakers, three males and three females positioned alternately. The loudspeakers were approximately located at $0°$, $60°$, $105°$, $180°$, $250°$, and $315°$ at a distance of 1.5 meters from the center of the array. The signal to noise ratio in the room was estimated at 25 dB. The DOA of all six sources is in general accurately estimated. The DOA estimation of the second speaker deviates slightly from the true DOA for some periods of time (e.g., around the sixth second of the experiment). This might be attributed to a lower energy of the signal of the particular speaker over these periods in comparison to the other speakers.

We also conducted experiments with moving sources. The scenarios followed the simulations (see Figures 4.10 and 4.11). However, since these are experiments in real conditions we do not show the ground-truth DOA of the sources, since this information cannot be exactly known. For these experiments, the signal to noise ratio in the room was, on average, 20 dB. We plot the DOA estimation in Figures 4.17 and 4.18. The DOA estimation is in general effective except for the areas around the crossing points. Nevertheless, as we stated for the corresponding simulations, our method shows the potential of localizing moving sources that cross each other.

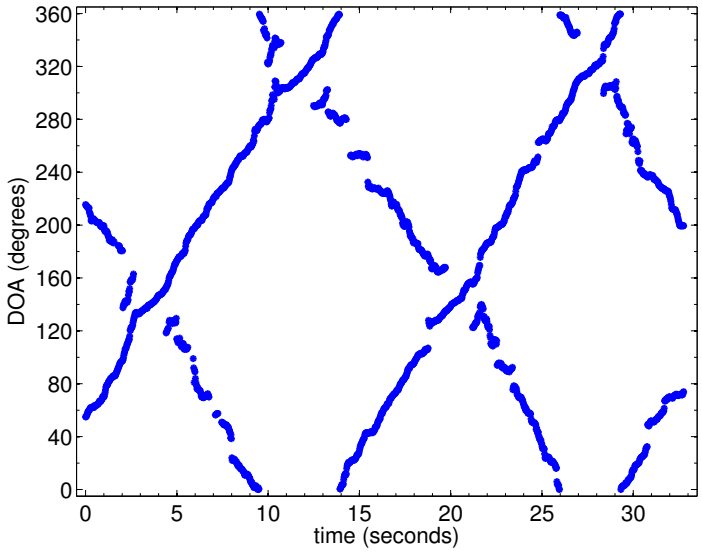


Figure 4.18: Estimated DOA of two moving speakers around the circular array in a real environment.

Chapter 5
Direction of arrival estimation in the three-dimensional space

This chapter presents multiple sound source localization and counting in the 3D space utilizing a compact spherical microphone array in the DRACOSS framework. Thus, as for the 2D DOA estimation, the proposed methods are based on detecting TF zones where one source is dominant over the others, i.e., SSZs. As a local DOA estimator for the second step of the DRACOSS framework, we use a sound intensity vector estimator, via the encoding of the signals of the spherical microphone array from the space domain to the spherical harmonic domain. A smoothed 2D histogram (the third DRACOSS step) of these estimates reveals the DOA of the present sources and through an iterative process, accurate 3D DOA information can be obtained. Additionally we incorporate beamforming and the MUSIC algorithm into the DRACOSS framework and enhance their performance. We show promising counting results by training a convolutional neural network with 2D histograms.

5.1 DRACOSS in three dimensions

In this section we describe the development of the DRACOSS framework in the 3D space. We are now looking to estimate not only the azimuth but also the elevation of an emitting source. These two estimates together define the 2D DOA, $\Omega = (\theta, \varphi)$, of the source in the 3D space as shown graphically in Figure 5.1.

For the estimation of the DOA we will follow the steps of the DRACOSS framework as described in Chapter 3. As a first step, for the exploitation of the sparsity of the sound sources, we will again use the MCC criterion that we used in DRACOSS in 2D spaces (see Chapter 4, Section 4.1.1) which we will recall here for the sake of completeness. As a local DOA estimator (second step) we will use the intensity vector $\mathbf{I}(\tau, k)$, which indicates the direction of sound flow at a TF point (τ, k) where τ indicates the timeframe and k the frequency index respectively. For the third and fourth steps of DRACOSS we will formulate

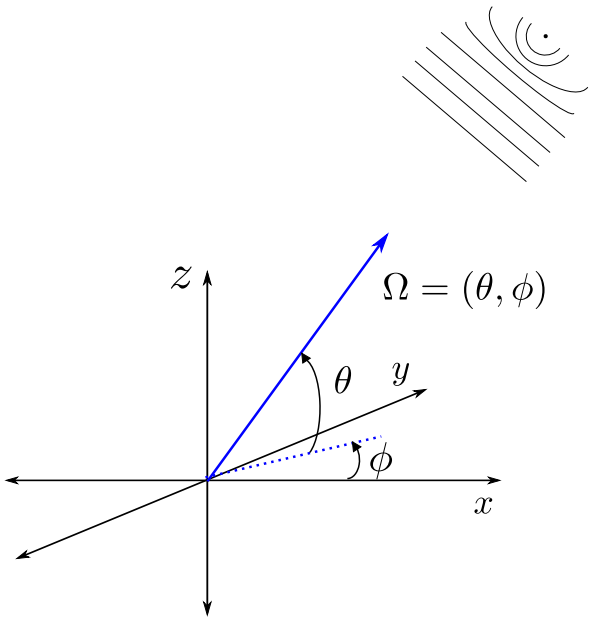


Figure 5.1: Direction of arrival of an emitting source in the 3D space, $\Omega = (\theta, \varphi)$, where $\theta \in [-\pi/2, \pi/2]$ denotes the elevation and $\varphi \in [0, 2\pi)$ denotes the azimuth.

and post-process 2D histograms as described later in Sections 5.1.3 and 5.1.4.

The intensity vector is formulated in the spherical harmonic domain (SHD), thus we will now provide an overview of the process of how to spatially encode the microphone array sensor signals to a set of spherical harmonic signals for the sake of completeness. For an extended overview of this process the reader is referred to [98].

Let us assume a microphone array that consists of Q microphones positioned at $\left(\Omega_q, r\right) = (\theta_q, \varphi_q, \mathrm{r})$. The spherical harmonic signals are approximated as:

$$s_{lm}(k, \mathrm{r}) \approx \sum_{q=1}^{Q} g_q(\Omega_q) x_q(k, \Omega_q, \mathrm{r}) Y_{lm}^*(\Omega_q), \tag{5.1}$$

where $x_q(k, \Omega_q, \mathrm{r})$ are the separate microphone signals for frequency k, $Y_{lm}^*(\Omega_q)$ are the complex conjugate spherical harmonic functions, and $g_q(\Omega_q)$ is selected so that it provides an accurate approximation of the spherical Fourier transform [114]. The accuracy of this approximation depends on how uniformly the microphones are distributed on the surface of the sphere, the type of the array, the radius r and the frequency k [96]. By omitting the frequency and radial dependency, the equalized spherical harmonic signals can be expressed in matrix form as:

$$\mathbf{s} \approx g_q \mathbf{B}^{-1} \mathbf{Y}^H \mathbf{x}, \tag{5.2}$$

where $(^H)$ denotes Hermitian transposition,

$$\mathbf{x} = [x_1, x_2, \ldots, x_Q]^T \in \mathbb{C}^{Q\times 1} \tag{5.3}$$

are the microphone array input signals,

$$\mathbf{s} = \quad [s_{00}, s_{1-1}, s_{10}, s_{11}, \ldots, s_{LL}]^T \in \mathbb{C}^{(L+1)^2\times 1} \tag{5.4}$$

are the spherical harmonic signals describing the decomposition of a soundfield comprised by plane waves (see also Appendix C.5), $g_q(\Omega_q) = \frac{4\pi}{Q}$, assuming a uniform distribution of microphones on the surface of a sphere (see also Appendix C.6.3),

$$\mathbf{B} = \quad \text{diag}\{[b_0, b_1, b_1, b_1, \ldots, b_L]\} \in \mathbb{C}^{(L+1)^2\times(L+1)^2} \tag{5.5}$$

is a diagonal matrix containing the equalization weights that depends on the array type, whether it is rigid or open, and is used in (5.2) to compensate for the effect of the microphone array [114]. $\mathbf{Y} \in \mathbb{C}^{Q\times(L+1)^2}$ is the matrix containing the spherical harmonics up to order L for the Q microphones

$$\mathbf{Y}(\Omega_q) = \begin{bmatrix} Y_{00}(\Omega_1) & Y_{00}(\Omega_2) & \ldots & Y_{00}(\Omega_q) \\ Y_{1-1}(\Omega_1) & Y_{1-1}(\Omega_2) & \ldots & Y_{1-1}(\Omega_q) \\ Y_{10}(\Omega_1) & Y_{10}(\Omega_2) & \ldots & Y_{10}(\Omega_q) \\ Y_{11}(\Omega_1) & Y_{11}(\Omega_2) & \ldots & Y_{11}(\Omega_q) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{LL}(\Omega_1) & Y_{LL}(\Omega_2) & \ldots & Y_{LL}(\Omega_q) \end{bmatrix}^T, \tag{5.6}$$

where $(^T)$ denotes transposition. The number of microphones to reconstruct L independent spherical harmonics signals is $Q \geq (L+1)^2$ [1].

5.1.1 Step 1: sound sources sparsity

We now recall the estimation of the MCC criterion for the detection of single source zones (SSZs) in the TF domain of the microphone array signals. A SSZ is a series of K frequency-adjacent TF points (τ, K) where one source is dominant over any other active source and satisfies the following criterion:

$$\overline{\rho}(\tau, K) \geq 1 - \epsilon, \tag{5.7}$$

where $\overline{\rho}(\tau, K)$ is the average correlation coefficient between pairs of observations of adjacent microphones and ϵ is a small user-defined threshold.

The correlation coefficient $\rho_{i,j}(\tau, K)$ is defined as

$$\rho_{i,j}(\tau, K) = \frac{R_{i,j}(\tau, K)}{\sqrt{R_{i,i}(\tau, K) \cdot R_{j,j}(\tau, K)}}, \tag{5.8}$$

where $R_{i,j}(\tau, K) = \sum_{k \in K} \left| X_i(\tau, k) \cdot X_j(\tau, k) \right|$ is the cross-correlation of the magnitude of the TF transform over an analysis zone for any pair of signals (x_i, x_j). Thus, SSZ detection takes place in the TF domain. $X_i(\tau, k)$ and $X_j(\tau, k)$ are the microphone signals of the i^{th} and the j^{th} microphones respectively in the TF domain. Note that $x_q(k, r, \Omega)$ in Equation (5.1) is now expressed in the TF domain as $X_q(k, n)$ for the q^{th} microphone by omitting the (r, Ω) parameters. The reader is also referred to Section 4.1.1 for a more detailed description of the definition and detection of SSZs.

5.1.2 Step 2: single-source local DOA estimator

The local DOA estimator is based on the sound intensity [37] and it has been utilized in parametric sound reproduction systems [95]. As in [57] (see also Section 2.3.2), the instantaneous active intensity vector can be approximated in the TF domain as

$$\mathbf{I}(\tau, k) = \frac{1}{2} \mathrm{Re} \left\{ \left[\frac{s^*_{00}(\tau, k)}{b_0(k)} \right] \begin{bmatrix} s_x(\tau, k) \\ s_y(\tau, k) \\ s_z(\tau, k) \end{bmatrix} \right\}, \tag{5.9}$$

where s^*_{00} is the complex conjugate of the 0^{th} order spherical harmonic signal, $b_0(k)$ is the mode strength compensation and s_x, s_y and s_z are averages of the 1^{st} order spherical harmonic signals with their positive phase towards the x, y and z-axis respectively. Each of these signals is calculated as

$$s_a(\tau, k) = \sum_{m=-l}^{l} \frac{Y_{lm}(\Omega_a)}{b_l(k)} s_{lm}(\tau, k), \quad a = \{x, y, z\}, \tag{5.10}$$

where $Y_{lm}(\Omega_a)$ is the spherical harmonic base function of order $l = 1$ and degree m, Ω_a is set to $(0, 0)$, $(0, \pi/2)$ and $(\pi/2, 0)$ for each axis and $b_l(k)$ is the mode strength compensation for the specific order and depends on the type of the array [114]. We graphically show the intensity vector and the corresponding DOA for an emitting source in Figure 5.2.

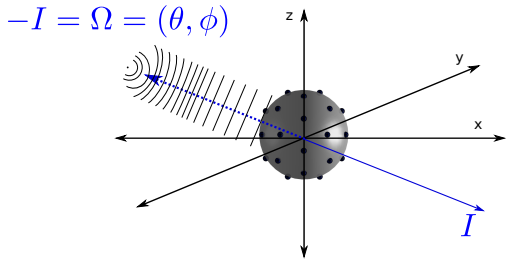


Figure 5.2: Vector $\mathbf{I}$ indicates the direction of sound flow, thus the DOA of the emmiting source is $-\mathbf{I} = \Omega = (\theta, \varphi)$, where $\theta \in [-\pi/2, \pi/2]$ denotes the elevation and $\varphi \in [0, 2\pi)$ denotes the azimuth.

5.1.3 Step 3: histogram formation

After the detection of a SSZ, we estimate the intensity vector $\mathbf{I}(\tau, k)$ at d selected frequency components belonging to the SSZ, i.e., those TF points that correspond to the indices of the d highest peaks of the magnitude of the cross-power spectrum over all microphone pairs. In this manner we have d DOA estimates at each detected SSZ .

Once we have estimated all the local DOAs in the SSZs (Sections 5.1.1 & 5.1.2), we form a 2D histogram, $\mathbf{h}(i, j)$, from the set of estimations in a block of N_T consecutive time frames which slides one frame each time. An example of such a histogram is shown in Figure 5.3. We smooth the 2D histogram by applying an averaging filter, e.g., a circularly symmetric Gaussian window $\mathbf{w}_{\mathrm{A}}(\theta, \varphi)$ of zero mean and standard deviation (std) equal to σ_{A}.

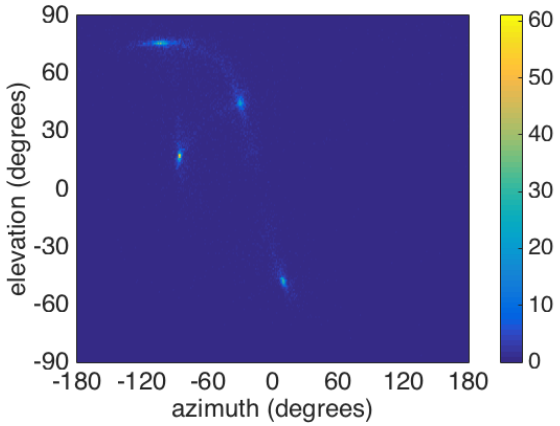


Figure 5.3: 2D histogram of four sources at $(43, -31)^\circ$, $(-48, 9)^\circ$, $(75, -104)^\circ$, and $(16, -86)^\circ$.

$$\mathbf{h}_{\mathrm{s}}(\theta,\varphi)=\sum_{i}\sum_{j}\mathbf{h}(i,j)\mathbf{w}_{\mathrm{A}}(\theta-j,\varphi-i), \tag{5.11}$$

where $\mathbf{w}(\theta,\varphi)=\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}\frac{\theta^2+\varphi^2}{\sigma^2}}$ is the Gaussian window, $\mathbf{h}(\theta,\varphi)$ is the original 2D histogram and $\mathbf{h}_{\mathrm{s}}(\theta,\varphi)$ is the smoothed one. The smoothed version of the histogram in Figure 5.3 is depicted in Figure 5.4.

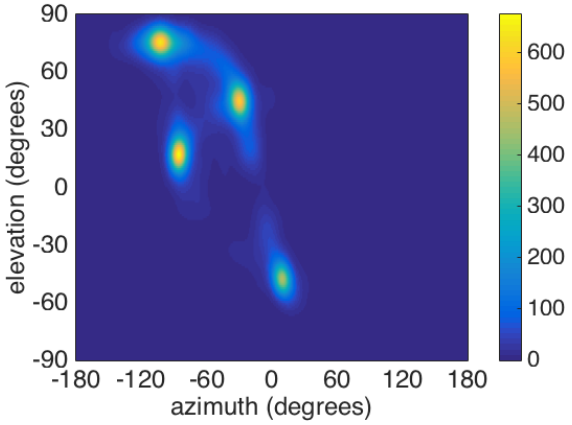


Figure 5.4: Smoothed 2D histogram of four sources at $(43,-31)^\circ$, $(-48,9)^\circ$, $(75,-104)^\circ$, and $(16,-86)^\circ$.

5.1.4 Step 4: histogram post-processing

In order to extract the final DOA estimates of the sources, we proceed further by processing the 2D smoothed histogram. We detect the highest peak of the smoothed histogram and we identify its index as the DOA of the first source. Then, we remove its contribution from the histogram by applying a Gaussian window $\mathbf{w}_{\mathrm{C}}(\theta,\varphi)$ of zero mean and std equal to σ_{C}. We proceed to the detection of the second peak and the removal of its contribution and iteratively to the next peak until we reach the number N_S of sources. The steps of the aforementioned iterative procedure are described in detail in Algorithm 1 and in Figure 5.5 we show an example for a scenario with four sources at $(54,82)^\circ$, $(43,118)^\circ$, $(-58,307)^\circ$, and $(-22,172)^\circ$ at RT_{60}=0.2 s and 45 dB SNR of additive white Gaussian noise.

We note here that we process the 2D-histograms in the same spirit as we did for the single-dimensional histograms in Chapter 4.1.4. Our approach is more of a practical, intuitive, yet efficient nature. One could choose and adopt a probabilistic framework, such as that of GMM, however we would like to avoid any approach that demands a priori the

Algorithm 1 2D Histogram Processing for 3D DOA estimation

1. Set the loop index $i = 1$.

2. Find $(\theta_i, \varphi_i) = \arg\max_{\theta,\varphi} \mathbf{h}_s^i(\theta, \varphi)$, where $\mathbf{h}_s^i(\theta, \varphi)$ is the smoothed histogram at the current iteration. The DOA of this source is (θ_i, φ_i).

3. Calculate the contribution of the current source as

$$\delta_i = \mathbf{y}_s(\theta, \varphi) \odot \mathbf{w}_C(\theta - \theta_i, \varphi - \varphi_i),$$

   where the operator $\odot$ stands for element-wise multiplication.

4. Remove the contribution of this source as

$$\mathbf{y}_s^{i+1}(\theta, \varphi) = \mathbf{y}_s^i(\theta, \varphi) - \delta_i.$$

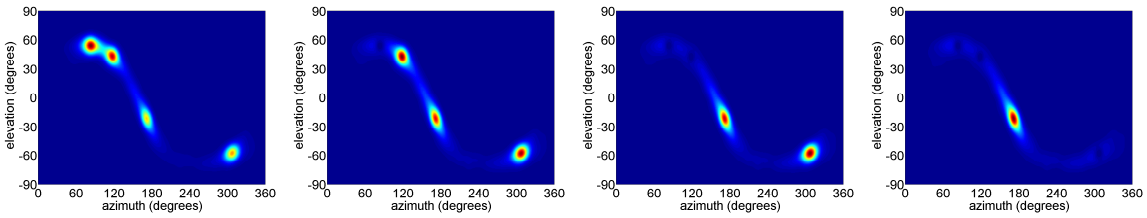
5. Increment i .

6. If $i < N_S$ go to step 2.

number of clusters, in our case the number of simultaneously active sources.

5.2 Evaluation

The performance of the DRACOSS framework in 3D spaces is investigated by conducting extended simulations in anechoic and reverberant environments. We have employed a rigid spherical microphone array comprising 32 microphones, placed according to the angular positions of the 3D sphere covering problem solutions [23] and radius equal to $r_a = 0.042$ m. We used the spherical microphone array room impulse response generator by Jarrett et al [58] which is based on the image method of Allen and Berkley [4] to simulate a room of $8 \times 8 \times 6$ meters. The spherical array was placed in the centre of the room, coinciding with the origin of the x, y and z-axis and the simulated sound sources were placed 1.5 m away from the centre of the array. The speed of sound was $c = 343$ m/s while the frequency range used was 500-3800 Hz to avoid aliasing phenomena [82]. In



(a) (b) (c) (d)

Figure 5.5: Visualization of Algorithm 1:
5.5a The 2D histogram given as input to Algorithm 1. Four sources are clearly visible.
5.5b The 2D histogram after the first iteration. The contribution of the first detected source at $(54, 82)^\circ$ has been removed while the DOAs of the three remaining sources are highlighted.
5.5c The 2D histogram after the second iteration.
5.5d The 2D histogram after the third iteration where only the contribution the fourth source at $(-22, 172)^\circ$ is present.

each simulation the sound sources had equal power and the signal-to-noise ratio at each microphone was estimated as the ratio of the power of each source signal to the power of the noise signal. Any other parameters and their corresponding values can be found in Table 5.1.

Table 5.1: Simulation parameters

parameter	notation	value
sampling frequency	F_s	48000 Hz
frame size		2048 samples
overlapping in time		50%
FFT size		2048 samples
TF zones width	K	375 Hz
overlapping in frequency		50%
SSZ threshold	ϵ	0.2
frequency bins/SSZ	d	2
histogram bin size		$0.5^\circ \times 0.5^\circ$
averaging window std	σ_A	5°
localization window std	σ_C	20°

We demonstrate the performance of our framework by the mean estimation error (MEE) which measures the angular distance between a unit vector pointing at the true DOA ($\mathbf{v}$)

and a unit vector pointing at the estimated DOA ($\hat{\mathbf{v}}$) [57] over all sound sources, all positions and all the frames of the source signals. The error is defined as

$$\text{MEE} = \frac{1}{N_\text{O} N_\text{F} N_S} \sum_{o,f,s} \cos^{-1}\left(\mathbf{v}_{osf}^T \hat{\mathbf{v}}_{osf}\right), \tag{5.12}$$

where $\cos^{-1}\left(\mathbf{v}_{osf}^T \hat{\mathbf{v}}_{osf}\right)$ expresses the angular distance between the true DOA, $\mathbf{v}_{osf}$ of the s^{th} active source in the o^{th} positioning in the f^{th} frame and the estimated one, $\hat{\mathbf{v}}_{osf}$. The association between the true and the estimated DOA of a source is determined based on the permutation that leads to the minimum error, given the permutations between the true DOAs and the estimated ones. N_O is the total number of different positions of the sound sources around the array, i.e., the sound sources were placed in different and random orientations in each simulation and the total number of different positions is $N_\text{O} = 10$. N_F is the total number of frames after subtracting $N_T - 1$ frames of the initialization period and N_S is the total number of active sources, which is assumed to be known. In all the simulations speech sound files were used of duration approximately equal to 9 seconds, leading to $N_\text{F} = 375$ frames. Any gaps or silent periods were manually removed. The block size is equal to 1 second, i.e., $N_T = 46$ frames.

In our first set of simulations we investigated the spatial resolution of our proposed method, i.e., how close two sources can be while accurately estimating their DOA. Figure 5.6 shows the MEE against the angular separation of two continuously active sound sources, one male and one female speaker, for SNR=$\{15, 20, 45\}$ dB of additive white Gaussian noise and reverberation time $\text{RT}_{60} = \{0.4, 0.8\}$ s. The MEE is very low even when the sources are very close to each other, e.g., for angular separation equal to 20°, RT_{60}=0.8 s and SNR=15 dB, the MEE is equal to 5.74°. With increasing SNR and decreasing reverberation time the MEE is improved as expected and shown in Figure 5.6.

Aiming at highlighting the impact of the SSZ selection on the DOA estimation robustness, we compare the MCC criterion against the W-disjoint orthogonality (WDO) assumption [122], where each TF point is assumed to be dominated by a single source (see also Section 2.4). Therefore $\mathbf{I}(\tau, k)$ is estimated for every TF point in the frequency range of interest. Figure 5.7 shows the MEE versus the reverberation time, $\text{RT}_{60} = \{0, 0.2, 0.4, 0.6, 0.8\}$ s, where RT_{60}=0 s corresponds to the anechoic case at SNR=$\{15, 20, 45\}$ dB for scenarios with four simultaneously and continuously active sources, two male and two female speakers. The minimum angular separation between the sources was 19.94°. As it is shown in Figure 5.7 the proposed SSZ criterion experiences lower error for all SNR and RT_{60} conditions which was expected since the MCC selects only TF points belonging to single-source zones and avoids the use of TF points with spurious $\mathbf{I}(\tau, k)$ information.

We show the performance of the proposed framework when the number of simultaneously active sources is considered to be relatively high, i.e., $N_S = 6$ in Figure 5.8. Three

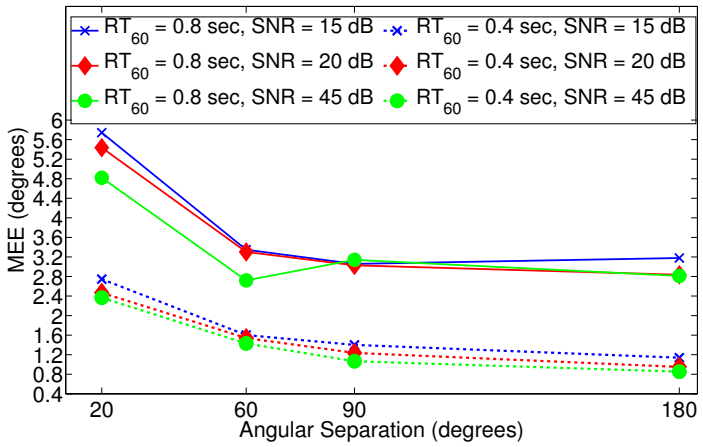


Figure 5.6: MEE versus angular separation between two sound sources in various SNR and reverberation conditions.

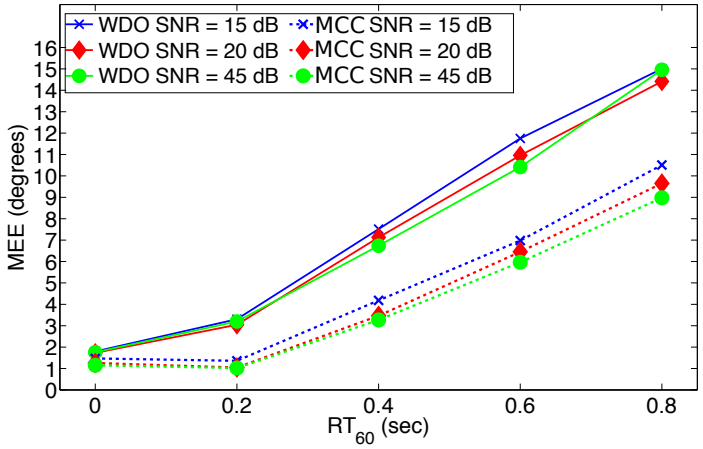


Figure 5.7: MEE versus RT_{60} for scenarios with four simultaneously active sound sources in various SNR conditions.

male and three female speech sound files were used with the minimum angular separation between them being at $21.14°$. Once again the MEE versus the reverberation time for various SNR conditions is shown. The MEE is higher compared to the four sources scenarios shown in Figure 5.7. The method performs robustly in moderate reverberation conditions, experiencing MEE equal to $19.57°$ in the worst case scenario, i.e., at $RT_{60} = 0.8\,s$ and SNR=15 dB.

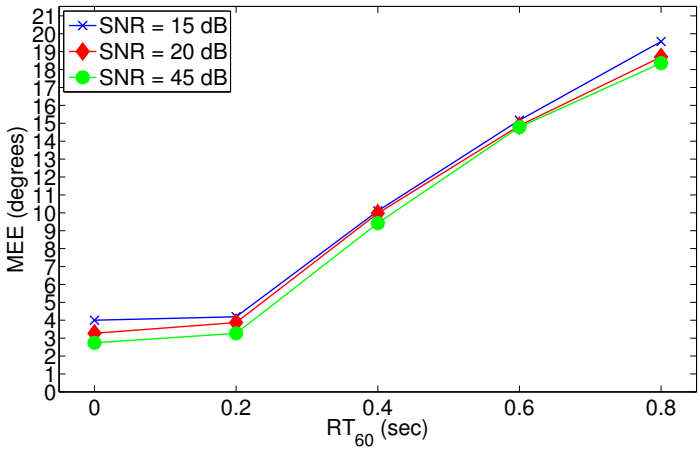


Figure 5.8: MEE versus RT_{60} for scenarios with six simultaneously active sound sources in various SNR conditions.

5.3 From intensity vector estimates to spatially constrained beamforming

Local DOA estimation through sound intensity vector possesses low computational complexity, since it can provide instantaneous time-frequency estimates. However, by definition, the intensity vector estimation exploits the spherical harmonic analysis of the sound field up to the first order (see also Eq. (5.9)), even though the available microphone array may provide higher orders for the analysis. On the other hand, DOA estimation relying on steered-response beamforming, even though it can exhibit high accuracy and exploits the full potential of the recording device, it suffers from high computational complexity due to the exhaustive search of the 3D space. These two different approaches motivated us to enhance the local DOA estimator of DRACOSS (step 2) by proposing a hybrid methodology that takes advantage of the simplicity of the intensity vector estimation and the accuracy of beamforming.

Assume that, although $-\mathbf{I}(\tau,k)$ might not point exactly to the DOA of a source, it will point towards the “neighborhood" of a true source, i.e., it aims near the true direction. We call this a coarse DOA estimation, $\Omega_c = (\theta_c, \varphi_c)$. We could then beamform around the area where $-\mathbf{I}(\tau,k)$ is pointing, i.e., perform spatially constrained beamforming (SCB), and thus obtain a refined DOA estimation. The beamforming is performed over the spherical sector defined by $-\mathbf{I}(\tau,k)$ and a vector of angle distance equal to γ_b from $-\mathbf{I}(\tau,k)$ (see also Fig. 5.9). The DOA, $\Omega_f = \left(\theta_f, \varphi_f\right)$, is then estimated as the index where the power of the SCB

gets maximized, i.e.,

$$\Omega_f = \arg\max_{\Omega_S} |p(\tau, k, \Omega_s)|^2, \tag{5.13}$$

where Ω_s belongs to the set of points in the spherical sector to be scanned, and $p(k, n, \Omega_s)$ is the beamformer's output for a regular beampattern steered at direction Ω_s [98] with $p(\tau, k, \Omega_s)$ given by

$$p(\tau, k, \Omega_s) = \mathbf{Y}^*(\Omega_s)\mathbf{s}. \tag{5.14}$$

$\mathbf{Y}^*(\Omega_s)$ and $\mathbf{s}$ are estimated using Eqs. (5.4) and (5.6).

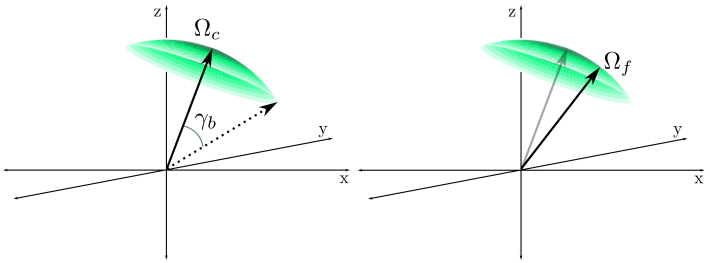


Figure 5.9: The gradient green spherical sector defines the beamforming area

Once we have estimated all the refined DOAs in the detected SSZs (first step of DRACOSS, see also Section 5.1.1), we proceed with the third and fourth step of the proposed framework as they were described in Sections 5.1.3 and 5.1.4.

5.3.1 Evaluation

We investigate the performance of the DRACOSS framework using the refined local DOA estimator by conducting extended simulations and real measurements in reverberant environments. For the simulations we used the spherical microphone array room impulse response generator by Jarrett et al [58] (as in Section 5.2) to simulate a room of $5.6 \times 6.3 \times 2.7$ meters in order to agree with the dimensions of the room where we conducted the real experiments. The angle for the spatially constrained beamforming was set to $\gamma_b = 10°$ and the spherical harmonic order was $L = 3$. The rest of the simulation parameters are the same as described in Table 5.1.

Results in a simulated environment

In our first set of simulations we investigate the performance of the refined local DOA estimator (denoted as "IVs+SCB" in the plots) for several angular distances between two continuously active sources for SNR={10, 15, 20} dB and $\mathrm{RT}_{60} = 0.4$ s in comparison with the local DOA estimator which relies solely on the intensity vector estimation (denoted as "IVs" in the plots). We show the results in Figure 5.11. In all examined cases, IVs+SCB

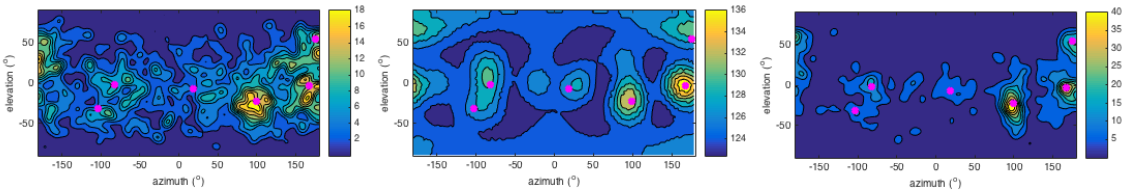


Figure 5.10: 2D histogram for six sound sources with the intensity vector (left), the corresponding pseudospectrum for the MUSIC subspace method with direct-path dominance test (middle) and the 2D histogram for intensity vector + SCB (right).

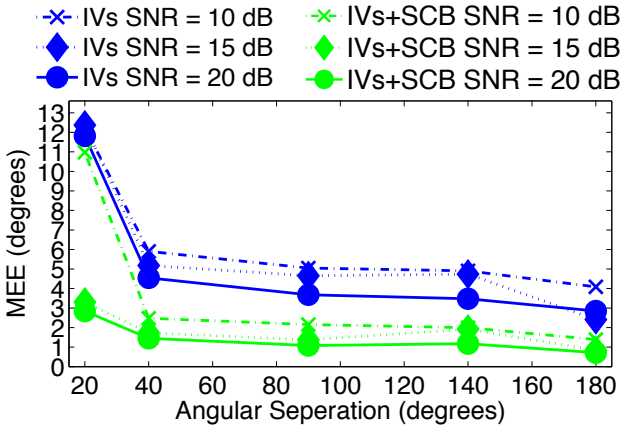


Figure 5.11: MEE versus angular separation between 2 sound sources for $RT_{60} = 0.4$ s and various SNR conditions.

exhibits better performance than the IVs one, for all SNR conditions and angular separations.

In our second set of simulations we compare the DRACOSS framework with the MUSIC algorithm as implemented in [82] and denoted as "DPD-MUSIC" (see also Section 2.1.1). In Figure 5.10 the 2D histograms for the IVs and the IVs+SCB methods and the pseudospectrum of the DPD-MUSIC are shown for a case of six simultaneously active sources in a simulated reverberant environment with $RT_{60} = 0.6$ s. The pink markers denote the true position of the sources. The processing of these representations of 2D estimates is based on one second history for all three methods assuming a known number of sources and follows the steps described in Section 5.1.4. Results in different acoustical conditions are shown in Fig. 5.12, for scenarios involving one, three and six simultaneously active speakers in highly reverberant conditions of $RT_{60} = 0.6$ s and SNR={10, 15, 20} dB. DPD-MUSIC

and IVs demonstrate similar performance while IVs+SCB exhibits a clear advantage especially for higher signal-to-noise ratios.

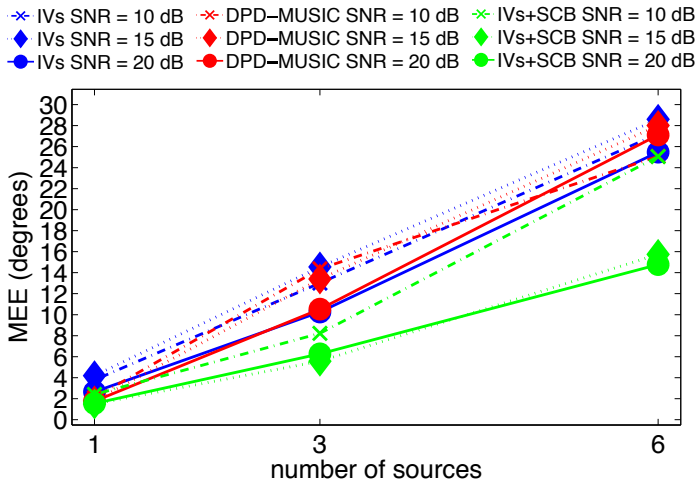


Figure 5.12: MEE versus number of sources for $RT_{60} = 0.6$ s and various SNR conditions.

Results in a real environment

For the conduction of real experiments we recorded room impulse responses using the EigenMike [77] spherical microphone array in a reverberant room of the same dimensions and reverberation time as in the simulations. We show our results in Figure 5.13 at the left plot, while at the right one we plot a simulated counterpart. DRACOSS with the refined DOA estimator, IVs+SCB, shows high accuracy for medium and higher SNR conditions even when six sources are simultaneously active. For lower SNR conditions and as the number of sources increases the performance degrades as it was expected, following similar tendency between the simulated and real results.

5.4 Beamforming in the DRACOSS framework

Steered-response power (SRP) methods for DOA estimation are based on scanning the sound field with a beamformer. The beamformer is steered in different directions of interest and the output power is then calculated. This signifies the SRP function which is utilized to identify the DOA of active sources as the indices of the local peaks of the SRP function. The results of the SRP can be enhanced by applying a phase transform [32]. SRP methods have been proposed for robust real-time applications using a coarse-to-fine search-grid contraction [21] or stochastic search-grid contraction [35].

Axis-symmetric beamforming in the spherical harmonic domain is performed by sim-

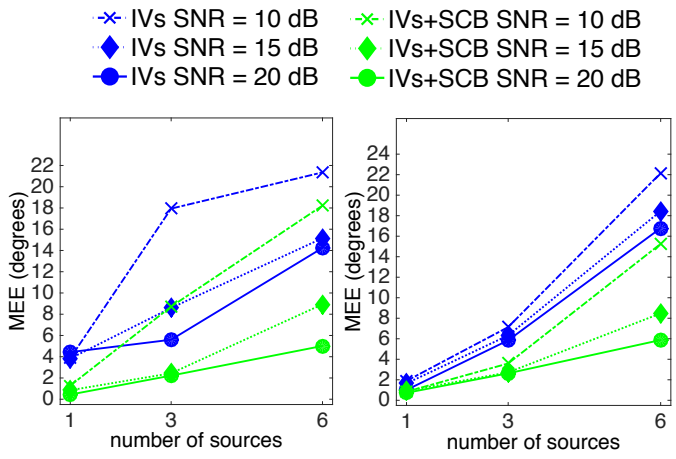


Figure 5.13: MEE versus number of sources for real RIRs in a room of $RT_{60} = 0.3$ s and various SNR conditions (left) and its simulated counterpart (right).

ply weighting and summing the spherical harmonic signals [97]. The single channel output, i.e., p, of the beamformer is

$$p(\Omega_s) = [\mathbf{y}(\Omega_l) \odot \mathbf{s}^T]\mathbf{d}, \tag{5.15}$$

where $\mathbf{y}(\Omega_l) \in \mathbb{C}^{1\times(L+1)^2}$ is a row of the spherical harmonics matrix (Eq. (5.6)), $\odot$ denotes the Hadamard product and $\mathbf{d}$ is a vector of weights defined as

$$\mathbf{d} = [d_0, d_1, d_1, d_1, \ldots, d_L, d_L, d_L]^T \in \mathbb{R}^{(L+1)^2\times 1}. \tag{5.16}$$

The power map of a beamformer can be provided by the output of the SRP function, which is defined as the energy of the beamformer in a grid of directions [57]

$$P(\Omega_l) = |p(\Omega_l)|^2, \tag{5.17}$$

where $\Omega_l = (\theta_l, \varphi_l)$ consists of all the elevation and azimuthal angles of the search grid. A power map is shown in Fig. 5.14, for six simultaneously active audio sources. The peaks in the power map indicate the DOAs of the audio sources.

Aiming at enhancing the presence of the DOAs in such 2D representations, we propose to first obtain single DOAs from power maps at each time-frequency point. That is, we include the beamforming into the DRACOSS framework by adopting the WDO assumption [122] as the first DRACOSS step and then we identify the highest peak of the power map as the DOA of the specific TF point which we define as a local DOA (DRACOSS second step). By collecting all DOAs from the TF points of interest, we proceed with steps 3 and 4, that is

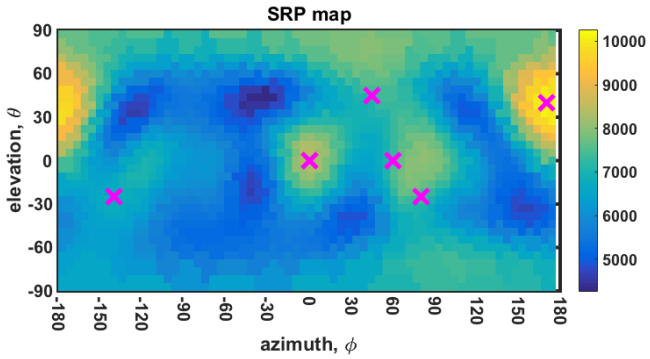


Figure 5.14: An SRP map snapshot of a scenario with six simultaneously active sources in a simulated environment of RT_{60}=0.3 s. The pink markers denote the actual positions of the audio sources.

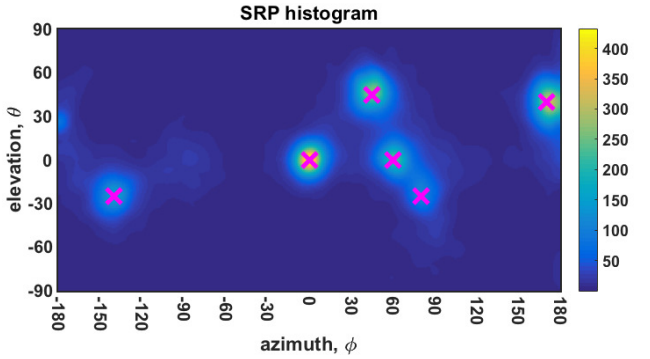


Figure 5.15: An SRP histogram for a scenario with six simultaneously active sources in a simulated environment of RT_{60}=0.3 s. The pink markers denote the actual positions of the audio sources.

we form 2D histograms which we process and infer the azimuthal and elevation angles of multiple sources, assuming their number to be known a priori. In Figure 5.15 we can see such a 2D histogram which corresponds to the same scenario as in Figure 5.14. Comparing the two representations, the advantage of the DRACOSS framework is already clear. We will show this positive effect reflected also at the evaluation results in Section 5.4.2. In the following section we present the formulation of four beamformers which we have used as local DOA estimators in the DRACOSS framework step 2.

5.4.1 Rotationally-symmetric beamformers

The types of axis-symmetric beamformers utilized in DRACOSS for local DOA estimation are:

- A *regular* beamformer, $\mathbf{d}_{\mathrm{r}}$, with unity gains [97]

$$d_{\mathrm{r}}(l) = 1, \ \forall l = [0, \ldots, L]. \tag{5.18}$$

- A *minimum-sidelobe* beamformer, $\mathbf{d}_{\mathrm{ms}}$, [26]. It smooths the sidelobes of the regular beamformer and provides complete sidelobe suppression with the cost of a wider main lobe. The weights are given by

$$d_{\mathrm{ms}}(l) = g_0 \frac{\Gamma(L+1)\Gamma(L+2)}{\Gamma(L+1+l)\Gamma(L+2+l)}, \tag{5.19}$$

  where Γ is the gamma function, and $g_0 = \sqrt{\frac{(2L+1)}{(L+1)^2}}$.

- A *maximum-energy* beamformer, $\mathbf{d}_{\mathrm{maxE}}$, that maximizes the energy concentration towards the look direction [14, 125]

$$d_{\mathrm{maxE}}(l) = C P_l(\mathrm{E}), \tag{5.20}$$

  where $P_l(\mathrm{E})$ is the l^{th} Legendre polynomial, E the largest root of P_{L+1} and C a normalization constant [27].

- A *Dolph-Chebyshev* beamformer

$$\mathbf{d}_{\mathrm{dc}} = \frac{2\pi}{R}\mathbf{PACT}\mathbf{x_0}, \tag{5.21}$$

  where $\mathbf{P}, \mathbf{A}, \mathbf{C}, \mathbf{T} \in \mathbb{R}^{(L+1)\times(L+1)}$, $\mathbf{x_0} \in \mathbb{R}^{(L+1)\times 1}$ and R are defined as in [97] in (6.66-6.71). The elements of the vector $\mathbf{d}_{\mathrm{dc}}$ are re-arranged so that they match (Eq. (5.16)), where the same weight is applied to the spherical harmonic signals of the same order.

The directivity patterns of the aforementioned beamformers are shown in Fig. 5.16. The input parameters of the Dolph-Chebyshev beamformer were: $R = 10^{\lambda/20}$ with $\lambda = 30$ dB.

5.4.2 Evaluation

The evaluation of DRACOSS with beamforming-based local DOA estimators is conducted with numerical simulation and real measurements in reverberant environments. We again use the MEE as a performance measure (see also Eq. (5.2)). The scanning area for the

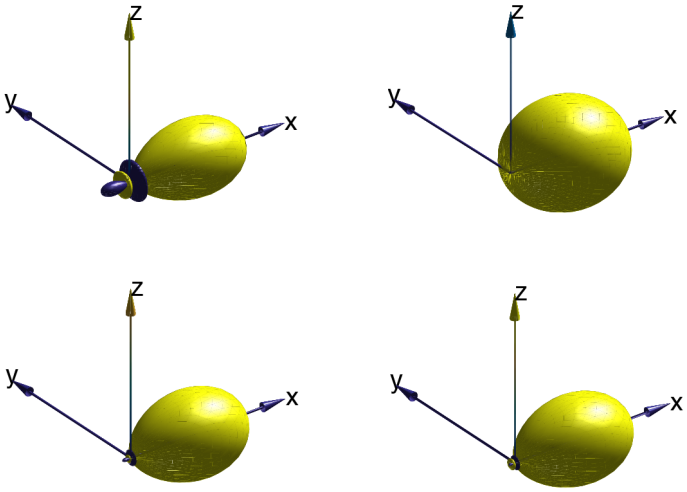


Figure 5.16: Directivity patterns of the axis-symmetric beamformers: regular (top left), minimum sidelobes (top right), maximum energy (bottom left) and Dolph-Chebyshev (bottom right).

beamformers comprises a set of 1002 points on a sphere. The distribution of the points is defined from a geodesic sphere constructed from an icosahedron with an iterative process [52]. The rest of the simulation and real experiments parameters are as those in Section 5.3.1. We note that the DOA estimation results for the four axis-symmetric beamformers (see also Section 5.4.1) that we present hereafter are solely from the proposed 2D DOA histograms processing. Due to the energy spread of the beamformers, obtaining multiple sources DOA estimates directly from the power maps (see Figure 5.14) had a significant error and was considered very inaccurate [36].

DOA results with simulated room impulse responses

In our first set of simulations, shown in Fig. 5.17, we plot the MEE versus the number of active sources in a simulated reverberant environment of RT_{60}=0.6 s for SNR={0, 10, 20} dB for the four different types of beamformers of Section 5.4.1. We notice that all four beamformers provide very good results for medium and high SNR conditions even when the number of active sources increases. However, their performance degrades as the SNR decreases with the minimum-sidelobe beamformer exhibiting the best performance.

DOA results with measured room impulse responses

The real measurements were performed by recording RIRs with the EigenMike [77] in a listening room with approximately the same dimensions as in the numerical simulations.

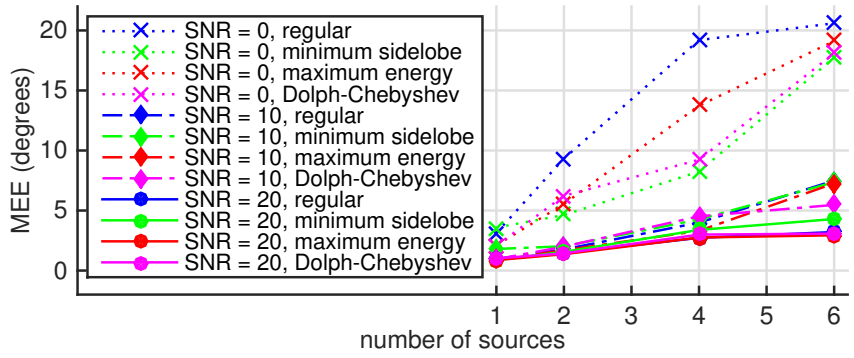


Figure 5.17: MEE versus number of sources for $RT_{60} = 0.6$ s and various SNR conditions for four types of axis-symmetric beamformers.

The reverberation time in the recording room was approximately equal to RT_{60}=0.3 sec. We show our first set of results in Figure 5.18 at the left plot, while at the right we plot a simulated counterpart. The SNR for both environments was at 45 dB. The performance of all beamformers is very good - the MEE is below three degrees in all cases - following similar tendency between the simulated and real results.

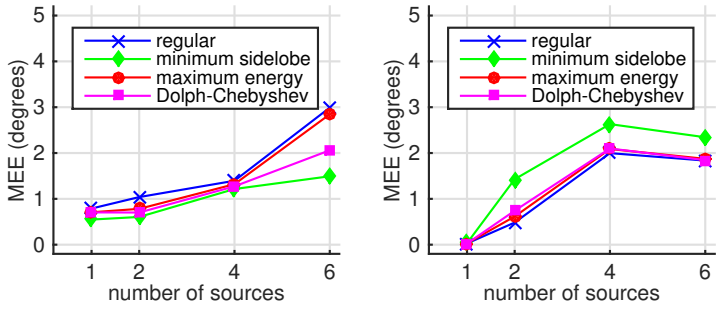
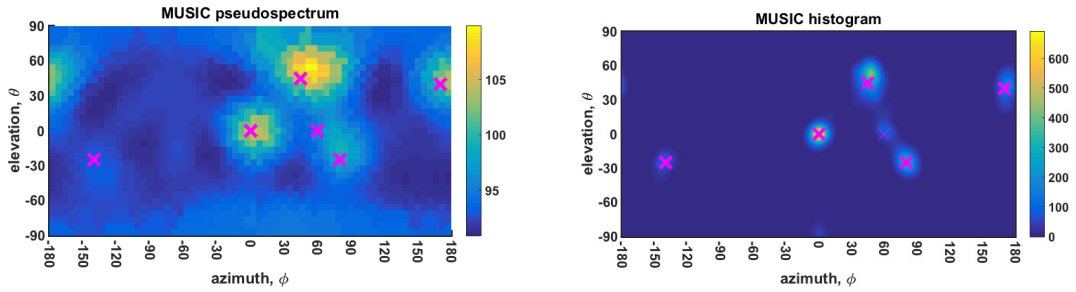


Figure 5.18: MEE versus number of sources for $RT_{60} = 0.3$ s for (a) real and (b) simulated measurements.

5.5 MUSIC in the DRACOSS framework

Apart from beamforming, a very popular algorithm for multiple source localization is the MUSIC algorithm, which has been recently formulated in the spherical harmonic domain [82] as we have previously mentioned in Sections 2.1.1 and 5.3.1 where we compared the DRACOSS framework with IVs and IVs+SCB as local DOA estimators with the MUSIC-DPD



(a) (b)

Figure 5.19: 5.19a A MUSIC-DPD pseudospectrum and 5.19b a MUSIC-DPD histogram for a scenario with six simultaneously active sources at a simulated environment of RT_{60}=0.3 s. The pink markers denote the actual positions of the audio sources.

formulation. As a reminder, in MUSIC-DPD [82] the authors propose to estimate the narrowband MUSIC pseudospectrum only in TF points that are identified as dominated by a single source using the direct-path dominance (DPD) test. In the aforementioned implementation of MUSIC-DPD algorithm, the selected incoherent narrowband pseudospectra are averaged to provide one pseudospectrum, the local peaks of which reveal the DOAs of the active sources.

Our work on 2D histograms of local DOAs motivated us to modify the MUSIC-DPD algorithm by incorporating it in the DRACOSS framework, aiming at improving its accuracy in localizing multiple sources. We, thus, propose to estimate a local DOA as the index of the highest peak of each narrowband pseudospectrum (step 2 of DRACOSS) at TF points approved by the DPD test (which now acts as step 1 of DRACOSS). All the local DOAs for a block of N_T consecutive time frames are then provided as input to the 2D histogram formation and processing steps of the proposed framework (see also Sections 5.1.3 and 5.1.4). In Fig. 5.19a we show an example of the MUSIC pseudospectrum as in [82] and in Fig. 5.19b the proposed MUSIC 2D histogram.

5.5.1 Evaluation

As for the previous DRACOSS setups, the evaluation of the DRACOSS-MUSIC was conducted with simulations and measurements in a real environments. Apart from other parameters that remain as in Section 5.3.1, the scanning area for the MUSIC pseudospectra comprises a set of 1002 points on a sphere with a distribution defined from a geodesic sphere constructed from an icosahedron with an iterative process [52]. The windows used at the histograms processing had std equal to $\sigma_A = 5^\circ$ and $\sigma_C = 20^\circ$. We have utilized speech files of duration approximately 7 seconds, with any silent periods removed. The

2D histograms and the MUSIC pseudospectra have resulted from 1 second of data, i.e., $N_T = 46$ frames.

DOA results with simulated room impulse responses

In Fig. 5.20 we explore the performance of the MUSIC algorithm when the DOA results from the averaged 2D pseudospectra, denoted as “MUSIC-SH DPD-incoh", and when it results from the 2D histograms, denoted as “MUSIC-SH DPD-2Dhist". It is clear that for high and moderate SNR values the proposed 2D histogram based processing exhibits an advantage versus the averaged pseudospectra approach. In low SNR conditions both approaches fail to provide a reasonable MEE especially when the number of active sources is increased.

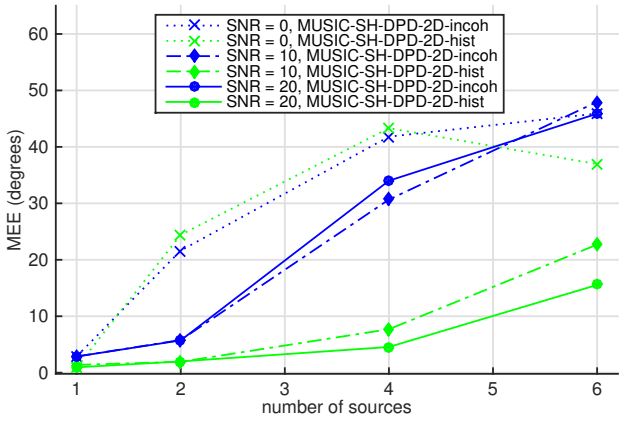


Figure 5.20: MEE versus number of sources for $RT_{60} = 0.6$ s and various SNR conditions for two approaches of the MUSIC-DPD algorithm.

We show the performance of the two MUSIC-DPD approaches along with the beamformers presented in Section 5.4 for various angular separation values between two sources in a reverberant environment of RT_{60}=0.4 sec and SNR=20 dB in Figure 5.21. All four beamformers as well as the MUSIC-SH DPD-2Dhist show very low MEE for all angular separations. As expected, when the sources get closer the MEE is higher but still in a very reasonable range of values except for the MUSIC-SH DPD-incoh which exhibits the highest error.

DOA results with measured room impulse responses

In our set of results with real RIRs we demonstrate the performance of the MUSIC-SH DPD-incoh and MUSIC-SH DPD-2Dhist approaches in Fig. 5.22 at the left plot with a simulated

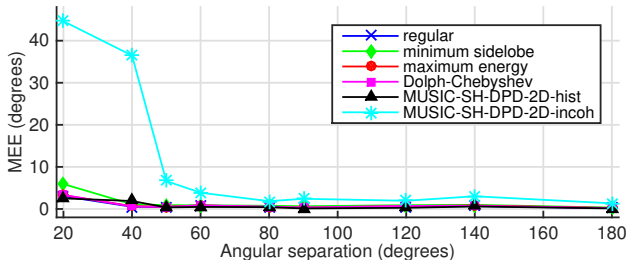


Figure 5.21: MEE vs angular separation for $RT_{60} = 0.4$ s and SNR=20 dB.

counterpart at the right side of the figure, also at SNR=45 dB. The MUSIC-SH DPD-2Dhist approach outperforms the MUSIC-SH DPD-incoh for all tested number of sources.

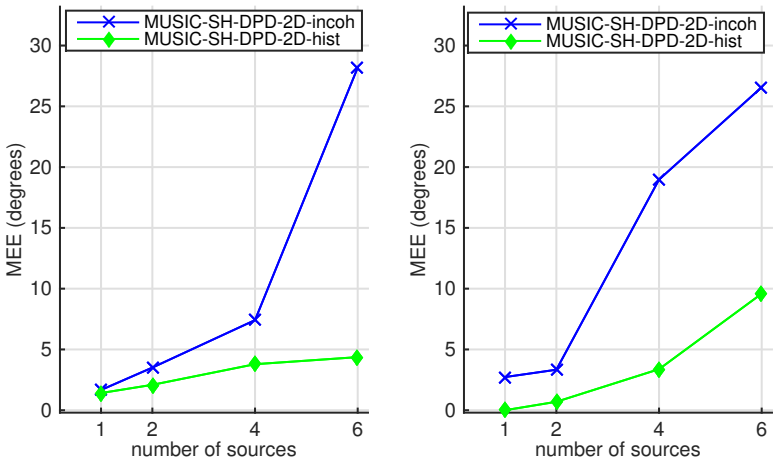


Figure 5.22: MEE versus number of sources for $RT_{60} = 0.3$ s for (a) real and (b) simulated measurements.

5.6 Counting with neural networks

In Chapter 4 we showed how we can post-process the one-dimensional histograms and estimate simultaneously not only the DOAs, but also the number of simultaneously active sources. We could follow an approach in the same spirit for the 2D histograms, but this would entail the decision of a vector of thresholds to account for the contribution of each source at the 2D histogram (see Section 4.1.4). Instead, we decided to follow a different approach inspired by the recent, ever-rising popularity of neural networks.

If we observe an adequate amount of 2D histograms such as those in Figures 5.4, 5.15, and 5.19b, we soon get the impression that the human eye gets trained and can easily identify the number of highlighted areas in the histogram, and consequently the number of active sources. This observation along with available, easy-to-use, powerful software tools, inspired us to to build and train a neural network (NN), where as training data we provide 2D histograms and as the outcome we get the number of active sources in the histogram. We have to note here that we do not aim in developing a novel neural network architecture. We instead treat the NN field as a black box and use already developed and mature libraries, freely available, such as the Theano framework [116] in order to experiment with different parameterizations and end-up with an efficient NN architecture.

With this in mind, we treated the 2D histograms as greyscale images and as proposed in the literature decided to focus on training a convolutional neural network (CNN). The concept of this idea is depicted in Figure 5.23.

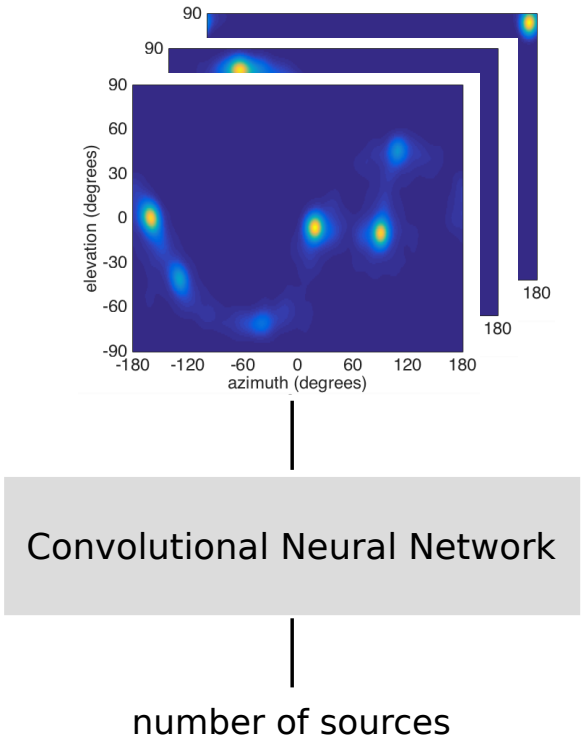


Figure 5.23: Counting using CNNs trained with 2D histograms.

5.6.1 Implementation specifics

For the implementation of our NN we have used the Keras library [20] with Theano backend [116]. We trained a convolution neural network which consists of two 2D-convolutional

layers, one LSTM layer and one Dense layer at the output. The first two convolutional layers produced six output filters and the convolution window was of size 30×30 for the first layer and 10×10 for the second layer. The dimensionality of the output of the LSTM layer was equal to 5. In order to avoid overfitting we have used a maxpooling layer followed by two dropout layers after the convolutional layers and before the LSTM one.

Although this architecture gave very promising results, as shown in Section 5.6.2, we have to note that most of our decisions are based on intuition and through a trial-and-error process. The interested reader may acquire more information from [20, 51, 89] and the references within.

5.6.2 Evaluation

We fed our CNN with 180000 histograms as training data. In this amount of data we find 2D histograms corresponding to one up to six simultaneously active sources and 18000 different arrangements of the sources around the spherical microphone array—3000 for each different number of sources. The validation set consisted of 18000 histograms and the test set includes 6000 histograms of new arrangements of the sources around the array.

All used histograms were generated using the DRACOSS framework as described in Section 5.3 and with the parameterization as described in Section 5.3.1 assuming noiseless and free field conditions. We show the counting results in the form of a confusion matrix in Table 5.2. The network exhibits a robust performance for all different number of sources. The accuracy in counting is very high when the number of sources is moderate to low, while it gets unstable when five or six sources are involved, still though with high accuracy. The overall accuracy of the trained NN reached as high as 88.9%.

Table 5.2: Confusion matrix

		$\hat{N}_S$					
		1	2	3	4	5	6
N_S	1	100%	0%	0%	0%	0%	0%
	2	4.3%	94.6%	1.1%	0%	0%	0%
	3	0%	1.4%	89.6%	9%	0%	0%
	4	0%	0%	2.7%	82.8%	14.5%	0%
	5	0%	0%	0%	8.4%	76.1%	15.5%
	6	0%	0%	0%	1%	8.6%	90.4%

The results we achieve with the specific CNN and the specific training data set are limited by the simulated noiseless and free field conditions. In future experiments we intent to further investigate the performance and suitability of a CNN for counting purposes with

various noise and reverberation conditions and with a wider collection of NN architectures in order to achieve high accuracy and stability results regardless of the simulation or real experiment conditions.

Chapter 6
Applications and DRACOSS elements in neighboring problems

DRACOSS is a framework which has been evolving in parallel with other, interesting problems of the audio signal processing area. Since the DOA estimation is a significant information to a wide range of applications, we have exploited the proposed framework and its features to a number of related problems. This chapter describes the contribution of DRACOSS in neighboring problems and related works.

6.1 Localization of sound sources with wireless acoustic sensor networks

In [42, 43] we proposed a method for multiple sound source exact location estimation utilizing a wireless acoustic sensor network (WASN), comprised by four uniform circular arrays. Each array was a four microphones UCA of radius equal to $r_a = 0.02\,\text{m}$, placed on the corners of a square-shaped area.

The proposed localization method relies on the estimation of the DOAs of each microphone array, separately, and for this purpose the DRACOSS framework was used, as developed in the 2D-space. The accuracy of the localization relies heavily on the DOA estimation accuracy. Moreover, since it was observed that when two sources are close together, an array may only detect one source, and thus sequentially affect the performance of the exact location estimation algorithm, in this work we extensively studied the DOA estimation error for the UCAs used in the WASN algorithm. The DOA estimation error at each node of the WASN was assumed to be normally distributed with a zero mean and a variance that was assumed to be dependent only upon the SNR at each sensor, which was in turn determined by the length of the path from the source to the sensor. We assumed an anechoic environment and simulated a speech source (male speaker) contaminated by white Gaussian noise at various SNR cases ranging from 5 dB to 20 dB. The noise at each microphone was uncorrelated with the speech source and with the noise at all the

other microphones. For each signal-to-noise ratio, the simulation was repeated with the source rotated in 1° increments around the array to avoid any orientation biasing effects. Figure 6.1 shows the standard deviations obtained when the DOA estimation error at each SNR was fitted with a Gaussian distribution. The fitted curve in figure 6.1 is given by:

$$\text{std(SNR)} = 1.979e^{-0.2815(\text{SNR})} + 1.884 \tag{6.1}$$

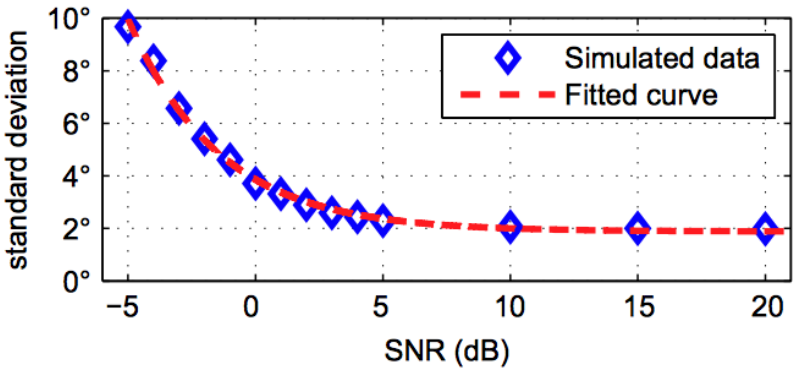


Figure 6.1: Modeling the effect of SNR on DOA estimation error standard deviation for a circular microphone array.

Since the proposed localization algorithm deals with multiple sources, it was also important to study the effect on DOA estimation when two sources were simultaneously active and close to each other, specifically when they were within the MASS, i.e., the minimum angular source separation of the UCA. Recall that for the 8-microphones circular array we have deployed for the development of DRACOSS in 2D-spaces, the MASS was estimated roughly to 25° (see also Figure 4.5). We performed a simulation study where two speech sources (one male, one female) were set at various separations of up to 20° and the energy of the second source was incrementally decreased so the signal-to-interferer ratio (SIR) seen by the first source varied from 0 dB to 20 dB. These simulations were then repeated with the sources being rotated around the array in 1° increments—while preserving their angular separation—to avoid any orientation biasing effects. In all simulations only one source was detected. Figure 6.2 shows the results of these simulations, where the DOA offset has been normalized by the separation between the sources. The fitted curve of the normalized DOA estimate, DOA_n (Figure 6.2) is given by:

$$\text{DOA}_\text{n}(\text{SIR}) = 0.5e^{-0.12987(\text{SNR})} \tag{6.2}$$

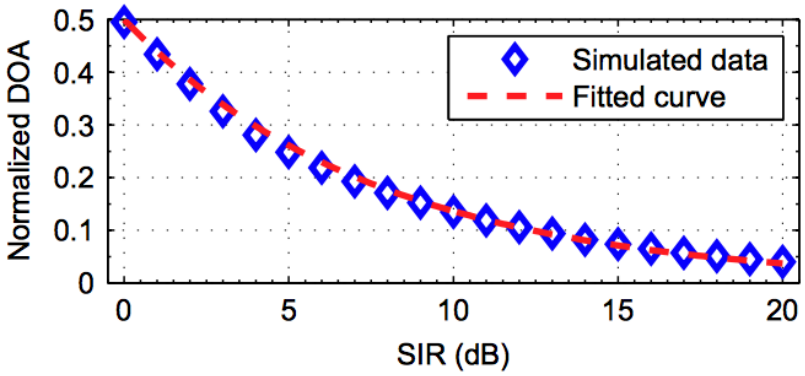


Figure 6.2: Modeling the effect of MASS and SIR on DOA estimation error for a circular microphone array.

6.2 ImmACS: an immersive audio communication system

ImmACS is an immersive audio communication system which performs real-time and efficient localization, coding, and reproduction of multiple sound sources and was developed by the SPL audio group at FORTH-ICS [110]. The system aims at performing real-time, accurate, and robust sound localization of multiple concurrent sources, indoors or outdoors, low-bitrate transmission of the original sound field, and interactive reproduction using a GUI where the user can enhance/isolate the sound(s) of his/her interest.

The DOA estimation algorithm we have used in ImmACS is DRACOSS as developed in 2D spaces, utilizing an 8-microphone UCA. ImmACS — and consequently DRACOSS — was demonstrated in real-time operation in the international conference ICASSP 2016 [75] utilizing an 8-microphone MEMS UCA. The system's operation and a comparison between an analog UCA and a digital MEMS counterpart (shown in Figure 6.3) was investigated in [3]. Our digital array was equipped with eight InvenSense ICS-43432 [76] digital MEMS bottom port microphones in a uniform circular arrangement. Its diameter was 0.06m, measured from the microphone ports, while the whole board's diameter was 0.067m. Our analog array was of the same geometry (8-microphones circular array with 0.03m radius) comprised of Shure SM93 omnidirectional microphones [107].

The experiment we performed included three different types of recordings, a 10- second rock music recording with one male singer at 0° and four instruments at 45°, 90°, 270°, and 315°, which is publicly available from the band "Nine Inch Nails", a 15-second classical music recording with four sources at 0°, 45°, 90°, and 270°, which is available from [90] and a 10-second speech recording with three speakers, where two speakers (one male at 225°, one female at 45°) are continuously and simultaneously active from the

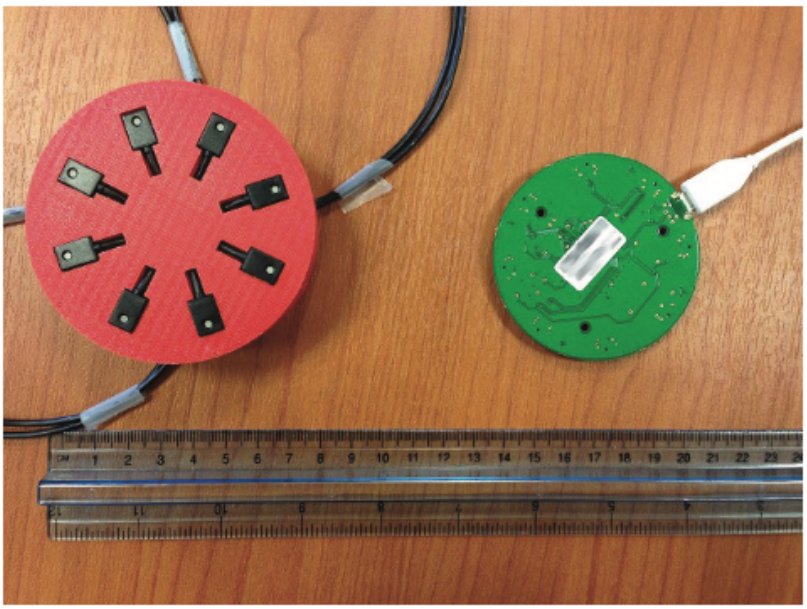


Figure 6.3: The analog microphone array (at the left) with its digital MEMS microphones counterpart (at the right).

beginning and a third speaker (female at 135°) starts speaking at the third second and remains active thereafter, resulting in three simultaneously active sound sources. The recordings were multi-track (each source on a separate track) and included both impulsive and non-impulsive sounds. Each source signal (track) was reproduced by a loudspeaker (Genelec 8050) located at the aforementioned directions at 2.10 m distance from the center of the array. The separate tracks were reproduced simultaneously and captured from each of the two aforementioned arrays. The sampling frequency for both arrays was set to 48 kHz. The recordings took place in a listening room, located at FORTH-ICS, which follows the ITU-R BS.1116 recommendation [56]. The reverberation time of the room was measured to be $RT_{60} = 0.27$ s. The DRACOSS performance is shown in Figure 6.4 which plots the DOA estimates obtained at each time frame, for all recordings, using the analog (top row) and the digital (bottom row) array. The signal of each source is plotted at its corresponding direction and the DOA estimates are overlayed on top. We observed consistent and smooth DOA estimates, especially for the speech and classical music recording. Some spurious estimates were evident in the classical recording which occurred due to the overestimation of the number of sources at these frames. A performance degradation is observed for the rock music recording, where the sources at 90°, 270°, and 315° failed to be estimated for short periods of time due to the challenging setup in terms of the number

of sources and their angular separation. Comparing the DOA estimation results between the two arrays, we can observe that the performance of the digital array is very similar to the analog array for all three recordings.

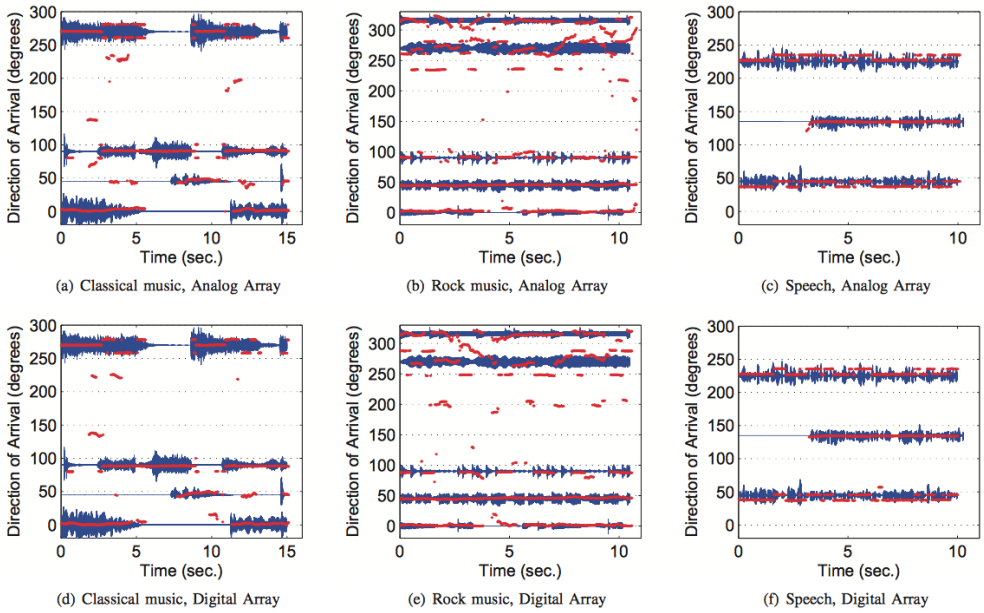


Figure 6.4: DOA estimates using the analog array signals (top row) and the digital array signals (bottom row).

6.3 MusiNet: a system for efficient networked music performance

MusiNet was a research project aiming to provide a comprehensive architecture and a prototype implementation of a complete networked music performance (NMP) system [81]. Such systems allow geographically distributed musicians to collaborate, or even perform a live concert, via computer networks. For the recording and reproduction of the interacting venues we proposed the use of spatial audio techniques, aiming at rendering the spatial attributes of each audio scene along with the audio data, thus achieving a more realistic audio impression as depicted in Figure 6.5. For this purpose we adopted ImmACS (Section 6.2) for spatial audio attributes estimation and spatial audio reproduction, respectively. Thus, DRACOSS, as an essential part of the ImmACS system, was used in the MusiNet project for acquiring the number and DOAs of involved musicians in a participating NMP venue.

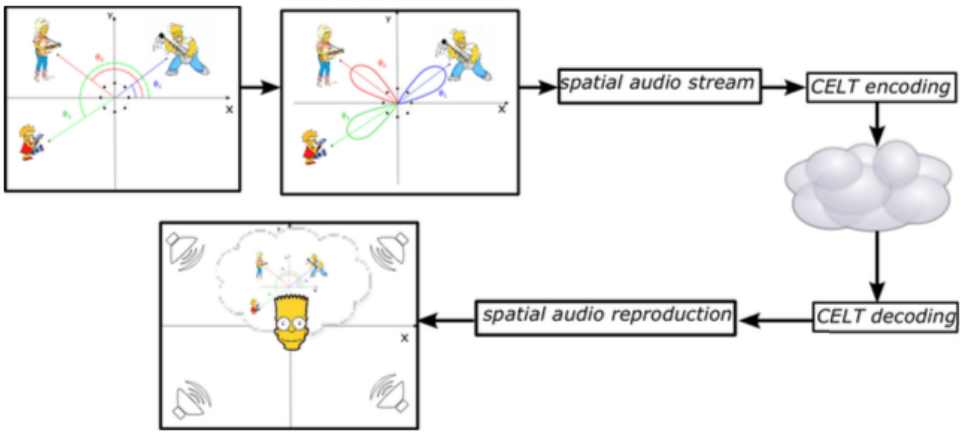


Figure 6.5: Spatial audio recording and reproduction in MusiNet

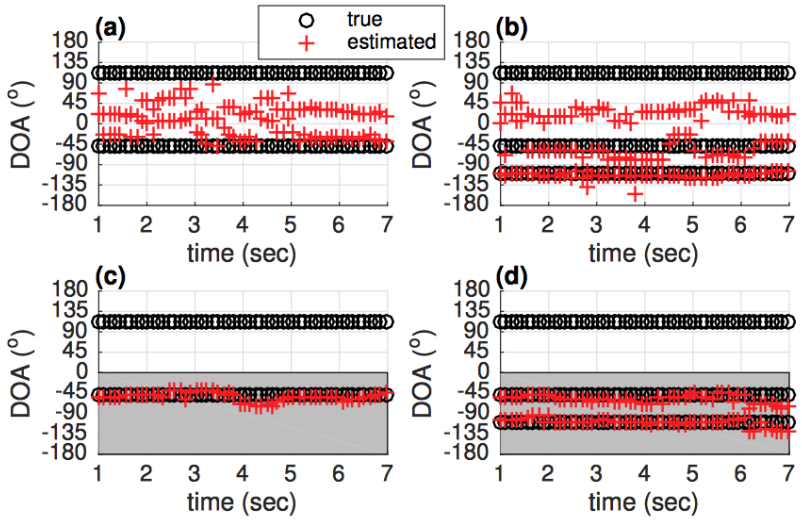


Figure 6.6: DOA estimation result with coherent sources when two and three sources are active with (a)-(b) the AIV estimator, and (c)-(d) the SCAIV estimator. The gray region in the plots indicates the analysis area.

6.4 Spatially constrained active intensity vectors for DOA estimation of coherent sources

In [29] we proposed a method suitable for DOA estimation of coherent sources, simultaneously active on the same plane. The method is based on the formulation of a higher-order

intensity vector estimator on spatially constrained regions of the plane which we have called SCAIV. SCAIV aims to overcome the weaknesses of first-order active intensity vector estimators. The local DOA estimates provided by SCAIV are then fed in a block-based manner to 1D histograms which reveal the final DOAs of the sources, by adopting the logic of steps 3 and 4 of DRACOSS for two-dimensional spaces. The advantages of SCAIV versus the first-order active intesity vector (AIV) are highlighted in Figure 6.6 where it is obvious that AIV fails completely at estimating the DOA of coherent sources.

6.5 Perpendicular cross-spectral fusion for sound source localization

We used ideas of the DRACOSS framework in our recent work on a new DOA estimation algorithm for small planar arrays, presented in [111]. The algorithm depends on the perpendicular cross-spectra fusion (PCSF) of DOA estimates from estimation subsystems which operate on each TF bin in parallel and on a coherence metric which decides upon the reliability of the TF bin and consequently on the quality of the local DOA estimate. The final set of local DOA estimates in provided as input to 2D DRACOSS steps 3 and 4 in order to form the one-dimensional histogram and post-process it to infer the number of the sources and their final DOAs.

PCSF achieves significant improvements in terms of the histograms' quality which is reflected to the MAEE as can be shown in Figures 6.7 and 6.8. The method was compared against intensity-based DOA estimation as the one used in the DirAC framework [95] (denoted as DIRAC in the figures) and the method in [60] (denoted as CICS in the figures, see also 4.1.2) as alternative local DOA estimators (DRACOSS second step) utilizing a 4 microphones UCA of radius equal to 0.02 m. However we have to note that this improvement comes with a cost at the computational complexity. Moreover the algorithm demands the existence of specific direction microphone pairs in order to ensure the presence of the DOA subsystems and consequently its functional behavior.

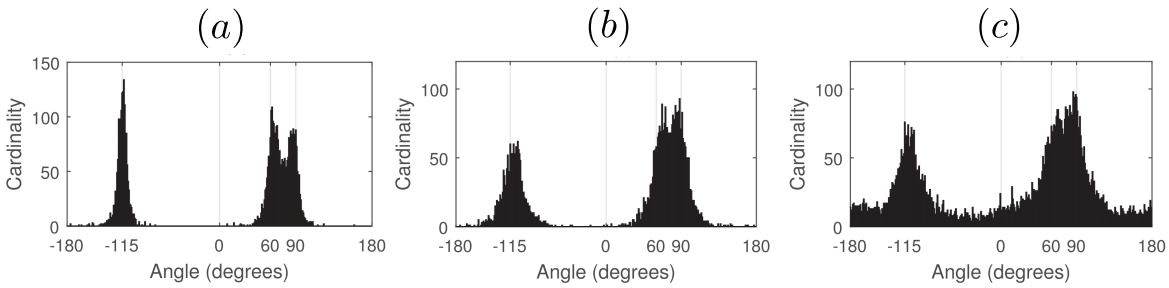


Figure 6.7: Histograms with DOAs obtained from (a) PCSF, (b) DIRAC, and (c) CICS. Results are shown for a simulated reverberant environment of RT_{60} = 0.3 s with three simultaneously active sources at −115°, 60° and 90°.

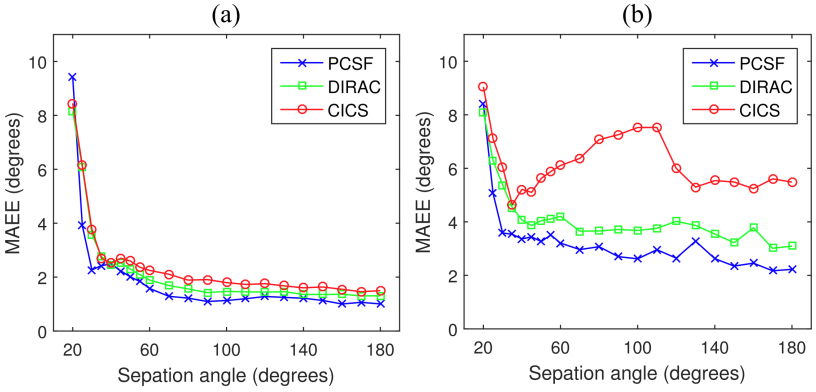


Figure 6.8: MAEE for two sources in a simulated environment as a function of the angular distance between the sources. Results are shown at 10 dB SNR for (a) RT_{60} = 0.2 s, (b) RT_{60} = 0.4 s.

Chapter 7
Conclusion

7.1 Synopsis of Contributions

This thesis presented DRACOSS, an integrated framework for direction of arrival estimation and counting of multiple sound sources. DRACOSS is comprised by four distinct steps, namely the exploitation of sparsity of the sound signals and the detection of areas of single source activity at the chosen transformation domain, the estimation of local DOAs at the detected areas, the formation of histograms of local DOA estimates and the post-processing of histograms in order to obtain the final DOA estimation and counting results.

DRACOSS was developed both in 2D and 3D spaces. In the first case the framework estimates the azimuth of the active audio sources and utilizes a uniform circular array. In the second case both the azimuth and the elevation are estimated, using an almost uniform spherical array. For both spaces we utilized a sparsity criterion based on the auto- and cross-correlations among pairs of microphone signals. For the 2D space the local DOA estimator was one specifically designed for UCA arrays while for the 3D space the DOA estimation was based on intensity vector estimates. The histograms in both spaces were smoothed in order to highlight the presence of the sources and were post-processed through an iterative procedure which led to the final DOA estimation and counting. For the 3D case the smoothed histograms were also fed to a convolutional neural network which provided very promising results in terms of counting.

DRACOSS was evaluated in a wide range of conditions, for varying number of sources, different levels of noise and reverberation and in comparison with other state-of-the-art methods. It showed very robust performance in terms of the mean absolute estimation error even when as many as six sources were simultaneously active. The framework outperformed known DOA methods of the literature, such as the MUSIC algorithm, while it presented excellent counting results, outperforming other known counting approaches. DRACOSS can operate with any topology of a compact microphone array, e.g., with linear, circular, cylindrical, spherical arrays, given that an appropriate local DOA estimator is chosen. Most importantly, it was shown that the performance of state-of-the-art DOA algorithms was significantly improved when they were adjusted in the DRACOSS frame-

work. Another beneficial characteristic of the framework is the distinct and independent nature of the four fundamental DRACOSS blocks which provides increased degrees of freedom to the audio engineer, such that she/he can modify the framework in order to fit best a potential application's needs. An additional significant benefit of DRACOSS is its low computational cost, which allowed the implementation and operation of the framework in real time.

7.2 Directions for Future Work and Research

Since DRACOSS is a four-step framework, decisions need to be made upon each step, thus a wide range of combinations exist, given these decisions. Thus, further work on examining all possible combinations at each step will reveal the full potential of the framework. Having examined thus far the MCC, WDO, and DPD sparsity criteria in varying scenarios and with different local DOA estimators, a structured and controlled comparison between the criteria could reveal the best candidate. The same holds for the local DOA estimators, i.e., an extensive comparison should be performed between the intensity vector estimator and the beamforming and the MUSIC-based estimators, along with comparative tests with other proposed estimators in the literature.

Using neural networks in the DRACOSS framework is a very recent idea which needs to be further explored. The first counting results with CNNs showed very promising performance, however additional CNNs-based architectures with data from adverse noise and reverberation conditions need to be examined. Furthermore we intent to investigate the possibility of using neural networks not only for counting but also for infering the DOAs from the histograms and not only for the 3D case but also for 2D spaces.

So far we have successfully implemented and demonstrated the real-time operation of the DRACOSS framework in the ImmACS system using both an analog and a digital MEMS circular microphone array. Our intention is to try and implement the 3D development of DRACOSS in real-time and built a digital spherical microphone array with more microphones, thus increased analysis capabilities.

Bibliography

[1] T. D. Abhayapala. Generalized framework for spherical microphone arrays: Spatial and frequency decomposition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 5268–5271, 2008.

[2] F. Abrard, Y. Deville, and P. White. From Blind Source Separation To Blind Source Cancellation In The Underdetermined Case: A New Approach Based On Time-Frequency Analysis. In *Proceedings of the 3rd International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 734–739, Oct 2001.

[3] A. Alexandridis, S. Papadakis, D. Pavlidi, and A. Mouchtaris. Development and evaluation of a digital mems microphone array for spatial audio. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 612–616, Aug 2016.

[4] J. B. Allen and D. A. Berkley. Image method for efficiently simulating smallb••room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[5] S. Araki, T. Nakatani, H. Sawada, and S. Makino. Stereo source separation and source counting with map estimation with dirichlet prior considering spatial aliasing problem. In *Independent Component Analysis and Signal Separation*, volume 5441 of *Lecture Notes in Computer Science*, pages 742–750. Springer Berlin Heidelberg, 2009.

[6] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1), January 2010.

[7] S. Argentieri and P. Danès. Broadband variations of the music high-resolution method for sound source localization in robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2009–2014, November 2007.

[8] M. Atiyah and P. Sutcliffe. Polyhedra in physics, chemistry and geometry. *Milan Journal of Mathematics*, 71(1):33–58, 2003.

[9] D. Bechler and K. Kroschel. Considering the second peak in the GCC function for multi-source TDOA estimation with microphone array. In *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 315–318, 2003.

[10] D. Bechler, M.S. Schlosser, and K. Kroschel. System for robust 3D speaker tracking using microphone array measurements. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2117–2122, September 2004.

[11] F. Belloni and V. Koivunen. Unitary root-music technique for uniform circular array. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, (ISSPIT)*, pages 451–454, December 2003.

[12] J. Benesty, J. Chen, and Y. Huang. Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on Speech and Audio Processing*, 12(5), September 2004.

[13] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer Topics in Signal Processing, Vol. 1. Springer, 2008.

[14] S. Bertet, J. Daniel, and S. Moreau. 3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone. In *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

[15] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Oct 2006.

[16] C. Blandin, A. Ozerov, and E. Vincent. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, October 2011.

[17] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, Aug 1969.

[18] J.F. Cardoso and A.Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[19] J. Chen, J. Benesty, and Y. Huang. Time delay estimation in room acoustic environments: An overview. *EURASIP Journal on Applied Signal Processing*, pages 1–19, 2006.

[20] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[21] M. Cobos, A. Marti, and J. J Lopez. A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. *Signal Processing Letters, IEEE*, 18(1):71–74, 2011.

[22] P. Comon and C. Jutten. *Handbook of blind source separation: independent component analysis and applications*. Academic Press. Elsevier, 2010.

[23] J. H. Conway et al. *Sphere packings, lattices and groups*, volume 3. Springer-Verlag New York, 1993.

[24] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10):1365–1376, Oct 1987.

[25] M. J. Crocker and F. Jacobsen. *Sound Intensity*, chapter 156, pages 1855–1868. Wiley-Blackwell, 2007.

[26] J Daniel. *Représentation de champs acoustiques, application á la reproduction eta la transmission de scenes sonores complexes dans un contexte multimédia*. PhD thesis, Ph. D. thesis, University of Paris 6, Paris, France, 2000.

[27] J. Daniel, J. B. Rault, and J. D. Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*. Audio Engineering Society, 1998.

[28] S. Delikaris-Manias. *Parametric spatial audio processing utilising compact microphone arrays*. PhD thesis, Department of Signal Processing and Acoustics, Aalto Finland, 2017. Aalto University publication series DOCTORAL DISSERTATIONS.

[29] S. Delikaris-Manias, D. Pavlidi, A. Mouchtaris, and V. Pulkki. Doa estimation with histogram analysis of spatially constrained active intensity vectors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 526–530, March 2017.

[30] Y. Deville and M. Puigt. Temporal and time-frequency correlation-based blind source separation methods. part i: Determined and underdetermined linear instantaneous mixtures. *Signal Processing*, 87:374–407, March 2007.

[31] Y. Deville, M. Puigt, and B. Albouy. Time-frequency blind signal separation: extended methods, performance evaluation for speech sources. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pages 255–260, July 2004.

[32] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. Robust localization in reverberant rooms. In *Microphone Arrays*, pages 157–180. Springer, 2001.

[33] J. Dmochowski, J. Benesty, and S. Affes. Direction of arrival estimation using the parameterized spatial correlation matrix. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1327–1339, May 2007.

[34] J. P. Dmochowski, J. Benesty, and S. Affes. Broadband music: Opportunities and challenges for multiple source localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 18–21, October 2007.

[35] H. Do, H. F. Silverman, and Y. Yu. A real-time SRP-PHAT source location implementation using stochastic region contraction (src) on a large-aperture microphone array. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–121. IEEE, 2007.

[36] C. Evers, A. H. Moore, and P. A. Naylor. Multiple source localisation in the spherical harmonic domain. In *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 258–262, Sept 2014.

[37] F. Fahy. *Sound intensity*. CRC Press, 2002.

[38] E. Fishler, M. Grosmann, and H. Messer. Detection of signals by information theoretic criteria: general asymptotic performance analysis. *IEEE Transactions on Signal Processing*, 50(5):1027–1036, may 2002.

[39] M. A. Gerzon. Periphony: With-height sound reproduction. *Journal of Audio Engineering Society*, 21(1):2–10, 1973.

[40] J.W. Goodman. *Introduction to Fourier Optics*. MaGraw-Hill, 2nd edition, 1996.

[41] R. Gribonval and M. Zibulevsky. Sparse Component Analysis. In *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pages 367–420. Academic Press, 2010.

[42] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris. Real-time localization of multiple audio sources in a wireless acoustic sensor network. In *The 22nd European Signal Processing Conference (EUSIPCO)*, pages 306–310, Sept 2014.

[43] A. Griffin, A.s Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris. Localizing multiple audio sources in a wireless acoustic sensor network. *Signal Processing*, 107:54 – 67, 2015. Special Issue on ad hoc microphone arrays and wireless acoustic sensor networks Special Issue on Fractional Signal Processing and Applications.

[44] A. Griffin, De. Pavlidi, M. Puigt, and A. Mouchtaris. Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 2303–2307, August 2012.

[45] S. Hafezi, A. H. Moore, and P. A. Naylor. 3d acoustic source localization in the spherical harmonic domain based on optimized grid search. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 415–419, March 2016.

[46] S. Hafezi, A. H. Moore, and P. A. Naylor. Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search. In *24th European Signal Processing Conference (EUSIPCO)*, pages 602–606, Aug 2016.

[47] S. Hafezi, A. H. Moore, and P. A. Naylor. Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1956–1968, Oct 2017.

[48] G. Hamerly and C. Elkan. Learning the k in k-means. In *Neural Information Processing Systems*, pages 281–288. MIT Press, 2003.

[49] R. H. Hardin and N. J. A. Sloane. Mclaren's improved snub cube and other new spherical designs in three dimensions. *Discrete & Computational Geometry*, 15(4):429–441, 1996.

[50] R.H. Hardin and Sloane N.J.A. Spherical designs. http://neilsloane.com/sphdesigns/.

[51] G. Hinton. Neural netowroks for machine learning. https://www.coursera.org/learn/neural-networks, 2017.

[52] F. Hollerweger. *Periphonic sound spatialization in multi-user virtual environments*. Citeseer, 2006.

[53] A. Hyvärinen, J. Karhunen, and E. Oja, editors. *Independent Component Analysis*. Wiley, 2001.

[54] C.T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita. Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS)*, pages 2027–2032, October 2009.

[55] N. Ito, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama. Crystal-MUSIC: Accurate localization of multiple sources in diffuse noise environments using crystal-shaped microphone arrays. In *LVA/ICA'10*, pages 81–88, 2010.

[56] ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1997.

[57] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor. 3D source localization in the spherical harmonic domain using a pseudointensity vector. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 442–446, 2010.

[58] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor. Rigid sphere room impulse response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America*, 132(3):1462–1472, 2012.

[59] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures. In *Proceedings of the IEEE International Co nference on Acoustics, Speech, and Signal Processing, (ICASSP)*, volume 5, pages 2985–2988, 2000.

[60] A. Karbasi and A. Sugiyama. A new DOA estimation method using a circular microphone array. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 778–782, 2007.

[61] H. Khaddour, J. Schimmel, and M. Trzos. Estimation of Direction of Arrival of Multiple Sound Sources in 3D Space Using B-Format. *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, 2(2), 2013.

[62] C.H. Knapp and G.C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4), August 1976.

[63] H. Krim and M. Viberg. Two decades of array signal processing research - the parametric approach. *IEEE Signal Processing Magazine*, pages 67–94, July 1996.

[64] M. Laitinen, F.n Kuech, S. Disch, and V. Pulkki. Reproducing applause-type signals with directional audio coding. *J. Audio Eng. Soc*, 59(1/2):29–43, 2011.

[65] E. A. Lehmann. Fast image-source method: Matlab code. http://www.eric-lehmann.com/.

[66] E.A. Lehmann and A.M. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1429–1439, August 2010.

[67] D. Levin, E. A. P. Habets, and S. Gannot. On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields. *The Journal of the Acoustical Society of America*, 128(4):1800–1811, 2010.

[68] C. L. Liu and H. M. Hang. Direction of arrival estimation of speech signals using ica and music methods. In *The 5th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1768 –1773, june 2010.

[69] B. Loesch, S. Uhlich, and Bin Yang. Multidimensional localization of multiple sound sources using frequency domain ica and an extended state coherence transform.

In *Proceedings of the IEEE/SP 15th Workshop on Statistical Signal Processing, (SSP)*, pages 677–680, September 2009.

[70] B. Loesch and B. Yang. Source number estimation and clustering for underdetermined blind source separation. In *Proceedings of the International Workshop for Acoustics Echo and Noise Control, (IWAENC)*, 2008.

[71] A. Lombard, Y. Zheng, and W. Kellermann. Synthesis of ica-based methods for localization of multiple broadband sound sources. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 157–160, May 2011.

[72] A. Lombard, Yuanhang Zheng, H. Buchner, and W. Kellermann. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1490–1503, August 2011.

[73] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[74] A. Masnadi-Shirazi and B. D. Rao. An ica-based rfs approach for doa tracking of unknown time-varying numberof sources. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 599–603, Aug 2012.

[75] Y. Mastorakis, A. Alexandridis, D. Pavlidi, S. Papadakis, and A. Mouchtaris. ImmACS: A Real-time Immersive Audio Communication System. http://www.icassp2016.org/ST-1.asp.

[76] InvenSense ICS-43432 Digital MEMS Microphone. http://www.invensense.com/products/digital/ics-43432.

[77] mh acoustics. EM32 eigenmike microphone array release notes (v17. 0). Technical report, mh acoustics, Summit, NJ, USA, 2013.

[78] Satish Mohan, Michael E. Lockwood, Michael L. Kramer, and Douglas L. Jones. Localization of multiple acoustic sources with small arrays using a coherence test. *The Journal of the Acoustical Society of America*, 123(4):2136–2147, 2008.

[79] A. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. In *23rd EuropeanSignal Processing Conference (EUSIPCO)*, pages 2296–2300, Aug 2015.

[80] S. Moreau, J. Daniel, and S. Bertet. 3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of Spherical Microphone. In *The 120th Convention of the Audio Engineering Society (AES)*, May 2006.

[81] MusiNet: Comprehensive design and implementation of a networked music performance system. http://musinet.aueb.gr.

[82] O. Nadiri and B. Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1494–1505, 2014.

[83] K. Nakadai, D. Matsuura, H. Kitano, H. G. Okuno, and H. Kitano. Applying scattering theory to robot audition system: Robust sound source localization and extraction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1147–1152, 2003.

[84] F. Nesta. http://bssnesta.webatu.com/software.html.

[85] F. Nesta and M. Omologo. Cooperative Wiener-ICA for source localization and Separation by distributed microphone arrays. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–4, March 2010.

[86] F. Nesta and M. Omologo. Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):246–260, January 2012.

[87] F. Nesta, P. Svaizer, and M. Omologo. Robust two-channel tdoa estimation for multiple speaker localization by using recursive ica and a state coherence transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4597–4600, April 2009.

[88] F. Nesta, P. Svaizer, and M. Omologo. Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):624–639, 2011.

[89] A. Ng. Machine learning. https://www.coursera.org/learn/machine-learning, 2017.

[90] J. Pätynen, V. Pulkki, and T. Lokki. Anechoic recording system for symphony orchestra. In *Acta Acustica united with Acustica*, pages 856–865, 2008.

[91] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris. Source counting in real-time sound source localization using a circular microphone array. In *Proceedings of the IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 521–524, June 2012.

[92] D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris. Real-time multiple sound source localization using a circular microphone array based on single-source confidence

measures. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2625–2628, March 2012.

[93] M. Puigt and Y. Deville. A time-frequency correlation-based blind source separation method for time-delayed mixtures. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 853–856, May 2006.

[94] M. Puigt and Y. Deville. A new time-frequency correlation-based source separation method for attenuated and time shifted mixtures. In *Proceedings of the 8th International Workshop (ECMS and Doctoral School) on Electronics, Modelling, Measurement and Signals*, pages 34–39, 2007.

[95] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[96] B. Rafaely. Analysis and design of spherical microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 13(1):135–143, 2005.

[97] B. Rafaely. *Fundamentals of Spherical Array Processing*, volume 8. Springer, 2015.

[98] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher. Spherical microphone array beamforming. In *Speech Processing in Modern Communication*, pages 281–305. Springer, 2010.

[99] B. Rafaely, B. Weiss, and E. Bachmat. Spatial aliasing in spherical microphone arrays. *IEEE Transactions on Signal Processing*, 55(3):1003–1010, March 2007.

[100] S. Rickard, R. Balan, and J. Rosca. Real-time time-frequency based blind source separation. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 651–656, Oct 2001.

[101] S. Rickard, T. Melia, and C. Fearon. DESPRIT - Histogram based blind source separation of more sources than sensors using subspace methods. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 5–8, Oct 2005.

[102] R. Roy and T. Kailath. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7):984–995, 1989.

[103] H. Sawada, R. Mukai, S. Araki, and S. Malcino. Multiple source localization using independent component analysis. In *IEEE Antennas and Propagation Society International Symposium*, volume 4B, pages 81–84, July 2005.

[104] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, March 1986.

[105] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proceedings of the nternational Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 261–266, 1999.

[106] S. Schulz and T. Herfet. On the window-disjoint-orthogonality of speech source in reverberant humanoid scenarios. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pages 241–248, 2008.

[107] Shure SM93 Lavalier Microphone. http://www.shure.com/americas/products/microphones/sm/sm93-lavalier-microphone.

[108] Z. I. Skordilis, A. Tsiami, P. Maragos, G. Potamianos, L. Spelgatti, and R. Sannino. Multichannel speech enhancement using mems microphones. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2729–2733, April 2015.

[109] D. Smith, J. Lukasiak, and I. S. Burnett. A two channel, block-adaptive audio separation technique based upon time-frequency information. In *Poceedings of the 12th European Signal Processing Conference (EUSIPCO)*, pages 393–396, Sept 2004.

[110] FORTH-ICS SPL. ImmACS: An immersive audio communication system. http://users.ics.forth.gr/~mouchtar/ImmACS/.

[111] N. Stefanakis, D. Pavlidi, and A. Mouchtaris. Perpendicular cross-spectra fusion for sound source localization with a planar microphone array. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1821–1835, Sept 2017.

[112] M. Swartling, B. Sällberg, and N. Grbić. Source localization for multiple speech sources using low complexity non-parametric source separation and clustering. *Signal Processing*, 91:1781–1788, August 2011.

[113] S. Tervo. Direction estimation based on sound intensity vectors. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 700–704, 2009.

[114] H. Teutsch. *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, volume 348. Springer Science & Business Media, 2007.

[115] H. Teutsch and W. Kellermann. EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 89–92, March 2005.

[116] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[117] T. Van den Bogaert, E. Carette, and J. Wouters. Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. *International Journal of Audiology*, 50(3):164–176, 2011.

[118] E. Vincent, S. Arberet, and R. Gribonval. Underdetermined instantaneous audio source separation via local gaussian modeling. In *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 775–782. Springer Berlin Heidelberg, 2009.

[119] D. B. Ward and T.D. Abhayapala. Range and bearing estimation of wideband sources using an orthogonal beamspace processing structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, (ICASSP)*, volume 2, pages ii–109–12 vol.2, 2004.

[120] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 33(2):387–392, 1985.

[121] M. Wax, T. K. Shan, and T. Kailath. Spatio-temporal spectral analysis by eigenstructure methods. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-32(4):817–827, August 1984.

[122] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.

[123] J.X. Zhang, M.G. Christensen, J. Dahl, S.H. Jensen, and M. Moonen. Robust implementation of the music algorithm. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 3037–3040, April 2009.

[124] Y. Zhang and B. P. Ng. MUSIC-Like DOA Estimation Without Estimating the Number of Sources. *IEEE Transactions on Signal Processing*, 58(3):1668–1676, March 2010.

[125] F. Zotter, H. Pomberger, and M. Noisternig. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica*, 98(1):37–47, 2012.

Part III

Appendices

Appendix A
Publications, Patents, and Systems

The research activity related to this thesis has so far produced the following publications (ordered by publication date):

(1) **D. Pavlidi**, M. Puigt, A. Griffin, and A. Mouchtaris, *"Real-Time Multiple Sound Source Localization and Counting using a Circular Microphone Array"*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, October 2013, pp. 2193 - 2206.

(2) D. Akoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lalioti, A. Mouchtaris, **D. Pavlidi**, G. C. Polyzos, P. Tsakalides, G. Xylomenos and P. Zervas, *"The MusiNet project: Towards unraveling the full potential of Networked Music Performance systems"*, 7th International Conference on Information, Intelligence, Systems and Applications (IISA), July 2014, pp. 1-6.

(3) A. Griffin, A. Alexandridis, **D. Pavlidi**, and A. Mouchtaris, *"Real-time localization of multiple audio sources in a wireless acoustic sensor network "*, 22nd European Signal Processing Conference (EUSIPCO), September 2014, pp. 306-310.

(4) A. Griffin, A. Alexandridis, **D. Pavlidi**, Y. Mastorakis and A. Mouchtaris, *"Localizing multiple audio sources in a wireless acoustic sensor network"*, Signal Processing, vol. 107, February 2015, pp. 54 - 67.

(5) D. Akoumianakis, C. Alexandraki, V. Alexiou, C. Anagnostopoulou, A. Eleftheriadis, V. Lalioti, Y. Mastorakis, A. Modas, A. Mouchtaris, **D. Pavlidi**, G. C. Polyzos, P. Tsakalides, G. Xylomenos and P. Zervas, *"The MusiNet project: Addressing the challenges in Networked Music Performance systems "*, 6th International Conference on Information, Intelligence, Systems and Applications (IISA), July 2015, pp. 1-6.

(6) **D. Pavlidi**, S. Delikaris-Manias, V. Pulkki and A. Mouchtaris, *3D localization of multiple sound sources with intensity vector estimates in single source zones*, 23rd European Signal Processing Conference (EUSIPCO), August 2015, pp. 1556-1560.

(7) **D. Pavlidi**, S. Delikaris-Manias, V. Pulkki and A. Mouchtaris, *3D DOA estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors*, IEEE International Conference on Acoustics, Speech and Signal Process-

ing (ICASSP), March 2016, pp. 96-100.

(8) A. Alexandridis, S. Papadakis, **D. Pavlidi**, and A. Mouchtaris, *"Development and Evaluation of a Digital MEMS Microphone Array for Spatial Audio"*, Proceedings of the 24^{th} European Signal Processing Conference (EUSIPCO), August-September 2016, pp. 612 –616.

(9) S. Delikaris-Manias, **D. Pavlidi**, V. Pulkki and A. Mouchtaris, *3D localization of multiple audio sources utilizing 2D DOA histograms*, Proceedings of the 24^{th} European Signal Processing Conference (EUSIPCO), August-September 2016, pp. 1473 –1477.

(10) S. Delikaris-Manias, **D. Pavlidi**, A. Mouchtaris, and V. Pulkki, "DOA estimation with histogram analysis of spatially constrained active intensity vectors", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pp. 526–530, 2017.

(11) N. Stefanakis, **D. Pavlidi**, and A. Mouchtaris, *"Perpendicular Cross-Spectra Fusion for Sound Source Localization with a small Microphone Array"*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 9, September 2017, pp. 1821 - 1835.

(12) A. Alexandridis, **D. Pavlidi**, N. Stefanakis, and A. Mouchtaris, "Parametric Spatial Audio Techniques in Teleconferencing and Remote Presence", *Parametric Time-Frequency Domain Spatial Audio*, eds. V. Pulkki, S. Delikaris-Manias, and A. Politis, John Wiley & Sons, October, 2017

Patents

(1) **D. Pavlidi**, A Griffin, A. Mouchtaris,"SOUND SOURCE CHARACTERIZATION APPARATUSES, METHODS AND SYSTEMS", USPTO Non-Provisional Application No. 14/038,726 (recently allowed, to be issued), filed September 26, 2013.

Systems

DRACOSS, as developed for 2D spaces is a fundamental part of the ImmACS [110] system which was demonstrated in the international conference ICASSP 2016, operating in real time.

Appendix B
Microphone arrays, theorems, and assumptions

B.0.1 Microphone arrays geometries

A microphone array consists of multiple microphones placed at different spatial locations in a way that the spatial information is well captured [13]. Literature review unearths a wide variety of geometries, i.e., in the 2D space we see linear, triangular, circular or planar arrays of random positioning of the microphones, in the 3D space, spherical, cubic or cylindrical microphone arrays.

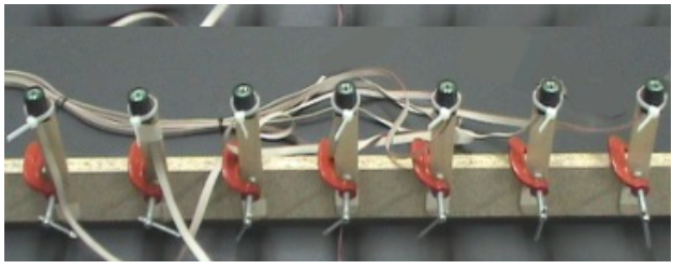


Figure B.1: A linear array comprised by 7 MEMS digital microphones built at the National Technical University of Athens (image taken from [108]).

Depending on the nature of the application, the geometry of the microphone array may play an important role in the formulation of the processing algorithms. For example linear arrays have been widely used because of the simplicity they provide and their convenient positioning due to their design. On the other hand this geometry suffers from the ambiguity of distinguishing the rear-front direction of a propagating source. Circular arrays can tackle the aforementioned ambiguity, hence they can be considered as the dom-

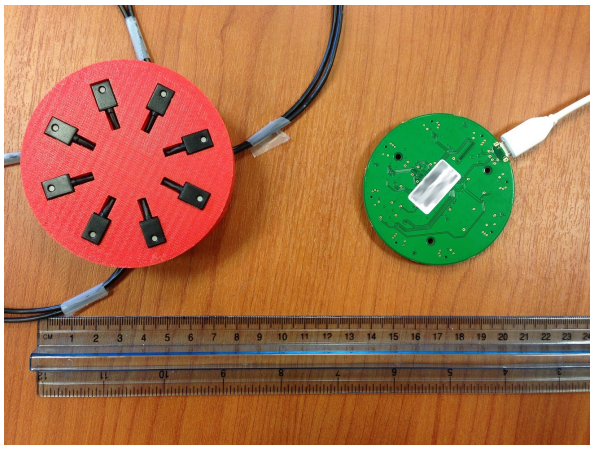


Figure B.2: Two UCAs comprised by 8 analog (left) and 8 MEMS digital (right) microphones, both built at FORTH-ICS [3].

Figure B.3: The Eigenmike comprised by 32 microphones, nearly uniformly placed on the surface of a 4.2cm-radius sphere [77].

inant array geometry in 2D spaces. In the 3D space interesting structures have appeared the recent years with the most dominant one being the sphere. Microphones mounted on cylinders have also been popular in the research community.

In Figures B.1, B.2, B.3, B.4 we show pictures of a linear, a circular, a spherical and a cylindrical array that are currently prototypes, built in educational institutions or are al-

Figure B.4: A six-element array configuration mounted on a rigid cylinder, built at Aalto University [28].

ready commercial products (in the case of the Eigenmike).

B.0.2 Far-field assumption

In all simulations and experiments it is assumed that the sources lie in the far-field of the microphone, that is they are far enough from the center of the array so that the microphones receive planar wavefronts from the emitting source. More formally, a source lies in the far-field (or Fraunhofer) region of a microphone array if the following condition, known as the “antenna designer’s formula", is satisfied [40]:

$$r_s > \frac{2d_a^2}{\lambda_{\min}}. \tag{B.1}$$

Thus,

$$r_s > \frac{2d_a^2 f_{\max}}{c}, \tag{B.2}$$

where r_s is the distance from the source to the center of the array, d_a is the largest linear dimension of the array, c is the speed of sound, $\lambda_{\min}$ is the source’s signal minimum wavelength and consequently $f_{\max}$ is the maximum frequency. Obviously, if the above condition is not satisfied, (i.e., $r_s \leq \frac{2d_a^2 f_{\max}}{c}$), then the source is located in the near-field.

If the source(s) lie in the near-field of the array, then the wave-fronts impinging on the

microphone are spherical and the signal's propagation vector (i.e., the signal's direction of propagation relative to the adopted coordinate system) varies across the array. On the other hand, if the source(s) lies in the far-field of the array, then the wave-fronts are planar and this vector is the same across all microphones of the array, independently of their location. The above is illustrated in Figure B.5.

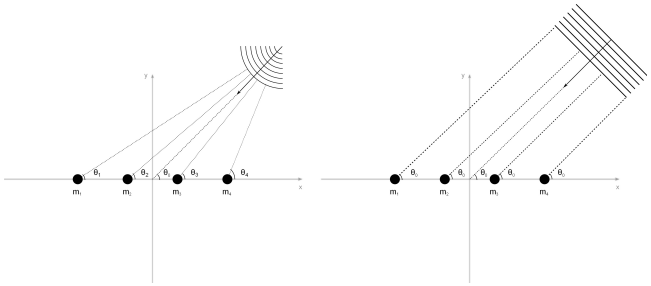


Figure B.5: The near (left) and far (right) field cases for a uniform linear array.

B.0.3 Mean correlation coefficient theorem

For completeness of the text and for the convenience of the reader, we here quote the theorem used for the detection of SSZs. All material was originally presented in [94].

Initial assumption: when several sources are active in the same analysis zone (K), they should vary so that the moduli of at least two observations are linearly independent

Theorem. *A necessary and sufficient condition for a source to be isolated in an analysis zone* (K) *is*

$$r'_{i,j}(K) = 1 \qquad \forall i, j \in \{1, \ldots, Q\}. \tag{B.3}$$

Proof. Suppose that $r'_{i,j}(K) = 1$, $\forall i, j \in \{1, \ldots, Q\}$ in a constant time analysis zone (τ, K). We have to prove that in that case only one source is active in that zone. We consider the moduli of the observations $X_i(\tau, k)$ as vectors of dimension K:

$$V_{|x_i|} = \Big[|X_i(\tau, k_1)|, \quad |X_i(\tau, k_2)|, \quad \ldots, \quad , |X_i(\tau, k_K)|\Big] \tag{B.4}$$

Using this notation the correlation coefficient, $r'_{i,j}(K)$, can be rewritten as:

$$r'_{i,j}(K) = \frac{\langle V_{|x_i|}, V_{|x_j|} \rangle}{\|V_{|x_i|}\| \cdot \|V_{|x_j|}\|} \tag{B.5}$$

Applying the Cauchy-Schwarz inequality to the eq. (B.5) we obtain:

$$r'_{i,j}(K) \leq 1 \tag{B.6}$$

The equality holds if and only if the vectors $V_{|x_i|}$ and $V_{|x_j|}$ are linearly dependent, i.e., if and only if there exists a real positive number μ such that:

$$V_{|x_i|} = \mu V_{|x_j|} \tag{B.7}$$

This would mean that $\forall k \in K$ it should hold that:

$$|X_i(\tau, k)| = \mu |X_j(\tau, k)|, \qquad \forall i, j \tag{B.8}$$

But this is not possible when more that one sources are active, because of the initial assumption that states that when several sources are active in the same zone, then the moduli of at least two observations are linearly independent.

Appendix C
Spherical harmonic domain analysis

The local DOA estimator used in 3D DRACOSS development was based on the estimation of the intensity vector using the spherical harmonic transform (SHT) of the acquired signal. Thus, in this section we provide basic theoretic elements of the spherical harmonic domain analysis field in order to facilitate the potential reader and also provide completeness to the text. Most of the presented material is taken from the textbooks on spherical harmonic analysis of B. Rafaely [97] and H. Teutch [114], thus for an in depth discussion please refer to the aforementioned books.

C.1 The acoustic wave equation

The acoustic wave equation in spherical coordinates is formulated as [114]:

$$\nabla^2_{\mathbf{r}} p(\mathbf{r},t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p(\mathbf{r},t) = 0, \tag{C.1}$$

where $p(\mathbf{r},t)$ expresses the sound pressure as a function of time (t) and space ($\mathbf{r}$) and c is the speed of sound. The position vector $\mathbf{r}$ is defined as $\mathbf{r} = (r,\theta,\varphi)$ in the spherical coordinate system with r being the radial distance from the origin, θ denotes the elevation angle, measured from the z-axis downwards, $\theta \in [0,\pi]$ and φ denotes the azimuthal angle, measured from the x-axis towards the y-axis and defined in $[0,2\pi)$ (see also Fig. C.1).

Considering steady-state conditions, by applying the temporal Fourier transform to Eq. (C.1), the homogeneous Helmholtz equation is obtained in spherical coordinates:

$$\nabla^2_{\mathbf{r}} p(k,\mathbf{r}) + k^2 p(k,\mathbf{r}) = 0, \tag{C.2}$$

where $k = \omega/c$ denotes the wavenumber and ω is the temporal radial frequency.

The inhomogeneous Helmholtz equation for a point source at $\mathbf{r}_0$ and an observation point at $\mathbf{r}$ is then

$$\nabla^2_{\mathbf{r}} p(k,\mathbf{r}|\mathbf{r}_0) + k^2 p(k,\mathbf{r}|\mathbf{r}_0) = -\delta(\mathbf{r}-\mathbf{r}_0), \tag{C.3}$$

where $\delta(\mathbf{r} - \mathbf{r}_0)$ denotes a three-dimensional Dirac delta function, representing a point source located at $\mathbf{r}_0$.

Solutions to the homogeneous and inhomogeneous Helmholtz equations can be formulated with Bessel and Hankel functions (Section C.4) and combinations of them. Specific solutions in the case of a sound-filed comprised by plane waves are presented in Section C.5.

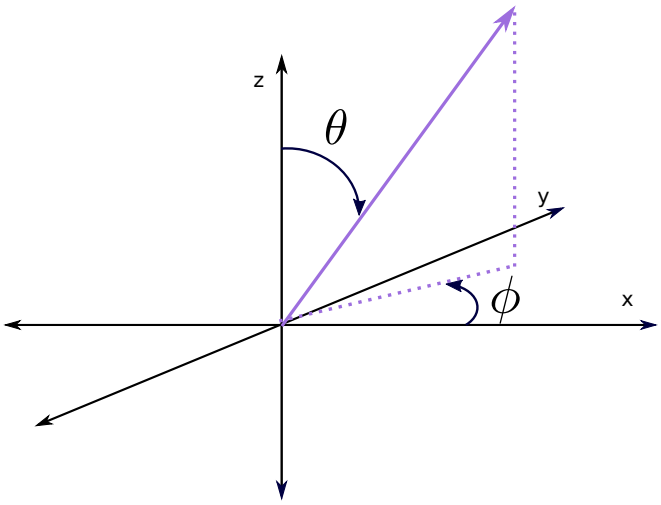


Figure C.1: The spherical coordinate system in relation with the standard Cartesian coordinate system.

C.2 Spherical Fourier transform

Consider a square integrable function $f(\theta, \varphi)$ on the unit sphere. The function $f(\theta, \varphi)$ can be represented as a infinite weighted summation of spherical harmonic functions (Section C.3) as:

$$f(\theta, \varphi) = \sum_{n=0}^{\infty} \sum_{m=-l}^{n} f_{lm} Y^{lm}(\theta, \varphi). \tag{C.4}$$

The terms f_{lm} are the weights and they represent the spherical Fourier transform (SFT) coefficients for the function $f(\theta, \varphi)$. The inverse of the SFT, i.e., the formula for the estimation of the weights f_{lm}, is:

$$f_{lm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) \left[Y_{lm}(\theta, \varphi)\right]^* \sin\theta d\theta d\varphi. \tag{C.5}$$

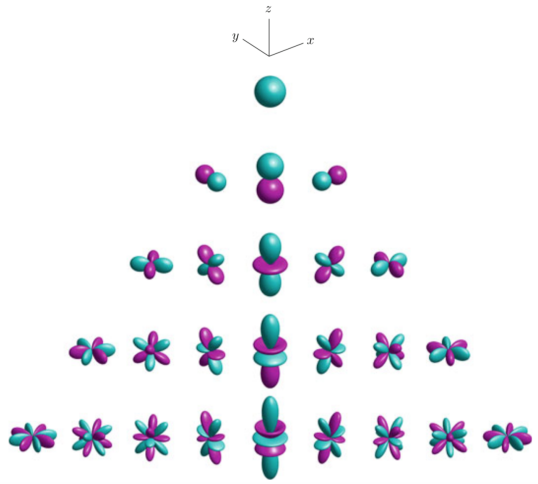


Figure C.2: Balloon plots of the imaginary (left) and real (right) parts of the spherical harmonic functions up to fourth order [97].

The basis functions for the SFT are the spherical harmonic functions. This is why the SFT is also referred as spherical harmonic transform (SHT).

C.3 Spherical Harmonic functions

The spherical harmonic functions (SHFs) are defined as:

$$Y_{lm}(\theta,\varphi) \equiv \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(\cos\theta)e^{im\varphi}, \tag{C.6}$$

where l denotes the order and m denotes the degree, with $-l \leq m \leq l$, $P_{lm}(\cdot)$ is the associated Legendre function and $(\cdot)!$ denotes the factorial function. Fig. C.2 shows the imaginary (left side) and the real (right side) parts of the SHFs for orders up to $l = 4$.

Some important and frequently used properties of the SHFs are:

- Orthogonality of the spherical harmonics: The spherical harmonic functions are orthogonal over the sphere surface, i.e.,

$$\int_0^{2\pi}\int_0^{\pi}[Y_{lm}(\theta,\varphi)]^* Y_{l'm'}(\theta,\varphi)\sin(\theta)d\theta d\varphi = \delta(l-l')\delta(m-m'), \tag{C.7}$$

  where $\delta(\cdot)$ is the Kronecker delta function.

- The completeness of the spherical harmonics: According to the completeness property

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{n}\left[Y_{lm}(\theta,\varphi)\right]^{*}Y_{lm}(\theta',\varphi') = \delta(\cos\theta-\cos\theta')\delta(\varphi-\varphi'), \tag{C.8}$$

  where $\delta(\cos\theta-\cos\theta')\delta(\varphi-\varphi')$ is the Kronecker delta function on the sphere.

- The spherical harmonics addition theorem: The addition theorem relates to the completeness theorem and it is formulated as:

$$\sum_{m=-l}^{n}\left[Y_{lm}(\theta,\varphi)\right]^{*}Y_{lm}(\theta',\varphi') = \frac{2l+1}{4\pi}P_l(\cos\Theta), \tag{C.9}$$

  where Θ is the angle between (θ,φ) and (θ',φ')

C.4 Spherical Bessel and Hankel functions

Spherical Bessel functions, along with spherical Neumann functions and spherical Hankel functions, are solutions to the differential equation

$$z^2\frac{\partial^2 w}{\partial z^2}+2z\frac{\partial w}{\partial z}+\Big(z^2-l(l+1)\Big)w=0. \tag{C.10}$$

Spherical Bessel functions and spherical Hankel functions are met in the sound pressure expansion expression (presented later in Eqs. (C.16) and (C.20)). Thus, it is of our interest to observe their behavior in Figs. C.3 and C.4, where we can see that the Bessel functions exhibit several nulling points, in contrast with the Hankel functions which diverge towards the origin and decay similarly for all orders l as x increases.

C.5 Plane wave decomposition

Since our interest focuses on far-field conditions, our focus is on plane waves instead of point sources. A unit-amplitude plane wave can be considered as a point source at infinity, that is, $|\mathbf{r}_0|\rightarrow\infty$. Taken this into account, a solution to the inhomogeneous Helmholtz equation (Eq. (1.3)) is formulated as:

$$p(k,\mathbf{r}|\mathbf{r}_0) = \frac{e^{ikr_0}}{r_0}e^{-i\mathbf{k}}, \tag{C.11}$$

where

$$k = \frac{2\pi}{\lambda}\begin{bmatrix}\sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta\end{bmatrix} \tag{C.12}$$

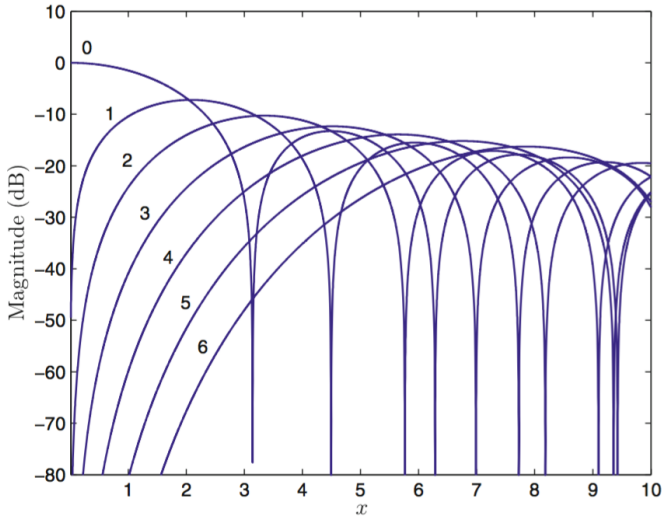


Figure C.3: Magnitude of the spherical Bessel function $|j_l(x)|$ for orders $l = 0, 1, \ldots, 6$ [97].

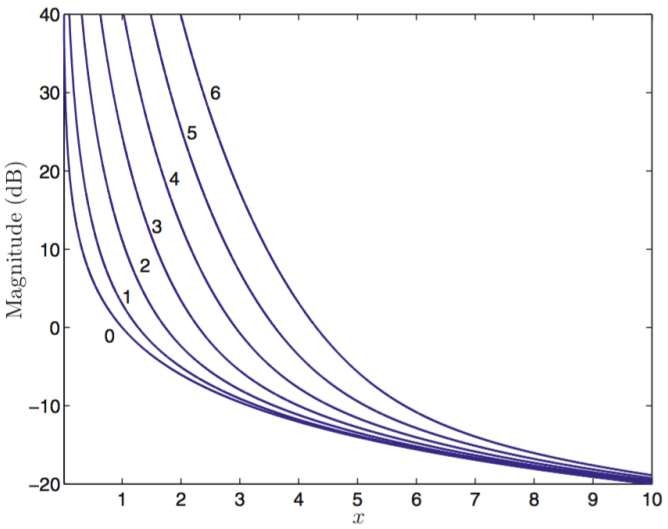


Figure C.4: Magnitude of the spherical Hankel function $|h_l(x)|$ for orders $l = 0, 1, \ldots, 6$ [97].

is the wavenumber vector which indicates the speed and direction of the wave propagation.

The solution can be expanded into spherical harmonic and spherical Bessel functions

as:

$$p(k,r,\theta,\varphi) = e^{-i\mathbf{k}^{\mathrm{T}}\mathbf{r}} = \sum_{n=0}^{\infty}\sum_{m=-l}^{l} 4\pi i^l j_l(kr) Y_{lm}(\theta_0,\varphi_0)^* Y_{lm}(\theta,\varphi), \tag{C.13}$$

where $i^2 = -1$, $(\cdot)^*$ stands for the conjugation operation, θ_0 and φ_0 denote the elevation and azimuth angle of the origin of the plane wave respectively (see also Fig. C.1), $j_n(\cdot)$ is the spherical Bessel function of order l (Section C.4) and $Y_{lm}(\theta,\varphi)$ denotes the spherical harmonic function of order l and degree m with $l \geq m$ (see Section C.3).

Taking the SFT of the left-hand side of Eq. (C.13) we get (see also Eq. (C.4))

$$p(k,r,\theta,\varphi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} p_{lm} Y_{lm}(\theta,\varphi), \tag{C.14}$$

and by comparing Eqs. (1.14) and (C.13) we find the spherical harmonic coefficients p_{lm} of a single amplitude plane wave as:

$$p_{lm} = 4\pi i^l j_l(kr) Y_{lm}(\theta_0,\varphi_0)^*. \tag{C.15}$$

When the sound field is composed by multiple plane waves of direction amplitude density $a(k,\theta_k,\varphi_k)$, the sound pressure is analyzed as:

$$p(k,r,\theta,\varphi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{n} 4\pi i^l j_l(kr) a_{lm} Y_{lm}(\theta,\varphi), \tag{C.16}$$

and analogously to the single plane wave, by evaluating the sound pressure on the surface of a sphere we take its SFT as

$$p_{lm} = 4\pi i^l j_l(kr) a_{lm}, \tag{C.17}$$

where a_{lm} are the spherical Fourier coefficients of $a(k,\theta_k,\varphi_k)$.

C.5.1 Soundfield decomposition around a rigid scatterer

In the preceding sections we referred to the soundfield decomposition around a sphere assuming a free space, i.e., we have referred to the so-called open sphere configuration. Of particular interest is the expression of the sound pressure of the incoming sound field around a rigid sphere. This is because, in practice, sound fields are measured with microphones mounted on such rigid spheres. In this case the soundfield on the rigid sphere can be expressed as

$$p(k,r,\theta,\varphi)_{\mathrm{tot}} = p(k,r,\theta,\varphi)_{\mathrm{in}} + p(k,r,\theta,\varphi)_{\mathrm{scat}}, \tag{C.18}$$

where $p(k, r, \theta, \varphi)_{\text{in}}$ and $p(k, r, \theta, \varphi)_{\text{scat}}$ express the incoming and the scattered soundfiled from the surface of the sphere respectively.

The expansion of the incoming soundfield, assuming it is comprised by plane waves of directional amplitude density $a(k, \theta_k, \varphi_k)$, was presented in Eq. (C.16). The scattered field can be expanded in a spherical harmonic series as:

$$p(k, r, \theta, \varphi)_{\text{scat}} = \sum_{l=0}^{\infty} \sum_{m=-l}^{n} -4\pi i^l \frac{j_l'(kr_a) h_l(kr)}{h_l'(kr_a)} a_{lm} Y_{lm}(\theta, \varphi), \tag{C.19}$$

where r_a is the array of the rigid spherical scatterer, $h_l(\cdot)$ denotes the Hankel function of the second kind, and $j_l'()$ and $h_l'()$ are the first derivatives of the Bessel and Hankel functions, leading to a total sound pressure around the rigid scatterer defined as

$$p(k, r, \theta, \varphi)_{\text{tot}} = \sum_{l=0}^{\infty} \sum_{m=-l}^{n} 4\pi i^l \left(j_l(kr) - \frac{j_l'(kr_a) h_l(kr)}{h_l'(kr_a)} \right) a_{lm} Y_{lm}(\theta, \varphi). \tag{C.20}$$

By defining an equalization term as:

$$b_l(kr) = \begin{cases} 4\pi i^l j_l(kr), & \text{open sphere} \\ 4\pi i^l \left(j_l(kr) - \frac{j_l'(kr_a) h_l(kr)}{h_l'(kr_a)} \right), & \text{rigid sphere} \end{cases}$$

we can write the spherical harmonic decomposition and the spherical harmonic signals in a more compact form as:

$$p(k, r, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{n} b_l(kr) a_{lm} Y_{lm}(\theta, \varphi), \text{ and} \tag{C.21}$$

$$p_{lm} = b_l(kr) a_{lm} \tag{C.22}$$

It is worth observing the behavior of the magnitude of $b_l(kr)$ for a rigid scatterer which is depicted in Fig. C.5. In the case of the spherical scatterer the equalization term does not have any zeros in contrast to the behavior of the Bessel function (see also Fig. C.3), which is very important when divisions are involved.

C.6 Sampling schemes on a sphere and spherical microphone arrays

In practical situations we cannot measure the sound pressure of a soundfield on every point on a sphere, thus we sample it using sensors located on a rigid or open structure constituting a rigid or open spherical microphone array. Sampling a sphere and placing the microphones is not as straight forward as for example with circular microphone arrays.

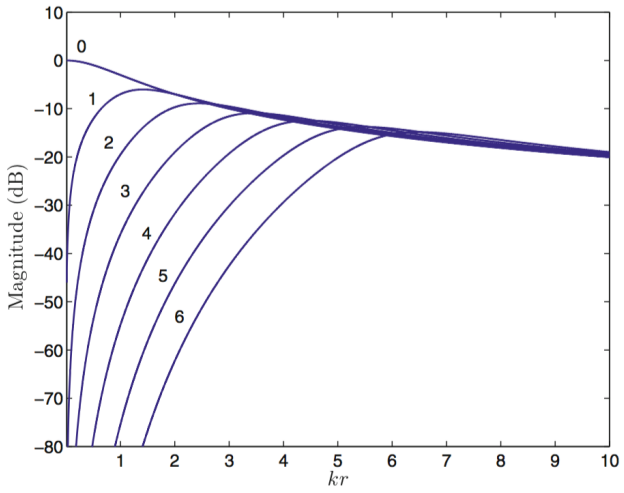


Figure C.5: The magnitude of the equalization term $|b_l|/4\pi$ for a rigid spherical scatterer with $r = r_a$ and $n = 0, 1, \ldots, 6$.

First, the number of sampling points has to be decided such that an accurate approximation of the SFT of the measured sound field can be obtained. In addition and in relation to the number of sensors, the sampling scheme, i.e., the positions of the microphones on the surface of the sphere, has to be designed.

Sampling theorems for functions on the sphere require the functions to be order-limited or represented by a finite number of spherical harmonics. That is, for the sound pressure evaluated on a sphere, we should be able to represent it by a finite summation of spherical harmonics as:

$$p(k, r, \theta, \varphi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} 4\pi i^l j_l(kr) a_{lm} Y_m(\theta, \varphi), \tag{C.23}$$

introducing though truncation errors, which according to the value of N can be significant or insignificant. Later, in Section C.6.4 we will briefly describe the nature of such errors.

If this approximation is feasible, then according to Eq. (C.5), the spherical harmonic coefficients of the measured soundfield $p(k, r, \theta, \varphi)$ are

$$p_{lm} = \int_0^{2\pi} \int_0^{\pi} p(k, r, \theta, \varphi) \left[Y_{lm}(\theta, \varphi)\right]^* \sin\theta d\theta d\varphi \tag{C.24}$$

which have to be approximated by a finite summation:

$$p_{lm} = \sum_{q=1}^{Q} g_q p(k, r, \theta_q, \varphi_q) \left[Y_{lm}(\theta_q, \varphi_q) \right]^*, \tag{C.25}$$

where Q is the number of microphones comprising the array, g_q are weights that assure the approximation is accurate. A basic property of an ideal sampling scheme is to maintain the orthogonality of the spherical harmonics (recall Eq. (C.7)), which means that the weights g_q, the number of the sensors Q as well as their positioning $\left(\theta_q, \varphi_q\right)$ has to be estimated.

Research has come up with a vast variety of different sampling schemes for a sphere that may or may not provide equidistant coverage of the sphere. The sampling schemes we will refer to are the (a) equal-angle sampling, (b) the Gaussian sampling and (c) the uniform and almost uniform schemes.

C.6.1 Equal-angle sampling

In equal-angle sampling the sampling points are placed on uniformly spaced angular positions along the elevation θ and azimuth φ. The scheme requires $(2L+2)(2L+2) = 4(L+1)^2$ sampling points in order to achieve a maximum order of reconstruction equal to L. Even though the sampling points are taken uniformly along the azimuth and elevation, they are not uniformly distributed on the surface of the sphere leading to a more dense distribution close to the poles. The weights g_q that guarantee the orthogonality of the spherical harmonics are independent of the azimuth for the equal-angle sampling, thus $(2L+2)$ weight values have to be estimated in total, provided by:

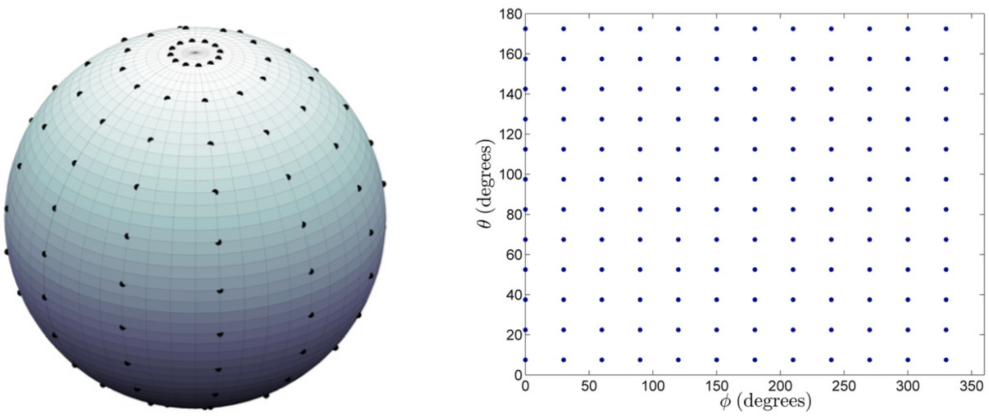
$$g_q = \frac{2\pi}{(L+1)^2} \sin(\theta_q) \sum_{q'=0}^{L} \frac{1}{2q'+1} \sin\left((2q'+1)\,\theta_q\right), \quad 0 \le q \le 2L+1 \tag{C.26}$$

In Figure C.6 we can see an example of an equal-angle sampling distribution with 144 sampling points on the surface of a unit sphere.

C.6.2 Gaussian sampling

Similar to the equal-angle sampling scheme, the Gaussian scheme[1] samples the azimuth in $2(L+1)$ equal-angle samples, but for the elevation it requires only $(L+1)$ samples, almost equally spaced. This leads to a total of $2(L+1)^2$ of sampling points, which is half the number required for equal-angle sampling in order to achieve order L of reconstruction. As with equal-sampling though the sampling is more dense close to the poles. The weights of the Gaussian sampling are evaluated as:

[1]The naming of the scheme comes from Gauss who, in 1814, answered the related question on how to discretize the Legendre transform with the minimum error and the minimum number of points. For more information, please refer to [114].

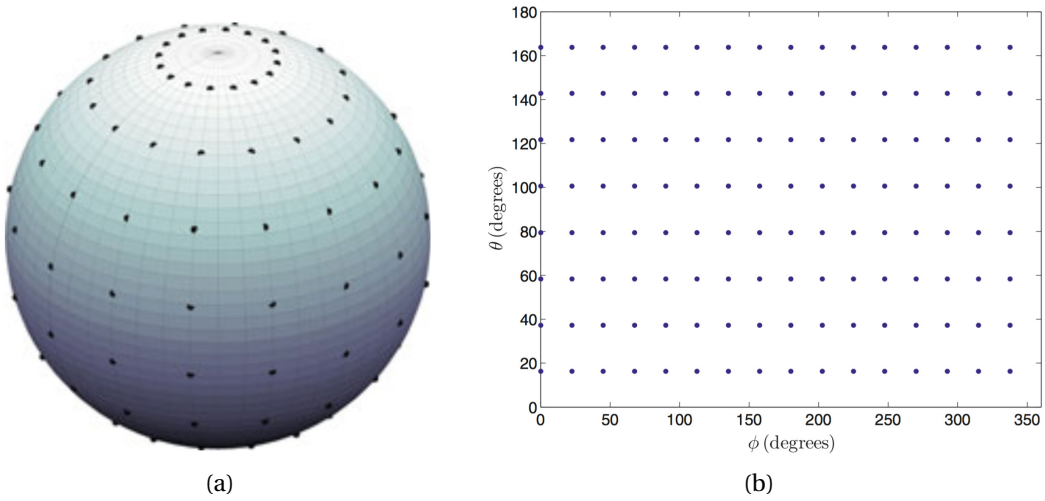


(a) (b)

Figure C.6: Equal-angle sampling with $L = 5$ and 144 sampling points in total, illustrated: C.6a on the surface of a unit sphere and C.6b over the $\theta\varphi$ plane.

$$g_q = \frac{\pi}{L+1} \frac{2\left(1-\cos^2\theta_q\right)}{(L+2)^2 P_{L+2}^2\left(\cos\theta_q\right)}, \quad 0 \le q \le L. \tag{C.27}$$

The weights and the positioning of the microphones can also be found in tables [114]. In Fig. C.7 we see an example of a Gaussian sampling scheme.



(a) (b)

Figure C.7: Gaussian sampling with $L = 7$ and 128 sampling points in total, illustrated: C.7a on the surface of a unit sphere and C.7b over the $\theta\varphi$ plane.

C.6.3 Uniform and almost uniform sampling

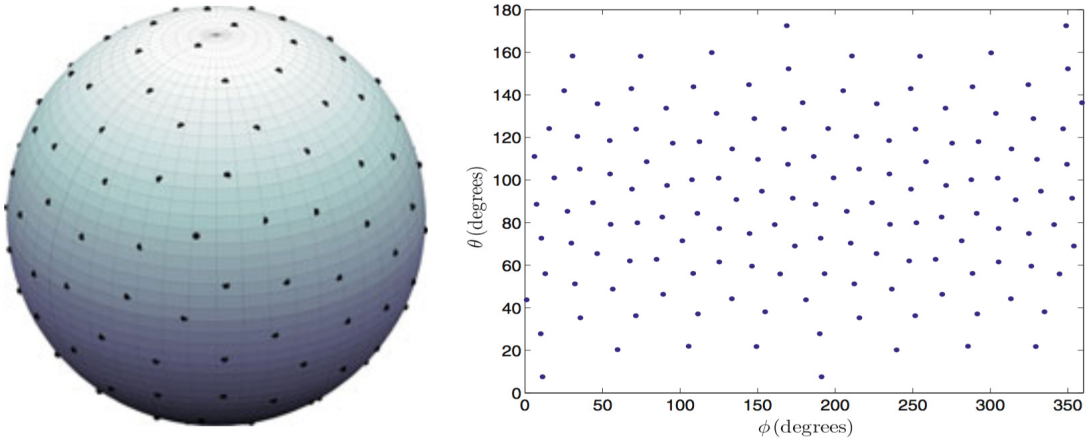
The equal-angle and Gaussian sampling schemes have a (almost) uniform distribution of the sampling points along the azimuth φ and the elevation θ, however the distributions are not uniform on the sphere, having more sampling points around the poles. An attempt to distribute the samples more evenly to the surface of the sphere leads to the five convex regular polyhedra (the Platonic solids) [8], the vertices of which can be taken as sampling points on a sphere, providing four (tetrahedron), six (octahedron), eight (hexahedron), twelve (icosahedron), and 20 (dodekahedron) sampling points. Taking sampling points at the vertices of the platonic solids satisfies the quadrature relation:

$$\int_0^{2\pi}\int_0^{\pi} c(\theta,\varphi)\sin\theta d\theta\, d\varphi = \frac{4\pi}{Q}\sum_{q=1}^{Q} c(\theta_q,\varphi_q). \tag{C.28}$$

If we replace the function $c(\theta,\varphi)$ with $f(\theta,\varphi)\,[Y_{lm}(\theta,\varphi)]^*$ in the above quadrature equation, the left-hand side of the equation becomes the SFT (eq. (C.4)), such that we can obtain the spherical harmonic coefficients of the function $f(\theta,\varphi)$ as :

$$f_{lm} = \int_0^{2\pi}\int_0^{\pi} f(\theta,\varphi)\,[Y_{lm}(\theta,\varphi)]^*\sin\theta d\theta\, d\varphi = \frac{4\pi}{Q}\sum_{q=1}^{Q} f(\theta,\varphi)\,[Y_{lm}(\theta,\varphi)]^*\,. \tag{C.29}$$

In the above equation we identify the weights $g_q = \frac{4\pi}{Q}$, which are constant, one of the benefits of the uniform sampling schemes. The problem however is the tight list of options for the number of sensors. The maximum number is 20 sensors (dodekahedron) which leads to a maximum order of reconstruction equal to $L = 2$. This stems from the t-design order specified for each platonic solid [97]. This is why researchers soon started looking for methods to enable the uniform or almost-uniform placement of more sampling points on the surface of a sphere, with Harding and Sloan extending the t-designs into a larger set of sampling configurations which satisfy Eq. (C.29) and retain the convenient constant sampling weights [49, 50]. An example of such an almost-uniform distribution is shown in Fig. C.8. This design achieves a reconstruction order of $L = 8$ utilizing 144 sampling points. Note that in order to achieve the same reconstruction order with an equal-angular sampling scheme one would require 324 sampling points while for the Gaussian scheme the corresponding number would be 162 sensors. As a general rule, we need at least $(L+1)^2$ sampling points for uniform schemes in order to achieve a reconstruction order equal to L [99].



(a) (b)

Figure C.8: Nearly-uniform sampling with $L = 8$ and 144 sampling points in total, illustrated: C.8a on the surface of a unit sphere and C.8b over the $\theta\varphi$ plane.

C.6.4 Spatial aliasing

The sampling methods that we presented in the previous sections guarantee zero or negligible error for order limited functions. However, acoustic sound fields, such as those produced by plane waves, are not order-limited on a sphere, since they are represented by infinite series of spherical harmonics. In practical implementations, though, the infinite summation is replaced by a finite one, introducing aliasing.

However the magnitude of the spherical harmonics coefficients of the sound pressure function is proportional to the magnitude of the Bessel function as indicated in Eq. (C.15). This means that the magnitude of the signals p_{lm} decays rapidly for $l > kr$, thus the error is expected to be negligible if the operating frequency range of the array satisfies $kr \ll N$ [99]. This is illustrated in Fig. C.3 and more explicitly in Fig. C.9 for $kr = 8$ and $kr = 16$.

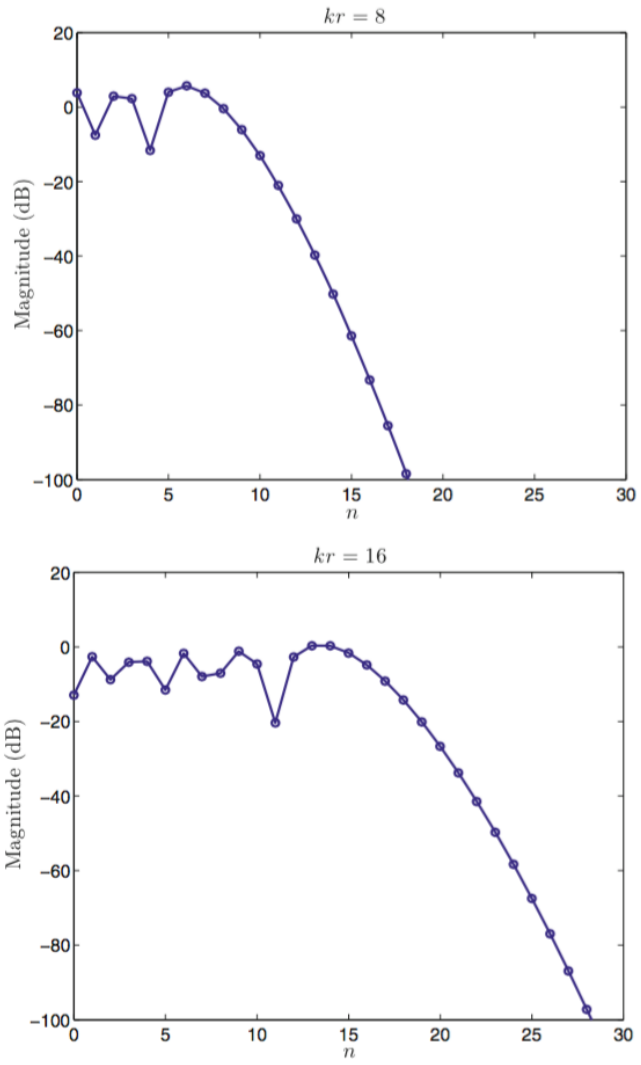


Figure C.9: The magnitude of the normalized spherical Bessel function $|4\pi i^l j_l(kr)|$ for $kr = 8$ and $kr = 16$.

Appendix D
Additional source counting methods

In this section we present our two alternative source counting methods, namely a peak search approach and a linear predictive coding approach. Both methods were originally presented in [91].

Peak Search

In order to estimate the number of sources we perform a peak search of the smoothed histogram at each time frame in the following manner

a) We assume that there is always at least one active source in a block of estimates, i.e. we always expect to find at least one peak at the histogram. So we set $i = 1$, where i corresponds to a counter of the peaks assigned to sources so far. We also set $u_i = u_1 = \arg\max h_i$, i.e., the histogram bin which corresponds to the highest peak of the smoothed histogram. Finally, we set the threshold $\gamma_{i+1} = \max\{h(u_i)/2, \gamma_{\text{static}}\}$, where γ_{static} is a user-defined static threshold.

b) We locate the next highest peak in the smoothed histogram, $h(u_{i+1})$. If the following three conditions are simultaneously satisfied:

$$h(u_{i+1}) \geq \gamma_{i+1} \tag{D.1}$$

$$u_{i+1} \notin \left[u_j - u_w, u_j + u_w\right], \qquad \forall u_j \tag{D.2}$$

$$j < (i+1), \tag{D.3}$$

then we proceed to the detection of the next peak, i.e., $i = i+1$ and $\gamma_{i+1} = \max\{h(u_i)/2, \gamma_{\text{static}}\}$. u_w is the minimum offset between neighbouring sources. (D.1) guaranties that the next located histogram peak is higher than the updated threshold γ_{i_s+1}. (D.2) and (D.3) guar-

antee that the next located peak is not in the close neighbourhood of an already located peak with $j_s = 1, \ldots i_s$ and u_{j_s} all the previously identified source peaks.

c) We stop when a peak in the histogram fails to satisfy the threshold γ_{i_s+1} or if the upper threshold $N_{S_{MAX}}$ is reached. The estimated number of sources is $\hat{N}_S = i$.

In Figure D.1 we can see how the Peak Search method is applied to a smoothed histogram where four sources are active. The black areas indicate the bins around a tracked peak of the histogram that are excluded as candidate source indicators.

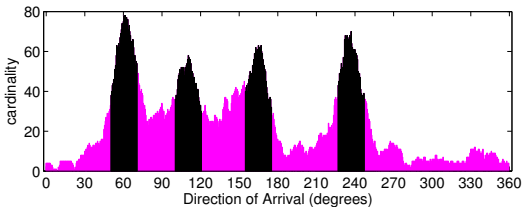


Figure D.1: Peak Search for source counting. The black areas indicate the bins around a tracked peak of the histogram that are excluded as candidate source indicators.

Linear Predictive Coding

Linear predictive coding (LPC) coefficients are widely used to provide an all-pole smoothed spectral envelope of speech and audio signals [73]. This inspired us to apply LPC to the smoothed histogram of estimates to emphasize the peaks and suppress any noisy areas. Thus, the estimated LPC envelope coincides with the envelope of the histogram. We get our estimate of $\hat{N}_S$ sources by counting the local maxima in the LPC envelope with the constraint that $\hat{N}_S \leq N_{S_{MAX}}$, where $N_{S_{MAX}}$ is user defined as in Section 4.1.4. In our estimation, we exclude peaks that are closer than u_w, as a minimum offset between neighboring sources.

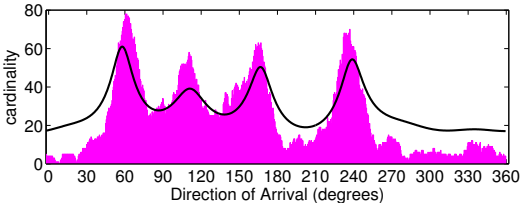


Figure D.2: LPC for source counting. The black curve corresponds to the LPC estimated envelope of the histogram.

A key parameter of this approach is the order of LPC. We want to avoid a very high order that will over-fit our histogram of estimates, in turn leading to an over-estimation of the true number of sources. On the other hand, the use of a very low order risks the detection of less dominant sources (i.e., sources with less estimates in the histogram, thus lower peaks). In order to decide on an optimum LPC order, we tested a wide range of values and chose the one that gave the best results in a wide range of simulated scenarios. In Figure D.2 we plot an example LPC envelope with order 16, along with the smoothed histogram.