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

Multiple Sound Source Location Estimation and Counting in Wireless Acoustic Sensor Networks

by

Anastasios Alexandridis

PhD Dissertation

Presented

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

Heraklion, January 2018

UNIVERSITY OF CRETE

DEPARTMENT OF COMPUTER SCIENCE

Multiple Sound Source Location Estimation and Counting in Wireless Acoustic Sensor Networks

PhD Dissertation Presented

by **Anastasios Alexandridis**

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

APPROVED BY:

Author: Anastasios Alexandridis

Supervisor: Athanasios Mouchtaris, Associate Professor, University of Crete

Committee Member: Panagiotis Tsakalides, Professor, University of Crete

Committee Member: Marc Moonen, Professor, KU Leuven

Committee Member: Yannis Stylianou, Professor, University of Crete

Committee Member: Maximo Cobos, Associate Professor, Universitat de Valencia

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

Committee Member: Emmanuel Vincent, Senior Research Scientist, Inria Nancy - Grand Est, France

Department Chairman: Bilas Angelos, Professor, University of Crete

Heraklion, January 2018

To my beloved grandmother, Maria

Στην αγαπημένη μου γιαγιά, Μαρία

Acknowledgments

Now, after 4.5 years, that this journey towards the completion of my PhD dissertation has come to an end, it is time to thank all those who contributed both to the technical aspects as well as with their moral support. First of all, my supervisor Prof. Athanasios Mouchtaris who showed belief in me, for all our constructive meetings and for the excellent and pleasant cooperation.

I would also like to thank the members of my dissertation committee for their valuable comments and questions and for the high-quality evaluation of this dissertation. More specifically, I would like to thank Prof. Panagiotis Tsakalides and Prof. Yannis Stylianou (Computer Science Department – University of Crete), Prof. Petros Maragos (School of Electrical and Computer Engineering – National Technical University of Athens), Prof. Marc Moonen (Electrical Engineering Department – K.U. Leuven), Prof. Maximo Cobos (Computer Science Department - Universitat de València), and Dr. Emmanuel Vincent (Inria Nancy – Grand Est., France).

Special thanks also go to my excellent colleague and beloved friend Dr. Anthony Griffin. His passion for research has always been an inspiration for me. Our cooperation during my MSc studies and during the first years of my PhD, his advice and comments had a significant contribution to this dissertation.

I would like to acknowledge the Institute of Computer Science of the Foundation for Research and Technology – Hellas (FORTH-ICS) for providing financial support and all the necessary equipment during this work.

A dedicated thank you to all my colleagues at the Signal Processing Laboratory and the Telecommunications and Networks Laboratory of FORTH-ICS for the very pleasant and friendly environment that they created at work. A special thank you to my colleague and friend Despoina Pavlidi for her support from the first moment that I joined our group at FORTH.

I would also like to thank all my friends for their moral support throughout all the years of my PhD studies. Maria, Giorgo, Paulo, Olina, Serafeim, Klairy, Sofia thank you all for your support and advice.

Last but not least I would like to thank my family for their support and help towards achieving my goals.

Ευχαριστίες

Τώρα, μετά το πέρας 4.5 χρόνων, που το ταξίδι για την ολοκλήρωση της διδακτορικής διατριβής έφτασε στο τέλος του, ήρθε η στιγμή να ευχαριστήσω τα άτομα που συνέβαλλαν με την στήριξη τους τόσο στον τεχνικό όσο και στον ηθικό τομέα. Πρωτίστως τον επόπτη της παρούσας διατριβής, αναπληρωτή Καθηγητή του τμήματος Επιστήμης Υπολογιστών κ. Αθανάσιο Μουχτάρη που έδειξε εμπιστοσύνη σε εμένα και μέσα από όλες τις συναντήσεις μας σε ένα πλαίσιο εξαίρετης και ευχάριστης συνεργασίας βοήθησε ώστε να γίνει πραγματικότητα αυτή η διατριβή.

Θα ήθελα επίσης να ευχαριστήσω τα υπόλοιπα μέλη της επταμελούς εξεταστικής επιτροπής που με τα πολύτιμα σχόλια τους διετέλεσαν σημαντικό ρόλο στην υψηλής ποιότητας αξιολόγηση της παρούσας διατριβής. Συγκεκριμένα, τους καθηγητές κ. Παναγιώτη Τσακαλίδη και κ. Ιωάννη Στυλιανού (τμήμα Επιστήμης Υπολογιστών – Πανεπιστήμιο Κρήτης), τον καθηγητή κ. Πέτρο Μαραγκό (Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών – Εθνικό Μετσόβιο Πολυτεχνείο), τον καθηγητή κ. Marc Moonen (Electrical Engineering Department – K.U. Leuven) , τον αναπληρωτή καθηγητή κ. Maximo Cobos (Computer Science Department – Universitat de Valencia) και τον Δρ. Emmanuel Vincent (Inria Nancy – Grand Est, France).

Φυσικά, δεν θα μπορούσα να ξεχάσω τον εξαιρετικό συνεργάτη και αγαπητό φίλο Δρ. Anthony Griffin που κατάφερε να μου μεταλαμπαδεύσει το πάθος του για την περιοχή της επεξεργασίας ήχου και που η συνεργασία μας κατά τη διάρκεια της μεταπτυχιακής και των πρώτων χρόνων της διδακτορικής μου διατριβής αποτέλεσε για μένα πηγή έμπνευσης.

Θα ήθελα επίσης να ευχαριστήσω το Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας (ΙΤΕ-ΙΠ) για την οικονομική ενίσχυση και την παροχή όλου του απαραίτητου εξοπλισμού κατά την εκπόνηση της παρούσας διατριβής.

Ένα μεγάλο ευχαριστώ αξίζει σε όλους τους συνεργάτες του Εργαστηρίου Επεξεργασίας Σήματος και του Εργαστηρίου Τηλεπικοινωνιών και Δικτύων του Ι.Τ.Ε. για το υπέροχο κλίμα συνεργασίας που δημιούργησαν όλα αυτά τα χρόνια. Ένα ιδιαίτερο ευχαριστώ στη συνεργάτη, φίλη και συνοδοιπόρο σε αυτό το ταξίδι της διδακτορικής διατριβής Δέσποινα Παυλίδη για τη στήριξη της από την πρώτη στιγμή που ξεκίνησα σαν μέλος της ομάδας μας στο Ι.Τ.Ε.

Ένα μεγάλο ευχαριστώ και σε όλους μου τους φίλους για την ηθική στήριξη που προσέφεραν όλα αυτά τα χρόνια. Μαρία, Γιώργο, Παύλο, Ολίνα, Σεραφείμ, Κλαίρη, Σοφία σας ευχαριστώ για την στήριξη και τις συμβουλές σας.

Τέλος, ξεχωριστές ευχαριστίες αξίζουν στους γονείς μου για την στήριξη και τη συνεχή παρακίνηση τους για την επίτευξη των στόχων μου.

Abstract

Wireless acoustic sensor networks (WASNs) represent a new paradigm for acoustic signal acquisition. Multiple acoustic nodes that feature processing and communication capabilities are distributed in the environment where typically multiple sound sources are active. In such a setup, inference of location information has always been an attractive research problem. Enabling machines to estimate the locations of the multiple simultaneously active sound sources from their acoustic emissions is crucial in many applications, such as wildlife monitoring and speech enhancement for robust signal acquisition.

Throughout the years, different localization methods have been proposed with the everlasting goal to achieve the lowest possible localization error. Although significant steps have been made towards this direction, another unexplored field concerns the practical limitations posed by the sensor network that restrict the application of such methods to real-life WASNs. Such limitations include the limited processing power and battery life of the nodes, the communication bandwidth that has to be attained at low levels, the real-time requirements and synchronization issues. In this thesis, we consider the problem of multiple source localization and we investigate the development of methods that not only achieve high accuracy in realistic scenarios but also attain low communication bandwidth, tolerate unsynchronized input and are computationally efficient to facilitate their application in real-life WASNs.

We consider a WASN where each node is a microphone array that estimates and transmits information related to the directions of arrival (DOAs) of the active sound sources. Such a scheme attains very low communication bandwidth, as only the DOAs need to be transmitted. Moreover, DOA-based localization methods can tolerate unsynchronized input, thus the acoustic signals need not be perfectly synchronized. We first focus on the single source case and propose a computationally efficient non-linear least squares estimator that can accurately estimate the source's location using an iterative grid-based approach. We then proceed to the multiple sources case, assuming that the number of sources is known. In this case, a core problem for DOA-based approaches is that the fusion center that receives the multiple DOA estimates from the nodes cannot know to which source each DOA belongs. This is known as the data-association problem. To address this problem we propose two solutions: the first concerns the extension of our grid-based approach to multiple sources and the second utilizes additional information, apart from the DOA estimates, in order to find the correct association of DOAs from the nodes to the sources. We then relax the assumption of known number of sources and propose another method that can jointly perform source counting and location estimation. Our method is based on clustering narrowband per-frequency location estimates which are inferred using narrowband per-frequency DOA estimates from the nodes.

Since a determinant factor that affects localization performance is the accuracy in which the DOA estimates are obtained, we also investigate how we can improve DOA estimation performance and we propose a methodology to infer more accurate and reliable DOA estimates. Finally, we investigate the potential use of location information to audio processing applications. We provide two examples of how location information can be used for spatial audio capturing and for the design of beamformers that leverage location information in order to estimate the steering vector of the target source. Our preliminary results reveal the potential of location-based approaches to provide improved performance.

Keywords: localization, direction-of-arrival estimates, wireless acoustic sensor networks, location information, microphone arrays

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

Περίληψη

Τα Ασύρματα Δίκτυα Ακουστικών Αισθητήρων αποτελούν μια νέα τεχνική λήψης ακουστικών σημάτων. Πολλαπλοί ακουστικοί αισθητήρες με επεξεργαστική ισχύ και ικανότητες μετάδοσης πληροφορίας διανέμονται σε ένα περιβάλλον όπου τυπικά πολλές ηχητικές πηγές είναι ενεργές. Σε τέτοιες περιπτώσεις, η εκτίμηση της θέσης των πηγών στο χώρο ήταν πάντα ένα ενδιαφέρον ερευνητικό πρόβλημα. Η πληροφορία της θέσης των πολλαπλών ενεργών ηχητικών πηγών είναι σημαντική σε μια πληθώρα εφαρμογών όπως η παρακολούθηση της άγριας πανίδας και η βελτίωση ποιότητας για την εύρωστη λήψη ηχητικών σημάτων.

Με την πάροδο των χρόνων αναπτύχθηκαν διάφορες μέθοδοι εύρεσης θέσης με τελικό στόχο την επίτευξη του χαμηλότερου δυνατού σφάλματος. Ενώ έχει γίνει σημαντική πρόοδος προς αυτή την κατεύθυνση, ένας άλλος τομέας που δεν έχει ευρέως μελετηθεί αφορά τους πρακτικούς περιορισμούς που προκύπτουν από το δίκτυο αισθητήρων, οι οποίοι περιορίζουν την πρακτική εφαρμογή τέτοιων μεθόδων σε πραγματικά δίκτυα ακουστικών αισθητήρων. Τέτοιοι περιορισμοί αφορούν την περιορισμένη επεξεργαστική ισχύ των αισθητήρων, τις απαιτήσεις σε εύρος ζώνης που πρέπει να είναι χαμηλές, τις απαιτήσεις για εφαρμογές πραγματικού χρόνου και τα ζητήματα συγχρονισμού μεταξύ των ηχητικών σημάτων. Σε αυτή τη διατριβή, μελετάμε το πρόβλημα της εύρεσης θέσης πολλαπλών ταυτόχρονα ενεργών ηχητικών πηγών σε ένα δίκτυο ακουστικών αισθητήρων και ερευνούμε την ανάπτυξη μεθόδων εύρεσης θέσης που είναι ικανές όχι μόνο να επιτυγχάνουν υψηλή ακρίβεια σε ρεαλιστικά περιβάλλοντα, αλλά επίσης έχουν χαμηλές απαιτήσεις σε εύρος ζώνης, μπορούν να λειτουργήσουν με μη-συγχρονισμένη είσοδο και είναι υπολογιστικά αποτελεσματικές, ώστε να καθιστούν δυνατή την εφαρμογή τους σε πραγματικά δίκτυα ακουστικών αισθητήρων.

Θεωρούμε ένα ασύρματο δίκτυο ακουστικών αισθητήρων όπου ο κάθε κόμβος είναι μια συστοιχία μικροφώνων η οποία εκτιμά και μεταδίδει πληροφορία σχετικά με την κατεύθυνση άφιξης των ηχητικών σημάτων των ενεργών ηχητικών πηγών στο περιβάλλον. Αυτή η προσέγγιση επιτυγχάνει χαμηλές απαιτήσεις σε εύρος ζώνης, αφού αρκεί μόνο η μετάδοση των εκτιμήσεων των κατευθύνσεων άφιξης. Επιπλέον, οι τεχνικές εύρεσης θέσης που βασίζονται σε εκτιμήσεις κατευθύνσεων άφιξης μπορούν να λειτουργήσουν όταν τα ηχητικά σήματα μεταξύ των διάφορων κόμβων του δικτύου δεν είναι τέλεια συγχρονισμένα. Αρχικά επικεντρωνόμαστε στο πρόβλημα εκτίμησης θέσης μιας ενεργής ηχητικής πηγής και προτείνουμε έναν υπολογιστικά αποτελεσματικό μη-γραμμικό εκτιμητή θέσης που είναι ικανός να εντοπίσει τη θέση της πηγής με ακρίβεια χρησιμοποιώντας μια επαναληπτική μέθοδο βασισμένη σε πλέγμα. Έπειτα, ασχολούμαστε με την περίπτωση όπου πολλαπλές ηχητικές πηγές είναι ταυτόχρονα ενεργές, θεωρώντας ότι ο αριθμός τους είναι γνωστός. Το βασικό πρόβλημα που προκύπτει στην περίπτωση των πολλαπλών πηγών είναι ότι ο κεντρικός κόμβος που λαμβάνει τις πολλαπλές εκτιμήσεις κατευθύνσεων άφιξης δεν γνωρίζει σε ποια πηγή αντιστοιχούν. Το πρόβλημα αυτό είναι γνωστό ως πρόβλημα αντιστοίχησης δεδομένων (δατα-ασσοσιατιον προβλεμ). Για να επιλύσουμε αυτό το πρόβλημα προτείνουμε δύο προσεγγίσεις: η πρώτη αφορά την επέκταση της βασισμένης σε πλέγμα τεχνικής σε πολλαπλές πηγές και η δεύτερη χρησιμοποιεί επιπλέον πληροφορία (εκτός των κατευθύνσεων άφιξης) από τους αισθητήρες με σκοπό να βρεθεί η σωστή αντιστοίχηση των κατευθύνσεων άφιξης από τους κόμβους στις ηχητικές πηγές. Έπειτα, θεωρούμε ότι ο αριθμός των πηγών είναι επίσης άγνωστος και προτείνουμε μια μέθοδο ικανή να εκτιμήσει τον αριθμό των πηγών που είναι ενεργές στο περιβάλλον και τις θέσεις τους. Η μέθοδος μας βασίζεται στην ομαδοποίηση εκτιμήσεων θέσης που προκύπτουν για κάθε συχνότητα των ηχητικών σημάτων και έχουν εκτιμηθεί χρησιμοποιώντας τις ανα-συχνότητα εκτιμήσεις κατευθύνσεων άφιξης.

Στη συνέχεια μελετάμε το πώς μπορούμε να βελτιώσουμε την ακρίβεια στην εκτίμηση των κατευθύνσεων άφιξής, αφού αυτή αποτελεί έναν σημαντικό παράγοντα που επηρεάζει την ακρίβεια της εκτίμησης θέσης. Προτείνουμε μια τεχνική που μπορεί να συνδυαστεί με οποιαδήποτε μέθοδο εκτίμησης κατευθύνσεων άφιξης για πιο ακριβείς και αξιόπιστες εκτιμήσεις. Τέλος, ερευνούμε και περιγράφουμε δύο παραδείγματα για την δυνητική χρήση της πληροφορίας σχετικά με τη θέση των ηχητικών πηγών σε διάφορες εφαρμογές επεξεργασίας ήχου. Το πρώτο παράδειγμα αφορά τη χρήση της πληροφορίας της θέσης για την παραγωγή ήχου με χωρική πληροφορία και το δεύτερο αφορά τη σχεδίαση ενός σχηματιστή λοβού (βεαμφορμερ) που χρησιμοποιεί την πληροφορία των θέσης για την ενίσχυση του σήματος μιας ηχητικής πηγής. Αρχικά αποτελέσματα στις δύο αυτές εφαρμογές δείχνουν ότι μέθοδοι που βασίζονται στη θέση των ηχητικών πηγών μπορούν δυνητικά να χρησιμοποιηθούν σε εφαρμογές διαχωρισμού πηγών και βελτίωσης της ποιότητας των λαμβανόμενων ηχητικών σημάτων.

Λέξεις κλειδιά: εύρεση θέσης, εκτιμήσεις κατεύθυνσης άφιξης, δίκτυα ακουστικών αισθητήρων, πληροφορία θέσης, συστοιχίες μικροφώνων

Επόπτης: Αθανάσιος Μουχτάρης
Αναπληρωτής Καθηγητής
Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Contents

List of Figures

List of Tables

Acronyms

DOA Direction of Arrival

DUET Degenerate Unmixing and Estimation Technique

CDF Cumulative Distribution Function

CFRC Coarse-To-Fine Region Contraction

CRLB Cramer-Rao Lower Bound

FIM Fisher Information Matrix

GB Grid-based

GCC-PHAT Generalized Cross-Correlation with PHAse Transform

ICA Independent Component Analysis

IP Intersection Point

ISRG Image Source Relative Gain

LLS Linear Least Squares

MASS Minimum Angular Source Separation

ML Maximum Likelihood

MUSIC Multiple Signal Classification

MVDR Minimum Variance Distortionless Response

NLS Non-linear Least Squares

OGCF Oriented Global Coherence Field

P-NLS Position Non-linear Least Squares

RMSE Root Mean Square Error

RSS Received Signal Strength

SCA Sparse Component Analysis

SIR Signal-to-interferer ratio

SNR Signal-to-Noise ratio

SOF Spatial Observability Function

SRC Stochastic Region Contraction

SRP Steered Response Power

TF Time-Frequency

TDOA Time Difference of Arrival

WASNs Wireless Acoustic Sensor Networks

WDO W-disjoint orthogonality

Chapter 1
Introduction

1.1 General Objective

Microphone arrays have become popular due to their ability to perform operations, such as direction of arrival (DOA) estimation and beamforming [5]. By exploiting the spatial diversity of microphones, microphone arrays are able to estimate the direction of arrival of incoming sound, enhance the sound coming from a specific direction, and/or cancel interfering sound sources from other directions. Microphone arrays are already used in numerous applications, such as distant sound acquisition, spatial audio, camera steering in videoconferencing, speech enhancement, and robust speech recognition. Despite their superior performance over traditional single-microphone systems, microphone arrays still have limitations, as they are capable of sampling the sound field only locally. A reduced performance is experienced when, for example, the sound source is located at a large distance from the array, as the recorded signals will have a low Signal-to-Noise ratio (SNR) to perform any operation.

In the past few years, the signal processing community has witnessed a new paradigm in the way that acoustic signals are acquired: traditional microphone arrays are being replaced by Wireless Acoustic Sensor Networks (WASNs), where a number of acoustic sensors—which can be microphones or microphone arrays—are distributed over an area. These sensors utilize wireless links to exchange information and also feature limited processing capabilities in order to perform signal processing tasks. WASNs have emerged from the need to provide better spatial coverage. Having multiple acoustic sensors distributed in an environment (which can be either indoor or outdoor) increases the probability of finding a microphone that is close to a source, thus capturing the sources' signals with higher signal-to-noise ratio. Moreover, as the nodes are connected over wireless links, the microphones can be placed at locations where it is difficult to place wired sensors. WASNs have attracted a lot of interest because they can be used in a variety of applications, such as in hearing aids, ambient intelligence, hands-free telephony and acoustic monitoring [6]. In this new paradigm of WASNs, a crucial task that has received significant research interest is that of estimating the locations in space of the multiple simultaneously active acoustic sources. Inference of location information is important for many applications, such as wildlife monitoring [7–9] and speech enhancement for robust signal acquisition [10].

In the general framework, the network consists of nodes that are comprised of a single or multiple microphones that passively monitor the acoustic environment. Note that, this problem of *passive localization* is different to that of *active localization* met in wireless sensor networks where the user device tries to estimate its own position based on information emitted by the sensors, such as Received Signal Strength (RSS), radio-frequency signals, etc. [11, 12]. In passive localization of acoustic sources, the sensor network must estimate the location of a source by using measurements (i.e., acoustic signals) gathered by the nodes. In contrast to active localization where each individual device estimates its own position, in passive localization the nodes are responsible for the localization task and thus the location estimation method must also be able to cope with multiple sources. Usually, a central node, known as the fusion center, collects all the measurements and performs the localization task. Moreover, as the number of active sound sources is generally unknown, the localization problem is tightly connected to that of source counting, i.e., estimating the number of sources that are active in the environment.

1.2 Motivation and Vision

The first challenge in sound source localization consists of developing methods that can accurately estimate the number of active sources and their locations in noisy and reverberant conditions. However, the real-life application of such methods in practical WASNs is another major challenge, especially because the particular nature of the sensor network poses several limitations that must be taken into account when designing the localization method—and any signal processing method in general—for WASNs. These limitations include:

- The **processing power** and the **battery life** of the nodes: A sensor network usually consists of low-cost, battery-powered nodes with reduced processing capabilities. Thus, the processing requirements at the nodes cannot be high or battery consuming.

- **Bandwidth limitations:** As the nodes are wirelessly connected, the amount of information that needs to be transmitted by a node must be attained at low levels. Consider for example that when the sensors are microphone arrays, the transmission of the multiple raw signals to a central node requires significant bandwidth that is usually not available in sensor networks, especially when the number of sensors is increased.

- **Synchronization issues:** Each sensor has its own clock for sampling the signals. As the clocks among sensors are not synchronized, the signal processing method must be able to tolerate unsynchronized input or provide some means to perform synchronization between the sensors.

- **Real-time requirements:** The majority of applications require the signal processing task to be performed in real-time. This highlights the need for computationally efficient methods.

Motivated by the aforementioned constraints posed by the sensor network itself, our vision is to develop localization methods that not only achieve high accuracy, but also attain low communication bandwidth, tolerate unsynchronized input, and are computationally efficient to facilitate their application in real-life WASNs.

1.3 The Approach

Throughout the literature various localization methods have been proposed that rely on various types of information transmitted by the sensors, however few of them consider the aforementioned practical challenges. In our approach, we utilize direction of arrival (DOA) information that describes the directions of the sound sources with respect to a sensor. To perform the DOA estimation task, each sensor must consist of multiple microphones, i.e., each sensor must be equipped with a microphone array. DOA-based approaches are attractive for localization because they have many advantages related to the practical challenges met in WASNs. First, they can attain very low bandwidth needs as only the DOA estimates need to be transmitted. Moreover, they do not require the sensors to be perfectly synchronized thus alleviating a very important constraint met in WASNs. Finally, the variety of broadband DOA estimation methods available in the literature [13–18] makes it easy to obtain such estimates both for the single and multiple sources case.

We first consider the single source case and design a computationally efficient non-linear least squares estimator that can accurately estimate the location of a source from its DOA estimates using an iterative grid-based procedure. We then proceed to the multiple sources case where a core problem for DOA-based approaches is that the central node that receives the multiple DOA estimates from the sensors cannot know to which source each DOA belongs. This is known as the *data-association problem*. We identify, define and propose solutions to this problem. One such solution derives from the extension of our method from the single source to the multiple sources case, where all possible DOA combinations from the sensors are examined in order to select the ones that correspond to the locations of the sources. This is done by utilizing heuristics that rely on the estimated locations and the corresponding DOA combinations.

To improve the performance of the localization method we then further examine the data-association problem and propose a solution that utilizes additional information from the sensors, apart from the DOA estimates. Each DOA from each sensor is associated with a feature that can be thought of as a “fingerprint” of the corresponding source. The association of DOAs from the sensors to the sources is then carried out by comparing the corresponding fingerprints and grouping them according to their similarity. In this way, the data-asssociation problem is addressed prior to the localization procedure and the multiple source localization problem decomposes into multiple single source localization problems that can be efficiently solved using any single-source localization method available in the literature. Keeping in mind the practical challenges that have to be considered in real-life WASNs, we propose a computationally efficient greedy algorithm to perform the association and we incorporate our method with a scheme to reduce the transmission requirements of the additional information that is necessary.

Taking also into account the problem of source counting, we then propose a DOA-based approach to jointly estimate the number of sources and their locations. Our approach is based on statistical clustering of narrowband per-frequency location estimates which are inferred using narrowband per-frequency DOA estimates from the sensors. The narrowband location estimates are expected to form clusters around the true sources' locations thus the source counting and location estimation problem reduces to the problem of estimating the number of clusters and the corresponding cluster centroids.

Of course, a very significant aspect that affects the localization performance of DOA-based methods is the accuracy in which the DOA estimates are inferred. Since the accuracy of the DOA estimation constitutes such an important factor, we also investigate how we can improve the performance of the DOA estimation. To do so, we propose a methodology that can be applied to any narrowband DOA estimation method in order to infer more accurate and reliable DOA estimates. Our approach is based on statistical modelling of the microphone array signals in the frequency domain with the complex Watson distribution and then applying the DOA estimation method to the distribution's mode vector instead of the raw microphone array signals. The choice of the distribution is motivated by directional statistics where it is used to model uncertainties in directions of complex unit-norm vectors.

The use of spatial features, typically consisting of direction of arrival estimates and estimates of how diffuse or directional the soundscape is, have been extensively used in numerous applications involving speech enhancement, source separation, spatial audio, and robust speech recognition. In the final part of this dissertation, we investigate the potential use of location information to applications involving speech enhancement and separation. We present preliminary results on the design of methods that leverage location information in order to separate the multiple active sound sources in a cocktail party situation. In our first example, we demonstrate how multiple microphone arrays can collaborate during beamforming and post-filter design in order to separate the multiple sources' signals for spatial audio acquisition. We show that when multiple arrays collaborate during the beamforming and post-filter design we can achieve better separation and more efficient capturing of the acoustic environment compared to using the closest array to the each source. In our second example, we consider the problem of beamforming using time-frequency masks for steering vector estimation. We show that these masks can be constructed using location information and we validate that our location-based time-frequency mask estimation can achieve satisfactory performance, revealing the potential of using location information as a new direction to speech enhancement in WASNs.

1.4 Contributions of this Dissertation

The key contributions of this thesis are the following:

- The development of localization methods for multiple simultaneously active sound sources that facilitate potential application in real-life wireless acoustic sensor networks. The main focus of this thesis was to develop localization approaches that are not only accurate in terms of localization performance, but also take into account the practical challenges that occur in real-life deployments, such as bandwidth limitations, synchronization issues, and computational efficiency.

- The investigation of the data-association problem that occurs when localizing multiple sources from DOA estimates and has severe implications in the performance of any DOA-based localization method. To the best of our knowledge, this problem has not been widely examined, especially under realistic conditions that include missed detections occurring at the sensors (i.e., some sensors cannot detect the DOAs of some sources), reverberation, noise, and moving sources. Our proposed approach to this problem is robust to noise, reverberation and missed detections and can be combined with any DOA-based localization method that exists in the literature.

- The development of online per-frame methods for the joint problem of source counting and location estimation in WASNs and the investigation of how statistical clustering and model selection can be combined with DOA-based localization in order to infer the number of active sources and their locations.

- A novel methodology based on statistical modelling of the microphone array signals that can be used to improve the accuracy of narrowband instantaneous DOA estimation. Our methodology can be used with any narrowband DOA estimation method and any microphone array geometry.

- A new perspective on the design of spatial filters for speech enhancement and separation that are based on location information. We believe that our preliminary results on this direction reveal that location-based approaches have the potential to result in improved speech enhancement and separation.

1.5 List of publications

The work presented in this thesis has resulted in 3 journal publications and 8 conference papers:

- A. Alexandridis, A. Mouchtaris, “Multiple sound source location estimation in wireless acoustic sensor networks: the data-association problem,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

- M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, B. Lee, “A Survey of sound source localization methods in wireless acoustic sensor networks”, *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 3956282, 24 pages, 2017.

- A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, A. Mouchtaris, “Localizing multiple audio sources in a wireless acoustic sensor network”, *Signal Processing, Special issue on wireless acoustic sensor networks and ad-hoc microphone arrays*, vol. 107, February 2015, pp. 54-67, ISSN 0165-1684, http://dx.doi.org/10.1016/j.sigpro.2014.08.013.

- A. Alexandridis, N. Stefanakis, A. Mouchtaris, “Towards wireless acoustic sensor networks for location estimation and counting of multiple speakers in real-life conditions”, *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2017.

- T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schluter, H. Ney, H. Reinhold, A. Mouchtaris, “The RWTH/UPB/ FORTH System Combination for the 4th CHiME Challenge Evaluation”, *The 4th International Workshop on Speech Processing in Everyday Environments*, 2016. **Scored 2nd in all tracks of the CHiME speech recognition challenge.**

- A. Alexandridis, A. Mouchtaris, “Improving narrowband DOA estimation of sound sources using the complex Watson distribution”, *European Signal Processing Conference (EUSIPCO)*, 2016.

- A. Alexandridis, S. Papadakis, D. Pavlidi, A. Mouchtaris, “Development and evaluation of a digital MEMS microphone array for spatial audio”, *European Signal Processing Conference (EUSIPCO)*, 2016.

- A. Alexandridis, A. Mouchtaris, “Multiple sound source location estimation and counting in a wireless acoustic sensor network”, *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* , 2015.

- A. Alexandridis, G. Borboudakis, A. Mouchtaris, “Addressing the data-association problem for multiple sound source localization using DOA estimates”, *European Signal Processing Conference (EUSIPCO)*, 2015 **Best Student Paper Award**

- A. Alexandridis, A. Griffin, A. Mouchtaris, “Breaking down the cocktail party: Capturing and isolating sources in a soundscape”, *European Signal Processing Conference (EUSIPCO)*, September 2014

- A. Griffin, A. Alexandridis, D. Pavlidi, A. Mouchtaris, “Real-time localization of multiple audio sources in a wireless acoustic sensor network”, *European Signal Processing Conference (EUSIPCO)*, September 2014

1.6 Outline of Dissertation

The rest of this dissertation is organized in the following way: Chapter 2 presents the taxonomy of sound source localization methods in wireless acoustic sensor networks. The taxonomy is built upon the nature of information that is transmitted by the sensors. Such information may include (i) Time Difference of Arrival (TDOA) measurements, (ii) Direction of Arrival (DOA) measurements, (iii) energy readings, or (iv) the use of the Steered Response Power (SRP) function. Since the focus of this dissertation is on DOA-based localization approaches, Chapter 3 reviews in more detail the localization approaches that are based on direction of arrival estimates both for the single and the multiple sources case. Chapter 4 presents our proposed Grid-based method for localizing a single and multiple sources. For the single-source case, our method is a computationally efficient non-linear least squares estimator. For multiple sources, heuristic measures are employed to surpass the data-association problem. Then, Chapter 5 presents our proposed approach to the data-association problem by utilizing additional information transmitted by the sensors, apart from the DOA estimates. The data-association problem is addressed prior to the localization procedure and the multiple source localization problem decomposes into multiple single source localization problems that can be efficiently solved using any single-source location estimator available in the literature. The joint problem of source counting and localization is considered in Chapter 6, which presents our proposed approach to estimating both the number of active sources and their corresponding locations. Chapter 7 then presents our proposed methodology to enhancing the narrowband per-frequency DOA estimation procedure, which can have beneficial effects on the final localization accuracy. Chapter 8 presents some examples of how location information can be used for speech enhancement and separation. The first example presents a method where multiple microphone arrays use location information and collaborate during beamforming and post-filter design in order to separate the multiple sources’ signals for spatial audio acquisition. The second example shows how location information can be used for time-frequency mask estimation in order to estimate the steering vector which is necessary when estimating the beamformer filter coefficients. Finally, Chapter 9 concludes the dissertation.

Chapter 2
A taxonomy of sound source localization methods

2.1 Introduction

A taxonomy of sound source localization methods for wireless acoustic sensor networks can be built upon the nature of information from the sensors that is utilized in order to infer the location(s) of the source(s). Such information includes (i) Time Difference of Arrival (TDOA) measurements, (ii) Direction of Arrival (DOA) measurements, (iii) energy reading, or (iv) the use of the Steered Response Power (SRP) function. The interested reader is referred to [1] for a complete literature review on the localization methods. In the following, Sections 2.2–2.5 briefly review the localization methods of each category highlighting their properties, advantages, and disadvantages. Then, Section 2.6 provides a literature review on source counting and Section 2.7 discusses works on real deployments of WASNs for localization.

2.2 TDOA-based localization

When each sensor consists of multiple microphones, it can estimate and transmit the time differences of arrival (TDOAs) between different microphone pairs. This strategy is a very good candidate to localization when no synchronization between the sensors is guaranteed and when low bitrate requirements have to be met. However, it results in increased computational complexity in the sensors in order to estimate the TDOAs.

2.2.1 TDOA estimation at the sensors

The TDOAs are usually estimated by the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) [19]. Assuming that the microphone locations are known, the GCC-PHAT function between two microphone signals x_i and x_j is given by:

$$R_{\mathrm{GCC}_{x_i x_j}}(\tau) = \int_{-\infty}^{\infty} \Phi(\omega) S_{x_i x_j}(\omega) e^{j\omega\tau} \, \mathrm{d}\omega \tag{2.1}$$

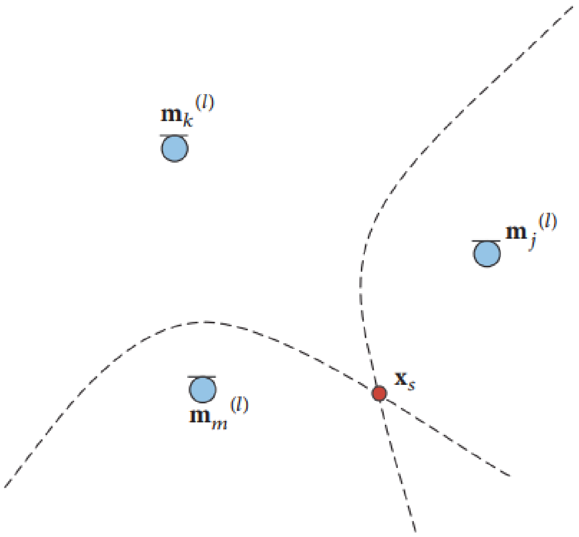


Figure 2.1: Figure taken from [1]. Localization using time-differences of arrival (TDOAs). The source location is found as the intersection of hyperbolas defined by the corresponding TDOAs.

where $S_{x_ix_j}(\omega)$ is the cross-power spectrum between the microphone signals and $\Phi(\omega)$ is a prefiltering function known as the Phase Transform and is given by:

$$\Phi(\omega) = \frac{1}{|S_{x_ix_j}(\omega)|} \tag{2.2}$$

The TDOA is then estimated from the peak of the GCC-PHAT function as:

$$\hat{\tau}_{ij} = \arg\max_{\tau} R_{\mathrm{GCC}_{x_ix_j}}(\tau) \tag{2.3}$$

However, TDOAs are very sensitive to noise and reverberation. As a result, in adverse environments, some TDOAs may not be reliable and may have to be discarded and not be taken into account by the location estimation process. In order to identify such outliers, several works [20–22] analyze the shape of the GCC function and propose reliability measures related to the maximum and second maximum peak location of the GCC. Other works use the observation that the TDOAs along closed paths of microphones sum to zero (e.g., $\tau_{ij} + \tau_{jk} + \tau_{ki} = 0$) [23] or combine all aforementioned approaches [24].

2.2.2 Source localization using TDOAs

For a given microphone pair, the estimated TDOA $\hat{\tau}_{ij}$ constraints the source to lie on the hyperbola whose foci are the microphone locations and whose vertices are $c\hat{\tau}_{ij}$ apart from each other with c being the speed of sound. When the source and the microphones are coplanar the source location can be obtained by the intersection of two or more hyperbolas, as shown in Fig. 2.1. However, due to noise in the TDOA measurements, usually multiple constraints from multiple microphone pairs are combined and the source location can be estimated by the intersection of the hyperbolas in the two dimensions or hyperboloids in the three dimensions.

However, intersecting hyperbolas comes with increased computational cost as the problem is highly non-linear. Several estimators have been proposed that are based on the minimization of different cost functions involving the source location and vary from closed-form to iterative solutions. Among them, the Spherical Intersection [25], Spherical Interpolation [26], Least Squares with Linear Correction [27], and Squared Range Difference Least Squares [28].

2.3 DOA-based localization

When each node in the network incorporates multiple microphones, the location of an acoustic source can be estimated using direction of arrival (DOA) (also known as *bearing*) measurements. DOA measurements describe the direction from which sound is propagating with respect to a sensor. Similar to TDOA-based localization, such approaches can attain very low transmission requirements—as only the DOA measurements need to be transmitted to the central node—and do not require the sensors to be perfectly synchronized. However, they come at the cost of an increased computational complexity at the sensors as DOA estimation has to be carried out.

Moreover, DOA-based approaches are attractive due to the ease in which DOA estimates can be obtained: a variety of broadband DOA estimation methods for acoustic sources is available in the literature, such as the broadband MUSIC algorithm [13], the ESPRIT algorithm [14], methods based on Independent Component Analysis (ICA) [15] or Sparse Component Analysis (SCA) [16–18]. It is also important to highlight that many of these methods are suitable for the multiple sources case, which makes DOA-based approaches a perfect candidate for the location estimation of multiple sources in an acoustic environment. In the general case, the location of a source can be estimated by the intersection of bearing lines, i.e., lines emanating from the sensors at the direction of the corresponding DOA, a method which is known as *triangulation.* Several estimators have been proposed to solve the triangulation problem which range from linear closed-form solutions to iterative non-linear estimators. DOA-based approaches to both single and multiple source localization are discussed in detail in Chapter 3.

2.4 Energy-based localization

Energy-based approaches for sound source localization build upon the fact that the energy of an acoustic signal attenuates as a function of distance from the source. Based on this property, the location of a sound source can be estimated by comparing energy readings from different microphones. Energy-based approaches are able to operate in sensor networks with distributed microphones, eliminating the need for multiple microphones in each sensor. Moreover, an energy reading is much simpler to obtain than a TDOA or DOA measurement.

For the energy attenuation model, generally the free-space model is assumed where the acoustic energy $s_i(t)$ received at time t from microphone i attenuates at a rate which is inversely proportional to the square of the distance [29]:

$$s_i(t) = g_i \frac{s(t)}{||\boldsymbol{p}_s - \boldsymbol{p}_i||^2} \tag{2.4}$$

where $s(t)$ is the energy emitted by the source, g_i is the microphone gain, and $\boldsymbol{p}_s$, and $\boldsymbol{p}_i$ are the location vectors of the source and the i-th microphone respectively. Based on energy readings between the microphones or energy ratios between pairs of microphones several linear and non-linear estimators have been proposed [30–33]. A recent review of energy-based source localization methods can be found in [34].

However, in order to achieve reasonable performance there are several practical challenges that must be overcome. First of all, the performance of an energy-based localization method depends on the validity of the energy attenuation model. In practise, room reverberation and echoes will severely reduce the localization performance in indoor environments. In outdoor environments the performance will also be degraded as the wind direction and noise affect the sound propagation. Another major constraint is the need for all microphones to be calibrated in order to produce consistent energy readings.

2.5 SRP-based localization

Approaches based on the SRP have gained attention due to their robustness in noisy and reverberant conditions. Similarly to TDOA and DOA based approaches, they require multiple microphones at each sensor. They are based on the construction of power maps that are defined over a set of possible sound source locations and describe the plausibility that a sound source is present at a specific location [35]. The power map (also known as Global Coherence Field) is expected to exhibit a peak at the true source's location.

For the generation of the power map, the use of the SRP-PHAT (Steered Response Power with PHAse Transform) [36] is the most dominant choice in the literature [35, 37–39]. The SRP-PHAT can be computed from the GCC-PHAT [19] between all possible microphone pairs at each sensor. Given L microphone pairs, the SRP-PHAT can be computed as:

$$P_{\mathrm{SRP}}(\boldsymbol{x}) = \sum_{l=1}^{L} \int_{-\infty}^{\infty} \frac{X_{l_1}(\omega) X_{l_2}^{*}(\omega)}{|X_{l_1}(\omega) X_{l_2}^{*}(\omega)|} e^{j\omega(\tau(\boldsymbol{x},l_1)-\tau(\boldsymbol{x},l_2))} \, \mathrm{d}\omega \tag{2.5}$$

where l_1 and l_2 are the indices of the microphones that belong to the l-th pair and $\tau(\boldsymbol{x}, k)$ is the theoretical time of travel from location $\boldsymbol{x}$ to microphone k. However, it is evident that these methods require high transmission bandwidth as the SRP power maps from the individual sensors have to be transmitted to the central node in order to construct the global SRP power map.

The location of the source $\boldsymbol{p}$ is then estimated using a grid-search over the set of all candidate locations $\mathcal{G}$. The estimated location is given by the candidate location where the SRP is maximized:

$$\boldsymbol{p} = \arg\max_{\boldsymbol{x} \in \mathcal{G}} P_{\mathrm{SRP}}(\boldsymbol{x}) \tag{2.6}$$

To further enhance the estimated power maps, Aarabi *et al.* [35] used a weighted addition in order to integrate the individual power maps from each sensor into a global power map. The weights define the Spatial Observability Function (SOF) of each sensor and are assigned according to how well each sensor is able to detect a sound source, depending on the physical characteristics of the environments, e.g., walls that are near a microphone array or obstacles that obscure the array to detect sources at certain locations. Lastly, it is worth mentioning the work of [40, 41] which proposes the Oriented Global Coherence Field (OGCF), a power map to estimate the head orientation along with the speaker location.

However, the SRP function may exhibit many local extrema and thus a complete search over the entire space must be performed to find the global maximum, which is very computationally demanding. Approaches to reduce the computational cost have also been investigated, such as the Stochastic Region Contraction (SRC) [42] and the Coarse-To-Fine Region Contraction (CFRC) [43]. The main idea behind these approaches is to start with a search volume that contains the global maximum and many local extrema and iteratively contract that space until a sufficiently small sub-volume is reached that contains the global maximum. A problem of these methods is that they may discard part of the available information which leads to some performance degradation. Other approaches [44, 45] tried to design modified versions of the SRP function in order to increase robustness and relax the computational complexity.

For multiple sources, a straightforward way is to detect the most dominant peaks of the power map and report those locations as the locations of the sources. However, the power map may exhibit many peaks as the surface is generally not smooth. Moreover, several of these peaks may not correspond to a location of a source [39]. An approach for localizing multiple sources is presented in [39] which attempts to de-emphasize the dominant source from the power map—after it has been detected—in order for the other sources to stand out.

2.6 Source counting

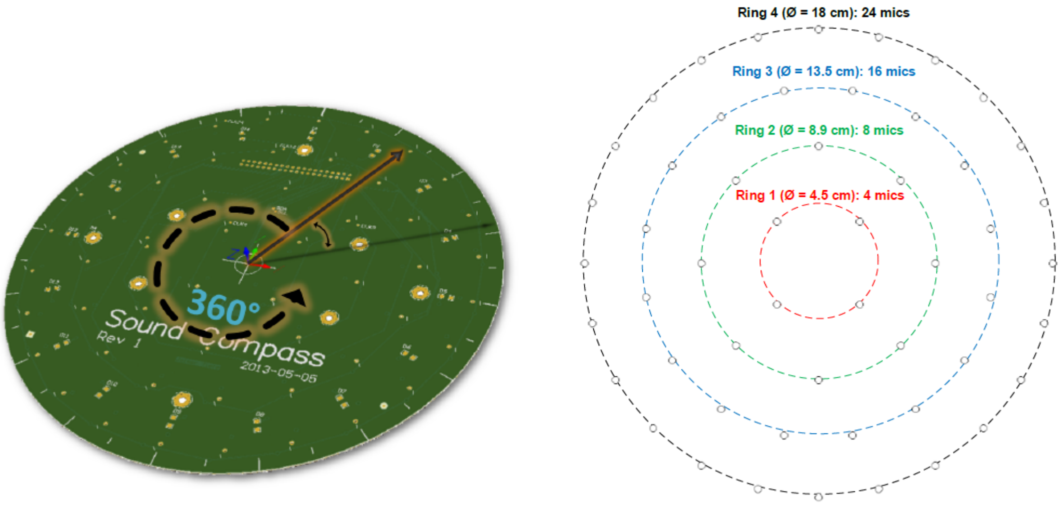
Assuming that the number of sources is also unknown and can vary arbitrarily in time, approaches were developed for WASNs in order to jointly solve the source counting and location estimation problem. In these approaches, the central idea is to utilize narrowband DOA estimates—for each time-frequency bin—from the sensors in order to estimate narrowband location estimates. Appropriate processing of the narrowband location estimates can infer the number and locations of the sound sources. A geometry-based approach that estimates the number of sources is presented in [46]. Another approach that processes the narrowband location estimates with statistical modeling methods is presented in [10]. In this work, the narrowband location estimates are modelled by a Gaussian Mixture Model (GMM), where the number of Gaussian components corresponds to the number of sources, while the means of the Gaussians determine the sources' locations. A variant of the Expectation-Maximization (EM) algorithm is proposed that incorporates empirical criteria for removing and merging Gaussian components. Another variant of the EM algorithm is discussed in [47], while distributed approaches are presented in [48, 49].

Other approaches to source counting have also been proposed, although they do not consider WASNs but rather a single microphone array. One class of such approaches perform source counting by modelling the microphone array signals in the Short Time Fourier Transform (STFT) domain [50–52]. In [50] a variational EM algorithm using complex Gaussian Mixture Models is used to model the complex STFT coefficients, while the authors in [51, 52] employ mixtures of complex Watson distributions. Remaining in the case of a single microphone array, other approaches rather than modelling the raw microphone array signals, perform modelling on the narrowband DOA estimates obtained in each frequency bin [53, 54]. Note that, although the same concepts can be extended from a single microphone array to a WASN, such extension is not straightforward. It would require investigation of whether such models can be applied to narrowband location estimates, as the distribution of such estimates can be very different to that of DOA estimates.

2.7 Deployment studies of real WASNs

Researchers have also considered the real deployment of acoustic sensor networks [2, 55–57], focusing on issues, such as the the hardware and software design of the sensors, the communication and synchronization protocols, and the accuracy of the localization task on real field measurements.

Recently, the SoundCompass [2] has been developed as a low-cost sensor capable of measuring the sound field directionality. The SoundCompass consists of 52 MEMS microphones


360°
Sound Compass
Rev 1
2013-05-05
Ring 4 (Ø = 18 cm): 24 mics
Ring 3 (Ø = 13.5 cm): 16 mics
Ring 2 (Ø = 8.9 cm): 8 mics
Ring 1 (Ø = 4.5 cm): 4 mics

(a) The SoundCompass sensor

(b) The SoundCompass microphone array geometry

Figure 2.2: Figure taken from [2]. The SoundCompass sensor and microphone array geometry

arranged in four concentric rings with the diameter of the circular circuit board being 20 cm (Fig. 2.2).

The sensor employs Delay-and-Sum beamformers [58] at fixed directions—with a certain angle resolution—along the 360^o circle to construct the Steered Response Power (SRP) which describes the signal energy as a function of direction. The SRP from different SoundCompass sensors can be transmitted to a central node and fused together to construct a probability map that exhibits local maxima at the locations of the sources. Their work has focused mostly on the development of the sensor and therefore communication aspects such as routing, bandwidth needs, and time synchronization have not been implemented yet.

A series of field studies for acoustic monitoring with an acoustic sensor network is presented in [57]. The study aims at locating marmot calls and woodpecker vocalizations in their natural environment. For the sensors, the ENSBox platform [59] was used where each node is a tetrahedral microphone array. The ENSBox system features self-configuration software and thus the sensors are able to estimate their locations to within a few centimetres and their orientations to within a few degrees. After deployment, the detection software in each node performs analysis of the audio stream. When a node detects an animal event it notifies the other nodes about the beginning and ending of the animal call signal, and all the nodes—using the integrated synchronization API of the ENSBox—synchronously record and store the specific range of the signal for offline processing. The localization is performed offline based on the recorded signals, thus the requirement for each node to transmit its likelihood function is not discussed. However, this can raise bandwidth concerns in a sensor network that performs the localization in real-time.

2.8 Conclusions

This chapter presented a taxonomy of localization methods in wireless acoustic sensor networks. Different classes of localization methods were discussed that utilize different types of information from the sensors, namely (i) Time Difference of Arrival (TDOA) measurements, (ii) Direction of Arrival (DOA) measurements, (iii) energy readings, and (iv) the Steered Response Power (SRP) function. Moreover, this chapter presented a literature review on source counting and on real deployments of WASNs for localization.

The next chapter will focus on the literature review of the DOA-based class of localization methods, which is also the main focus of this thesis. We will discuss the fundamental estimators for single source localization using DOA measurements and their extensions to the multiple sources case.

Chapter 3
Source Localization using Bearing-Only Measurements

3.1 Introduction

The use of DOA measurements—also known as *bearing* measurements—for sound source localization is an attractive approach as it can attain very low transmission bandwidth and it does not require the sensors to be perfectly synchronized. In this chapter, we review the background and state-of-the-art methods for the localization of a single and multiple sound sources using DOA estimates.

3.2 General problem formulation

Our framework is a wireless acoustic sensor network whose M nodes are each equipped with a microphone array—which we will also refer to as a sensor. This enables each node to generate a DOA estimate for any source that it can detect. The locations of the nodes are known. In the following the terms sensor, node, and microphone array will be used interchangeably. It is important to note that each node estimates consist of azimuth directions only, thus we consider location estimation in the two-dimensional space.

Let the x- and y-coordinates of the location of the m-th node be given by

$$\boldsymbol{q}_m = \begin{bmatrix} q_{\mathrm{x},m} & q_{\mathrm{y},m} \end{bmatrix}^T, \tag{3.1}$$

and, similarly, let the x- and y-coordinates of the location of the k-th source be given by

$$\boldsymbol{p}_k = \begin{bmatrix} p_{\mathrm{x},k} & p_{\mathrm{y},k} \end{bmatrix}^T. \tag{3.2}$$

Let K denote the number of active sound sources that are simultaneously present in the acoustic environment monitored by the WASN. The number of sources is assumed to be known. Then the $2K \times 1$ position vector of all the sources can be written as

$$\boldsymbol{p} = \begin{bmatrix} \boldsymbol{p}_1^T & \boldsymbol{p}_2^T & \dots & \boldsymbol{p}_k^T & \dots & \boldsymbol{p}_K^T \end{bmatrix}^T, \tag{3.3}$$

and we can define the DOA vector for sensor m as

$$\boldsymbol{\theta}_m(\boldsymbol{p}) = \begin{bmatrix} h_{m,1} & h_{m,2} & \dots & h_{m,k} & \dots & h_{m,K} \end{bmatrix}^T, \tag{3.4}$$

where

$$h_{m,k} = \arctan\left(\frac{p_{\mathrm{y},k} - q_{\mathrm{y},m}}{p_{\mathrm{x},k} - q_{\mathrm{x},m}}\right) \tag{3.5}$$

with $\arctan(\cdot)$ being the four-quadrant inverse tangent function.

In the ideal scenario where the microphone array at each node is able to detect all sources, the m-th array outputs a $K \times 1$ vector of DOA measurements

$$\hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_m(\boldsymbol{p}) + \boldsymbol{\eta}_m, \tag{3.6}$$

where $\boldsymbol{\eta}_m$ is the noise at the m-th sensor, which models the inaccuracies in DOA estimation and is assumed to be zero-mean Gaussian with covariance matrix $\boldsymbol{\Sigma}_m = \mathrm{diag}(\sigma_{m,1}^2, \sigma_{m,2}^2, \dots, \sigma_{m,K}^2)$.

However, a node may not be able to detect some of the sources. We refer to these situations as *missed detections*. Missed detections can occur for several reasons: due to the challenging setup in terms of reverberation and noise, because some sources may be located close together in terms of their angular separation for a node to discriminate between them, or because they are located far away from a node. As a result, the number of DOAs each node estimates may be less than K and may also vary in time and across the nodes. In general, each node can detect *up to K* sources.

We also assume the presence of a central node (fusion center). The nodes in the network are connected with the fusion center over wireless links. Each node estimates the DOAs of the detected sources in each time instant and transmits these estimates to the fusion center which is responsible for the localization task.

3.3 Single-source localization

In the single-source case, the location can be estimated as the intersection of bearing lines (i.e., lines emanating from the locations of the sensors at the directions of the sensors' estimated DOAs), a method which is known as *triangulation*. An example of triangulation is illustrated in Fig. 3.1. The problem closely relates to that of target motion analysis, where the goal is to estimate the position and velocity of a target from DOA measurements acquired by a single moving or multiple observers. Hence, many of the methods were proposed for the

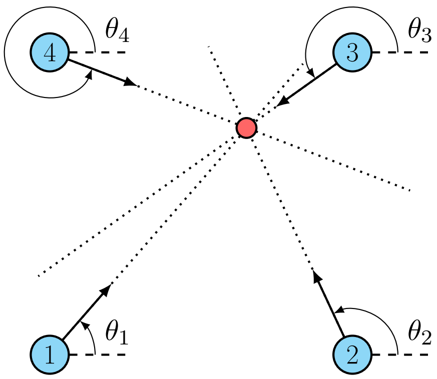


Figure 3.1: Example cell with four sensor nodes (blue circles, numbered 1 to 4), and the DOAs (θ_1–θ_4) to a source (the red circle).

target motion analysis problem, but outlined here in the context of sound source localization in WASNs. Note that as we are considering only one source here, (3.2) and (3.3) reduce to

$$\boldsymbol{p} = \begin{bmatrix} p_\mathrm{x} & p_\mathrm{y} \end{bmatrix}^T . \tag{3.7}$$

and the measured DOA vector $\hat{\boldsymbol{\theta}}_m$ at each sensor reduces to the scalar $\hat{\theta}_m$, which according to the model of (3.6) is now given by:

$$\hat{\theta}_m = \theta_m(\boldsymbol{p}) + \eta_m \tag{3.8}$$

where η_m is the noise at the m-th sensor which for the single-source case is assumed to be univariate zero-mean Gaussian with standard deviation σ_m.

As the DOA estimates will be contaminated by noise, triangulation will not be able to yield a unique solution, craving for the need of statistical estimators to optimally tackle the triangulation problem. This scenario is illustrated in Fig. 3.2. The remainder of this section reviews the fundamental estimators for single source localization.

3.3.1 Traditional Maximum-Likelihood (ML) Estimator

When the DOA estimates are modelled according to (3.8) and the DOA noise is assumed to be Gaussian, the likelihood function of the DOA measurements can be written as:

$$F(\theta_1, \theta_2, \cdots, \theta_M; \boldsymbol{p}) = \prod_{m=1}^{M} \frac{1}{\sigma_m \sqrt{2\pi}} e^{-\frac{1}{2\sigma_m^2}(\hat{\theta}_m - \theta_m(\boldsymbol{p}))^2} . \tag{3.9}$$

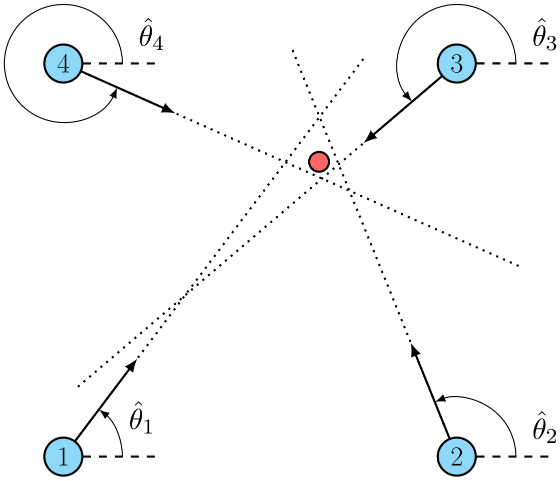


Figure 3.2: Triangulation using DOA estimates contaminated by noise ($\hat{\theta}_1 - \hat{\theta}_4$) in a WASN of 4 nodes (blue circles, numbered 1 to 4) and the estimated location of the sound source (red circle).

The Maximum Likelihood (ML) estimator can be obtained by maximizing the log-likelihood function:

$$\ln F(\theta_1, \theta_2, \cdots, \theta_M; \boldsymbol{p}) = \sum_{m=1}^{M} \left[-\ln(\sigma_m \sqrt{2\pi}) - \frac{1}{2\sigma_m^2} (\hat{\theta}_m - \theta_m(\boldsymbol{p}))^2 \right]. \tag{3.10}$$

The first term in (3.10) can be omitted as it is independent of $\boldsymbol{p}$ and the maximum likelihood estimate of the source location can be written as:

$$\hat{\boldsymbol{p}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{p}} J_{\mathrm{ML}}(\boldsymbol{p}) \tag{3.11}$$

where:

$$J_{\mathrm{ML}}(\boldsymbol{p}) = \sum_{m=1}^{M} \frac{1}{2\sigma_m^2} (\hat{\theta}_m - \theta_m(\boldsymbol{p}))^2. \tag{3.12}$$

The cost function described in (3.12) needs information about the measurement error variance at the sensors. As this information is rarely available in a practical system, (3.12) can be modified to [60]

$$J_{\mathrm{NLS}}(\boldsymbol{p}) = \sum_{m=1}^{M} (\hat{\theta}_m - \theta_m(\boldsymbol{p}))^2 \tag{3.13}$$

The estimator described by (3.13) is the maximum likelihood estimator when the DOA error variance is assumed to be the same at all sensors and is often termed as *Non-linear Least Squares (NLS) estimator*.

The main advantage of these estimators is that they are asymptotically unbiased. On the other hand, due to the non-linear nature of the aforementioned functions, one has to resort to numerical search algorithms for non-linear optimization to find the minimum. As a result, these methods are vulnerable to convergence problems and may diverge under bad initialization, poor geometry between the source and the sensors, and insufficient number of measurements. For the initialization, Linear Least Squares (LLS) estimators described in the next section are used.

To surpass the convergence problems of non-linear least squares estimators, Bishop *et al.* [61] form geometric constraints between the measured data and formulate the problem as a constraint optimization task. Although this method is equivalent to the traditional maximum-likelihood in terms of accuracy, the geometric constraints result in better convergence properties. As a result this method can still work in situations where the traditional maximum-likelihood approach would diverge.

Another variant of the ML estimator is proposed in [62]. The traditional ML estimator minimizes the total bearing error under the assumption of Gaussian noise. The work in [62] proposes an estimator that directly minimizes the mean squared position error. The authors show that this estimator is more accurate and is able to achieve a lower Cramer-Rao Lower Bound (CRLB) than that of the traditional ML approach.

3.3.2 Linear Least Squares (LLS) Estimator

The aforementioned constraints of the maximum likelihood class of location estimators have driven researchers to derive linear closed-form solutions to the localization problem. The pioneering work of Stansfield [63] in 1947 is among the first linear estimators for localizing a single source using DOA measurements. Stansfield developed a weighted linear least squares estimator that can be viewed as a small error approximation of the maximum likelihood function under the assumption that range information r_m is available.

Assuming that the DOA errors are small i.e., $\hat{\theta}_m - \theta_m(\boldsymbol{p}) \approx 0$, the following approximation holds:

$$\hat{\theta}_m - \theta_m(\boldsymbol{p}) \approx \sin(\hat{\theta}_m - \theta_m(\boldsymbol{p})) \tag{3.14}$$

and the cost function of (3.12) can be written as:

$$J_{\mathrm{ST}}(\boldsymbol{p}) = \sum_{m=1}^{M} \frac{1}{2\sigma_m^2} \sin^2(\hat{\theta}_m - \theta_m(\boldsymbol{p})). \tag{3.15}$$

Using the relation

$$\sin(\hat{\theta}_m - \theta_m(\boldsymbol{p})) = \sin\left[\hat{\theta}_m - \arctan\frac{\Delta y_m}{\Delta x_m}\right] = \frac{\Delta x_m \sin\hat{\theta}_m - \Delta y_m \cos\hat{\theta}_m}{r_m} \tag{3.16}$$

where:

$$\Delta x_m = p_{\mathrm{x}} - q_{\mathrm{x},m} \tag{3.17a}$$

$$\Delta y_m = p_{\mathrm{y}} - q_{\mathrm{y},m} \tag{3.17b}$$

$$r_m = \sqrt{\Delta x_m^2 + \Delta y_m^2} \tag{3.17c}$$

we can write (3.15) as:

$$\begin{aligned} J_{\mathrm{ST}}(\boldsymbol{p}) &= \frac{1}{2}\sum_{m=1}^{M}\frac{(\Delta x_m \sin\hat{\theta}_m - \Delta y_m \cos\hat{\theta}_m)^2}{r_m^2\sigma_m^2} \\ &= \frac{1}{2}(\boldsymbol{A}\boldsymbol{p}-\boldsymbol{b})^T\boldsymbol{W}^{-1}(\boldsymbol{A}\boldsymbol{p}-\boldsymbol{b}) \end{aligned} \tag{3.18}$$

where:

$$\boldsymbol{A} = \begin{bmatrix} \sin\hat{\theta}_1 & -\cos\hat{\theta}_1 \\ \vdots & \vdots \\ \sin\hat{\theta}_M & -\cos\hat{\theta}_M \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} q_{\mathrm{x},1}\sin\hat{\theta}_1 - q_{\mathrm{y},1}\cos\hat{\theta}_1 \\ \vdots \\ q_{\mathrm{x,M}}\sin\hat{\theta}_M - q_{\mathrm{y,M}}\cos\hat{\theta}_M \end{bmatrix} \text{ and } \boldsymbol{W} = \begin{bmatrix} r_1^2\sigma_1^2 & & & \mathbf{0} \\ & r_2^2\sigma_2^2 & & \\ & & \ddots & \\ \mathbf{0} & & & r_M^2\sigma_M^2 \end{bmatrix}. \tag{3.19}$$

The cost function $J_{\mathrm{ST}}(\boldsymbol{p})$ is now linear and has a closed-form solution:

$$\hat{\boldsymbol{p}}_{\mathrm{ST}} = (\boldsymbol{A}^T\boldsymbol{W}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}^{-1}\boldsymbol{b}. \tag{3.20}$$

The solution of (3.20) requires knowledge of the range information, i.e., the distances between the source and the sensors $r_m,\ m = 1, \cdots, M$. When this information is not available the weighting matrix $\boldsymbol{W}$ can be replaced by the identity matrix, resulting in the following solution [64]:

$$\hat{\boldsymbol{p}}_{\mathrm{OV}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{3.21}$$

The estimator of (3.21) is referred to as the *Orthogonal Vectors Estimator* or the *Pseudolinear Estimator* and is more preferred in practice as the range information and the error variances are usually unknown.

While simple in their implementation and computationally very efficient, these estimators suffer from increased estimation bias [65, 66]. For this reason, the non-linear and maximum-likelihood estimators are often preferred as they are asymptotically unbiased. A comparison between the Stansfield estimator and the ML estimator [67], reveals that the Stansfield es-

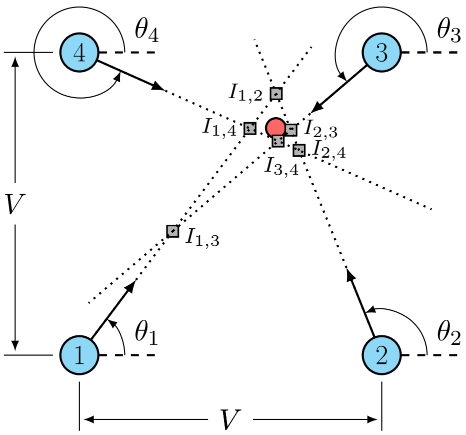


Figure 3.3: Example square cell with four sensor nodes (blue circles, numbered 1 to 4), the DOAs (θ_1–θ_4) to a source (the red circle), and the intersection points (grey squares, labeled $I_{1,2}$–$I_{3,4}$) of DOA vector pairs.

timator provides biased estimates even for a large number of measurements and that the bias does not vanish as the number of measurements increases. To reduce that bias various methods have been proposed based on instrumental variables [66, 68, 69] or total least squares [64, 70].

3.3.3 Intersection point method

Motivated by the need for computational efficiency, the Intersection Point (IP) method [71] is based on finding the location of a source by taking the centroid of the intersections of *pairs* of bearing lines. The centroid is simply the mean of the set of intersection points, and minimizes the sum of squared Euclidean distances between itself and each point in the set. This method can be thought of as sub-optimal version of the LLS method of Section 3.3.2, but we will show later it extends more easily to the multiple source case.

Fig. 3.3 illustrates this method with an example, where the DOA estimates have an error of up to ±5°, and the intersection points are labeled $I_{1,2}$–$I_{3,4}$. The locations of sensors 1 to 4 are: (0, 0), (4, 0), (4, 4), (0, 4), respectively, and the source is at (2.6, 3.0). The estimated location from the centroid of the intersection points is (2.40, 2.77), which is a distance error of 0.43, or 11% of the inter-sensor spacing, V. Further inspection of Fig. 3.3 reveals that the effect of $I_{1,3}$ is significant. By excluding this point from the centroid, the estimated location now becomes (2.64, 2.99) and the error drops to 0.03, or 1% of V.

A question that then naturally arises is: how can we detect and exclude outliers such as $I_{1,3}$? It can be shown that these outliers are caused by bearing lines that are almost parallel. A small change in the slope of either of these lines—due to DOA estimation error—can move their point of intersection significantly. Thus excluding the intersection points of pairs of bearing lines that are almost parallel improves the accuracy of the location estimation.

Before proceeding, let us first define the function $A(X, Y)$, the *minimum angular distance* between X and Y, whose output will be in the range $[0, \pi]$. An elegant—if somewhat inefficient—method to calculate the minimum angular distance between two angles X and Y and return a value in the range $[0, \pi]$ is given by

$$A(X, Y) = 2\sin^{-1}\frac{|\exp(jX) - \exp(jY)|}{2}. \tag{3.22}$$

An equivalent but programmatically more efficient way is to first ensure that X and Y are in the range $[0, 2\pi)$, then by defining

$$A_{X,Y} = (X - Y) \pmod{2\pi} \tag{3.23}$$

$$A_{Y,X} = (Y - X) \pmod{2\pi} \tag{3.24}$$

and then the minimum angular distance is given by

$$A(X, Y) = \min\left(A_{X,Y}, A_{Y,X}\right) \tag{3.25}$$

Now let $\gamma_{\|}$ be a "parallelness" threshold measured in radians, source localization using the intersection point method can then be summarized as:

1. Collect the M DOA estimates.
2. Take each of the pairs of DOA estimates $\theta_i, \theta_j,\ i \neq j$ for sensors i and j and discard it if either of the two conditions are met:

$$A(\theta_i, \theta_j) \ < \ \gamma_{\|}, \tag{3.26}$$

$$A(\theta_i, \theta_j) \ > \ \pi - \gamma_{\|}. \tag{3.27}$$

3. Calculate the points of intersection of the remaining pairs.
4. The estimate of the source location $\hat{\boldsymbol{p}}_{\mathrm{IP}}$ is then given by the centroid of the points of intersection.

3.3.4 Cramér-Rao Lower Bound

This section explores the optimal performance possible for the bearings-only framework we are considering, by presenting a lower error bound for this problem. The Cramér-Rao Lower Bound (CRLB) represents the minimum localization error covariance for any unbiased estimator and is defined as the inverse of the Fisher Information Matrix (FIM) $\boldsymbol{J}(\boldsymbol{p})$ [72]:

$$\mathbb{E}\{(\hat{\boldsymbol{p}} - \boldsymbol{p})(\hat{\boldsymbol{p}} - \boldsymbol{p})^T\} \geq \boldsymbol{J}^{-1}(\boldsymbol{p}) \tag{3.28}$$

where $\hat{\boldsymbol{p}} = [\hat{p}_\mathrm{x} \;\; \hat{p}_\mathrm{y}]^T$ is the estimated source location, and $\mathbb{E}\{\cdot\}$ is the expectation operator. The CRLB for the bearing-only localization of a single source is reported in [73], which derives the following expression for the FIM:

$$\boldsymbol{J}(\boldsymbol{p}) = \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \left[\nabla_{\boldsymbol{p}}\, \theta_m(\boldsymbol{p})\right] \left[\nabla_{\boldsymbol{p}}\, \theta_m(\boldsymbol{p})\right]^T \tag{3.29}$$

with the gradient $\nabla_{\boldsymbol{p}}\, \theta_m(\boldsymbol{p})$ being defined as:

$$\nabla_{\boldsymbol{p}}\, \theta_m(\boldsymbol{p}) = \begin{bmatrix} -\dfrac{p_\mathrm{y} - q_{\mathrm{y},m}}{(p_\mathrm{x} - q_{\mathrm{x},m})^2 + (p_\mathrm{y} - q_{\mathrm{y},m})^2} \\ \\ \dfrac{p_\mathrm{x} - q_{\mathrm{x},m}}{(p_\mathrm{x} - q_{\mathrm{x},m})^2 + (p_\mathrm{y} - q_{\mathrm{y},m})^2} \end{bmatrix} \tag{3.30}$$

3.4 Multiple source localization

The localization of multiple simultaneously active sound sources is not a straightforward extension of the single source case as it poses many challenges. The most fundamental of them is that the correct association of DOAs from the sensors to the sources is unknown, which is known as the *data-association problem.* Also, in realistic scenarios missed detections can occur and thus some sensors may not be able to detect some sources thus underestimating their number.

3.4.1 The data-association problem

When localizing multiple sources, a fundamental problem is that the fusion center receiving the multiple DOA estimates from each sensor (one DOA for each detected source) cannot know to which source each DOA belongs. This is known as the *data-association problem.* The correct association of DOAs from the nodes that correspond to the same source must be found, otherwise location estimation will result in “ghost” sources, i.e., locations not corresponding to real sources.

The data-association problem is illustrated in Fig. 3.4 with an example in a two-node WASN with two active sound sources. The solid lines show the DOAs to the first source and the dashed lines show the DOAs to the second source. Intersecting the DOA lines from the sensors results in 4 possible source locations. When intersecting the DOAs that correspond to the same source, i.e., the two solid lines and the two dashed lines, the correct source locations are estimated (red circles). When the erroneous combination of DOAs is used, the estimation results in “ghost” sources (white circles). When the correct association of DOAs from the sensors to the sources is found, the multiple source localization problem decomposes into multiple single-source localization problems which are straightforward to solve by

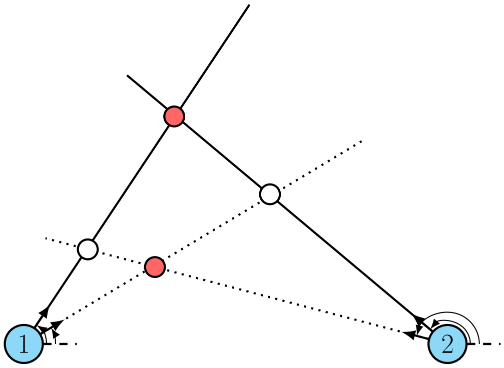


Figure 3.4: Illustration of the data-association problem in a two-node WASN with two active sound sources. The four possible source locations may either be the true sources' locations (red circles) or locations of "ghost" sources (white circles) as the result of using bearing lines that do not correspond to the same source.

applying any single-source location estimator proposed in the literature to the resulted DOA associations.

3.4.2 Missed detections

When multiple sources are active, some sensors may not be able to detect some sources, thus underestimating their number. As a result of such *missed detections*, the number of detected sources—and thus the number of estimated DOAs—can vary across the sensors and through time. This can occur for several reasons:

- Due to the challenging setup in terms of acoustic conditions, such as reverberation and noise.

- Because some sources may be located close together in terms of their angular distance for a sensor to discriminate between them.

- Because some sources may be located far away from a sensor.

Thus any localization algorithm must deal with the ambiguity that each DOA estimate may originate from any source, and that some (or even all) of the sensor nodes may underestimate the number of sources.

3.4.3 Approaches to the multiple source localization and data-association problem

This section reviews approaches to address the data-association problem and perform localization of multiple simultaneously active sound sources. Some of these approaches utilize additional information transmitted by the sensors in order to tackle the data-association problem, while others rely solely on the transmitted DOA estimates.

Position Non-linear Least Squares

An extension of the single-source non-linear least squares estimator for scenarios with no missed detections is discussed in [74]. The method, termed *Position Non-linear Least Squares (P-NLS)*, relies solely on the DOA estimates from the sensors. It enumerates all possible DOA combinations from the sensors and designs a quasi-ML estimator able to estimate the locations of multiple sources. Each combination is an $M \times 1$ vector of DOAs where each sensor contributes with a DOA. In general, if C_s denotes the number of sensors that detected s sources, the number of DOA combinations can be computed by:

$$N_{comb} = \prod_{s=1}^{K} s^{C_s} \tag{3.31}$$

The method incorporates the data-association procedure to the ML cost function, which takes the form:

$$C_{\text{P-NLS}}(\boldsymbol{p}) = \sum_{m=1}^{M} \min_{i} |\hat{\theta}_{m,i} - \theta_m(\boldsymbol{p})|^2 \tag{3.32}$$

where $\hat{\theta}_{m,i}$ is the i-th DOA estimate of sensor m. To minimize (3.32), N_{comb} initial locations are estimated (one for each DOA combination) using a linear least squares estimator, such as the pseudolinear transform. Then, the cost function (3.32) is minimized—using numerical search methods—N_{comb} times, each time using a different initial location estimate.

Each time, for each sensor, the DOA closest to the DOA of the initial location estimate is used to take part in the minimization procedure. In that way, for all initial location estimates, the estimator is expected to converge to a location of a true source.

Intersection point method

The extension of the Intersection Point method (Section 3.3.3) to the multiple source case is relatively straightforward. The method takes advantage of the fact that each DOA estimate from a sensor can only belong to one source. By dividing the possible locations for sources into the K^{C_K} unique combinations of DOA estimates, up to K^{C_K} regions can be obtained, (the word "up to" means that some of these regions may be null, depending on the orientation of the DOA estimates). By counting the number of intersection points in each region, and choosing the one that contains the most intersection points, we obtain the one that is most likely to contain one of the sources. Once we have chosen a region—and thus one of the combinations of DOA estimates—we then choose the next most likely, and so on, until we are left with only one remaining possible combination of DOA estimates pointing to the final source. The intersection point method to localize K sources can be more formally stated as:

1. Find the intersection points of all of the pairs of DOA lines, removing any pair whose lines are too parallel, as in step 2 of the single-source algorithm of Section 3.3.3.
2. Determine the set of sensors that detected K sources X_{K} and then number of sensors that detected K sources C_{K}, set the counter k to zero.
3. Find the $C_{\mathrm{K}}(K-1)$ means of the adjacent pairs of DOAs from the sensors in X_{K}.
4. The vectors of these circular means form $C_{\mathrm{K}}K$ half-planes, find the regions defined by all the intersection of all the possible combinations of pairs of half-planes from different sensors. There will be $K^{C_{\mathrm{K}}}$ of them.
5. Find the region with the most intersection points. If there is a tie, choose the region whose intersection points have the minimum variance. The location of the k-th source is given by the centroid of the intersection points in this region. Increment k.
6. If $k < K$, remove all regions that are not distinct from the already chosen region(s) and go to the previous step.

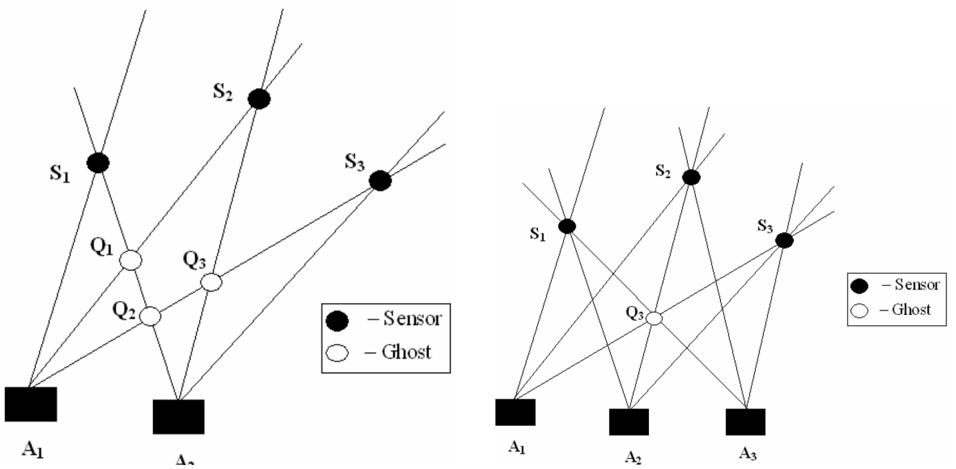
Note that we have described this algorithm conceptually, but it can be implemented very efficiently by using line tests—testing whether a point is above, below, or on a line—and binary masks.

View of the data-association as an assignment problem

The work of [75] views the data association problem as an assignment problem where the goal is to associate a list of measurements with a target in the presence of clutter and false-alarms. The problem is formulated as a statistical estimation problem, involving the maximization of the ratio of the likelihood that the measurements come from the same target to the likelihood that the measurements are false-alarms. However, the proposed solution is *NP-hard* when 3 or more sensors are used ($M \geq 3$). Sub-optimal solutions [76, 77] tried to solve the problem in pseudo-polynomial time.

A ghost elimination approach

An approach to solve the data association problem with successive elimination of ghost sources in time is presented in [3, 78]. This approach assumes a mobile node that moves along a known trajectory, taking n consecutive DOA measurements at different node locations. It is based on the principle that in noiseless environments, the bearing lines from the n measurements will create n-fold intersections only at the locations of the true sources.



(a) True and ghost sources found from $n = 2$ measurements.

(b) True and ghost sources found from $n = 3$ measurements.

Figure 3.5: Figure taken from [3] illustrating the concept of the ghost elimination approach.

This principle is illustrated in Fig. 3.5. The mobile sensor takes $n = 2$ measurements at two distinct locations (A_1 and A_2 in 3.5(a)) and estimates the intersections of the bearing lines emanating from A_1 and A_2. Except from the three source locations S_1, S_2, and S_3, three ghost sources also appear at locations Q_1, Q_2, and Q_3.

At the next time instant, the mobile sensor moves at location A_3 where it takes the third measurement. Based on $n = 3$ measurements now, the locations are estimated by the 3-fold intersections of bearing lines from the three sensor locations, resulting in the elimination of some ghost sources as they represent intersections of two but not three bearing lines. The procedure continues until the number of intersections equals the number of true sources which is assumed to be known. Simulation results show that the method is computationally efficient as generally a small number of measurements is required to eliminate all ghost sources. The major drawback of the method is that it assumes idealized conditions and thus it fails on practical settings where noise is introduced in the measurements.

Statistical clustering of the intersections of bearing lines

The principles of clustering have also been applied as a solution to the data association problem [79]. The main concept behind this approach is that the intersections of bearing lines that correspond to the same source will be close to each other, while intersections from bearing lines that correspond to ghost sources will be randomly distributed in space.

Based on this observation, the approach of [79] recursively incorporates the DOA measurements of one sensor at a time in order to associate each bearing line with one source.

The algorithm starts by finding the intersections of bearing lines from two arbitrary sensors. Then a third sensor is considered and the new intersections of bearing lines from the previous sensors and the new one are found. Each bearing line is then associated with a target based on the fact that the intersections that correspond to the true sources are expected to be close to the intersections found in the previous step. In this way, a bearing line can be associated with one target based on the total distance of the intersections of that bearing line with the intersections of the previous step. The procedure continues until all sensors are used.

Blind signal separation

Allowing the transmission of additional information—apart from the DOA estimates—can generally lead to more efficient solutions at the expense of increased bandwidth utilization in the sensor network. The method of [80, 81] addresses the data-association problem prior to localization. Once the association of DOAs to the sources is found, the multiple source localization problem decomposes into multiple single-source localization problems which can be efficiently solved using any single-source location estimator. The method is based on blind signal separation to address the data-association problem. It associates each detected DOA with a binary mask in the frequency domain that can be used to separate the corresponding source signal.

The binary masks for each source and the corresponding DOA estimates are estimated using the Degenerate Unmixing and Estimation Technique (DUET) [82]. The algorithm relies on the W-disjoint orthogonality (WDO) [83] assumption, which for any two sources can be stated as:

$$S_1(\tau,\ell)S_2(\tau,\ell) = 0 \tag{3.33}$$

where $S_i(\tau,\ell)$ $i = 1, 2$, is the windowed Fourier transform of the two source signals at frequency bin ℓ and time frame τ. This means that there is no energy overlap between the two signals (i.e., only one source is dominant in each time-frequency bin), an assumption which has been shown to be valid especially for the case of speech signals [84].

Given K active sources and two microphone arrays, the work in [80, 81] attempts to solve the data association problem by finding the binary masks between the two arrays that correlate the most. This can be achieved by transmitting the DOAs and the binary masks $\Omega_{k,m}$ $k = 1, \cdots, K,$ $m = 1, 2$ to the central processing node where a matrix $\boldsymbol{A}$ is formed as follows:

$$A = \begin{bmatrix} \alpha_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \alpha_{N,N} \end{bmatrix} \tag{3.34}$$

where

$$\alpha_{i,j} = ||\Omega_{i,1} \odot \Omega_{j,2}|| \tag{3.35}$$

where $\odot$ denotes elementwise multiplication.

To estimate a probable DOA pair that corresponds to the same source, the largest elements from $\boldsymbol{A}$ can be successively picked. When an element is selected, its corresponding row and column are removed from the next iteration in order to ensure that each DOA from each sensor will be used only once. However, the method assumes scenarios with no missed detections, while the association algorithm is designed for the limiting case of two nodes.

3.4.4 Cramér-Rao Lower Bound

This section extends the Cramér-Rao Lower Bound (CRLB) of Section 3.3.4 to the multiple source case. We consider an ideal system, in that we assume that each sensor is able to detect every source and that the association of the DOAs to the sources is known.

The multiple-source version of the likelihood function for the Gaussian error model of (3.6) can be written as,

$$\begin{aligned} F(\boldsymbol{p}) &= p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M; \boldsymbol{p}) \\ &= \prod_{m=1}^{M} \frac{1}{\sqrt{2^K \pi^K \det\left(\boldsymbol{\Sigma}_m\right)}} e^{-\frac{1}{2}\left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m(\boldsymbol{p})\right)^T \boldsymbol{\Sigma}_m^{-1} \left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m(\boldsymbol{p})\right)}, \end{aligned} \tag{3.36}$$

and using $f(\boldsymbol{p}) = \ln F(\boldsymbol{p})$ and $\mathbb{E}\{\cdot\}$ as the expectation operator, the FIM of (3.29) becomes

$$\boldsymbol{J}(\boldsymbol{p}) = \mathbb{E}\left\{ \left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right) \left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)^T \right\}, \tag{3.37}$$

with the derivatives being evaluated at the true values of $\boldsymbol{p}$. This then reduces to

$$\boldsymbol{J}(\boldsymbol{p}) = \sum_{m=1}^{M} \left(\frac{\partial \boldsymbol{\theta}_m(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)^T \boldsymbol{\Sigma}_m^{-1} \left(\frac{\partial \boldsymbol{\theta}_m(\boldsymbol{p})}{\partial \boldsymbol{p}}\right), \tag{3.38}$$

where $\frac{\partial \boldsymbol{\theta}_m(\boldsymbol{p})}{\partial \boldsymbol{p}}$ is the $K \times 2K$ Jacobian of the DOA vector defined in (3.4). The CRLB is then given by the inverse of $\boldsymbol{J}(\boldsymbol{p})$ as in (3.28).

3.5 Conclusions

This chapter reviewed the current state-of-the-art of DOA-based localization methods. It discussed the fundamental estimators for inferring the location of a single source using DOA measurements and their extensions to the multiple source case, which mainly focus on approaches to address the data-association problem prior or during the localization procedure.

The next chapter will present our proposed grid-based method for sound source localization in WASNs using DOA estimates.

Chapter 4
The Grid-based Method

4.1 Introduction

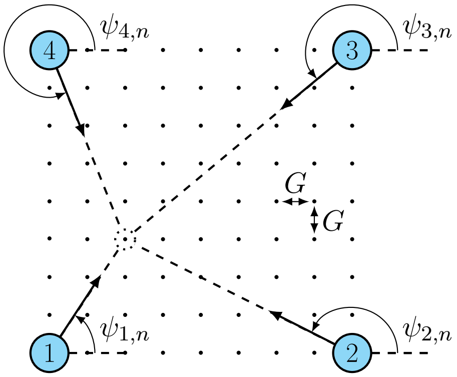
Due to the advantages of DOA-based localization methods that we discussed in Chapter 3, we now focus on this specific class of methods and propose a novel Grid-based (GB) method for the localization of multiple sound sources using DOA estimates. The work presented in this chapter has been published in [85, 86]. For localizing a single source, the grid-based method can be though of as an alternative solution to the non-linear least squares (NLS) estimator of (3.13) that performs much better in terms of computation time without sacrificing any accuracy. The computational efficiency allows this approach to be extended for localizing multiple sources. To do so, we apply the single-source grid based method to each possible combination of DOA measurements from the sensors and solve the data association problem with a computationally efficient method which relies on the estimated locations and the corresponding DOA combinations to decide on the actual source locations. Our approach is real-time and as our simulations and real experiments show, it performs better than state-of-the-art approaches.

Our simulations use new results that we present here to model the DOA estimation error of the algorithm of [18] and consider the problem of missing DOAs—due to missed detections—as a function of source locations which, to the best of our knowledge, has not been widely considered so far. The problem of missing DOAs when the sources are close together occurs very often in practise as our real experiments in this chapter suggest.

This chapter starts by introducing the single-source grid-based method in Section 4.2. Section 4.3 then discusses the extension of the method to multiple sound sources. Finally, Section 4.4 presents simulations and real experiments that evaluate the accuracy and computational complexity of our method.

4.2 Single source grid-based method

The single source grid-based method is an alternative formulation of the NLS estimator of Section 3.3.1, which tries to alleviate the major weaknesses of that approach, namely the



$$\psi_{*,n} = \begin{bmatrix} \psi_{1,n} \\ \psi_{2,n} \\ \psi_{3,n} \\ \psi_{4,n} \end{bmatrix}$$

Figure 4.1: Example cell with four nodes, showing the DOAs to the n-th grid point, and their associated column vector of $\boldsymbol{\Psi}$.

need for a good initial point so as the estimator does not converge to any local minimum, and the computational burden of the minimization procedure.

4.2.1 Single-iteration grid-based method

Our approach is based on making the search space discrete by constructing a grid of N points over the area of interest, and then find the grid point whose DOAs most closely match the estimated DOAs. Moreover, since our measurements are angles, we propose the use of the Angular Distance—defined in Section 3.3.3—as a more proper measure of "similarity" than the absolute distance utilized in (3.13). As we will show later (Section 4.4), this approach is much more computationally efficient, particularly in the multiple source case.

We first form the $(M \times N)$ matrix,

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{1,1} & \psi_{1,2} & \dots & \psi_{1,n} & \dots & \psi_{1,N} \\ \psi_{2,1} & \psi_{2,2} & \dots & \psi_{2,n} & \dots & \psi_{2,N} \\ \vdots & \vdots & & \vdots & & \vdots \\ \psi_{m,1} & \psi_{m,2} & \dots & \psi_{m,n} & \dots & \psi_{m,N} \\ \vdots & \vdots & & \vdots & & \vdots \\ \psi_{M,1} & \psi_{M,2} & \dots & \psi_{M,n} & \dots & \psi_{M,N} \end{bmatrix}, \tag{4.1}$$

where $\psi_{m,n}$ is the DOA from the m-th sensor to the n-th grid point. Note that the n-th column of $\boldsymbol{\Psi}$ is formed from the M DOAs to the n-th grid point, as illustrated in Fig. 4.1.

We then find the index of the grid point whose DOAs most closely match the estimated

DOAs by solving

$$n^* = \arg\min_n \sum_{m=1}^{M} \left[A(\hat{\theta}_m, \psi_{m,n})\right]^2, \tag{4.2}$$

where $\hat{\theta}_m$ is the DOA estimate from the mth sensor and $A(X, Y)$ denotes the angular distance between angles X and Y.

The source location estimate $\hat{\boldsymbol{p}}_{\text{GB}}$ is simply given as the co-ordinates of the n^*-th grid point. Thus, the grid-based method is based on making the space discrete by inserting a grid of candidate locations and finding the source location as the grid-point that minimizes (4.2) through an exhaustive search. We call this the *single-iteration* grid-based approach.

4.2.2 Grid-based error bound

A potential issue with this method is the localization error introduced by the discrete nature of our approach. Any grid-based localization method's accuracy will be limited by the density of its grid points. If we assume that the method works perfectly then the method will exhibit localization error occurred by discretizing the area. We will refer to that error as the *bias* introduced from the use of the grid.

To investigate this further, we now calculate a lower bound for the root mean squared location error. If we assume single source localization, and that the method works perfectly—or that there are no DOA errors—then the method will always choose the grid point closest to the source's location. Let the grid points be uniformly spaced, with G being the inter point spacing in the x and y directions (see Fig. 4.1). Without loss of generality, let us consider a grid point at (0, 0), then due to symmetry, we only need to analyze the squared error in the square defined by (0, 0) and ($G/2$, $G/2$). Let us also assume that a source may be located anywhere in the square under consideration, with a uniform probability density function given by

$$p(x, y) \;=\; p(x) \cdot p(y) \;=\; \frac{2}{G} \cdot \frac{2}{G} \;=\; \frac{4}{G^2}, \tag{4.3}$$

due to the independence between $p(x)$ and $p(y)$. The squared error between (0, 0) and a point (x, y) is simply $x^2 + y^2$, and the mean squared error is then given by

$$E^2_{\text{GB}} = \int_0^{G/2}\!\!\int_0^{G/2} (x^2 + y^2)\, p(x, y)\, dx\, dy \;=\; \frac{G^2}{6}, \tag{4.4}$$

with the root mean square error being

$$E_{\text{GB}} = \frac{G}{\sqrt{6}}. \tag{4.5}$$

If the inter sensor spacing in the x (and y) direction is defined as V (see Fig. 3.1), the number of grid points can be written as

$$N = \left(\frac{V}{G} + 1\right)^2, \tag{4.6}$$

and from (4.4), we can write

$$E_{\text{GB}} = \frac{V}{\sqrt{6}(\sqrt{N} - 1)}. \tag{4.7}$$

Note that this analysis is independent of the method, and should apply to any grid-based localisation method.

4.2.3 Iterative grid-based method

From (4.7) it should be clear that for a cell of given dimensions V, the number of grid points N—determined by the resolution of the grid G (Fig. 4.1)—will determine the method's bias. Increasing N can decrease the location estimation error, as it can make the error occurred from sampling the area significantly small, but it will also increase the complexity of the algorithm.

To maintain a very computationally efficient method when a very dense—i.e, large number of N—grid is considered we propose an iterative solution to (4.2) which starts with a coarse grid (low value of N), and once the best grid point is found, a new grid centered on this point is generated, with a smaller spacing between grid points, but also a smaller scope. Then the best grid point in the new grid is found. This may be repeated until the desired accuracy is obtained, while keeping the complexity under control, as it does not require the exhaustive search over all grid points of the final resolution grid. A possible implementation of the *iterative* grid-based method can be summarized in the following steps:

1. Denote the initial resolution of the grid as G_{initial}, the target resolution as G_{target}
2. Set $G = G_{\text{initial}}$
3. Construct a grid over the area of interest with resolution G
4. Find the grid point n^* by using (4.2)
5. If $G \leq G_{\text{target}}$ goto to step 9
6. Set $V = G$, $G = G/2$
7. Construct a square grid of dimensions V and resolution G centered on n^*
8. Go to step 4
9. Output the co-ordinates of n^* as the estimated location

4.3 Grid-based method for multiple sources

For multiple sources, the grid-based method must account for the fact that the correct association of DOAs to the sources is unknown. The localization consists of a two-step procedure: in the first step, an initial candidate location is estimated for each possible combination of DOA measurements, while in the second step, the final K source locations must be chosen from the candidate locations.

Let $\mathcal{J}$ denote the set of all possible unique combinations of DOA estimates and j enumerate the combinations. Moreover, let $\hat{\boldsymbol{\theta}}^{(j)}$ be the $M \times 1$ vector of DOAs for the j-th combination, and let $\hat{\theta}_m^{(j)}$ denote the DOA of sensor m for the j-th combination. The cardinality of $\mathcal{J}$ depends on the number of sources each sensor is able to detect and can be computed as:

$$|\mathcal{J}| = \prod_{s=1}^{K} s^{C_s} \tag{4.8}$$

As the correct association of the DOAs from the sensors to the sources cannot be known, the single-source GB method of Section 4.2 is applied to each element of $\mathcal{J}$ and the set $\mathcal{L}$ of candidate source locations is formed with $|\mathcal{L}| = |\mathcal{J}|$. Note that this multiple source localization algorithm increases complexity by at least $|\mathcal{J}| - 1$ times that of the single source algorithm, which highlights even more the need for a computationally efficient method to perform the localization of each DOA combination. As we will show later, using the standard NLS approach will significantly increase the computational burden without gaining any increase in accuracy over our approach.

In the next step, the final K source locations must be identified from the set of candidate locations $\mathcal{L}$ by solving the data association problem.

4.3.1 Brute-force approach to the data association problem

A brute-force solution to the data association problem is to perform an exhaustive search over all possible K-tuples of DOA combinations and select the most likely one. A K-tuple of DOA combinations is defined as the list of K DOA combinations (K elements of $\mathcal{J}$) each of them being an $M \times 1$ vector of DOA measurements from the M sensors. Moreover, in forming a K-tuple each sensor must contribute to each of the K DOA combinations with a different estimate, as the same DOA cannot belong to more than one source. In the case where a sensor has not detected all sources the same DOA can be repeated.

The brute-force approach can be summarized in the following steps:

1. Form all possible K-tuples of DOA combinations by combining the elements of set $\mathcal{J}$. The i-th K-tuple will be of the form:

$$T_i = \left\{ \hat{\boldsymbol{\theta}}^{(1)}, \quad \hat{\boldsymbol{\theta}}^{(2)}, \quad \cdots, \quad \hat{\boldsymbol{\theta}}^{(K)} \right\} \tag{4.9}$$

Note that each DOA combination $\hat{\boldsymbol{\theta}}^{(j)}$ is associated with a candidate source location $\boldsymbol{p}^{(j)} = \begin{bmatrix} p_{\mathrm{x},j} & p_{\mathrm{y},j} \end{bmatrix}^T$ in the set $\mathcal{L}$.

2. For each K-tuple i, calculate the sum of residuals of each DOA combination in the tuple as:

$$r_i = \sum_{j=1}^{K} \sum_{m=1}^{M} \left[A(\hat{\theta}_m^{(j)}, \theta_m(\boldsymbol{p}^{(j)})) \right]^2 \tag{4.10}$$

3. Choose the K-tuple that yields the minimum residual and output the corresponding candidate locations from that tuple as the final source locations.

This approach suffers from very high complexity as the number of tuples that need to be tested can grow as high as $\mathcal{O}((S!)^M)$, making this method highly impractical even for a moderate number of sources and sensors. In the next section we propose an alternative way of solving the data association problem that approximates the performance of this brute-force method and is much more computationally efficient.

4.3.2 Sequential approach to the data association problem

In this section, we propose a computationally efficient approach to solve the data association problem. It is a sub-optimal approach to the brute-force method of Section 4.3.1 that relies on a sequential procedure to find the K DOA combinations that approximate the minimum residual of (4.10) without testing all possible K-tuples of DOA combinations.

Our sequential approach can be stated as:

1. Create a set $\mathcal{J}' = \mathcal{J}$.
2. For each DOA combination j in the set $\mathcal{J}'$ compute the residual:

$$r_j = \sum_{m=1}^{M} \left[A(\hat{\theta}_m^{(j)}, \theta_m(\boldsymbol{p}^{(j)})) \right]^2 \tag{4.11}$$

3. Choose the DOA combination j^* with the minimum residual and output the corresponding location $\boldsymbol{p}^{(j^*)}$ as the location of one of the sources.
4. Update $\mathcal{J}'$ by subtracting all DOA combinations that contain DOAs that are part of the previously chosen combination j^*. Only DOAs of the sensors that have not detected all sources are allowed to take part in other combinations.
5. Repeat steps 2–4 until $\mathcal{J}' = \emptyset$ i.e., all K sources have been found.

Note that this approach does not need to test all possible K-tuples of DOA combinations, significantly reducing the computational burden to that of testing only up to $\mathcal{O}(S^M)$ DOA combinations.

4.4 Results and Discussion

In order to investigate the performance of our proposed localization method, we performed simulations and real measurements of a square 4-node cell of a WASN with dimensions of $V = 4$ meters, similar to that of Fig. 3.3.

4.4.1 DOA Estimation Error Modeling

As discussed in Section 3.2, the DOA estimation error at each sensor was assumed to be normally distributed with a zero mean and a certain variance. We also assume that the error variance is dependent only upon the SNR at each sensor, which is in turn determined by the length of the path from the source to the sensor. Given that the signal of a source radiates as a spherical wave, the attenuation experienced by the source signal in travelling from r_1 meters from the source to r_2 meters from the source is given by [87]

$$a = 20 \log_{10} \frac{r_2}{r_1} \quad \text{dB}. \tag{4.12}$$

Thus by specifying a signal-to-noise ratio of the source signal at the sensors when the source is at a specific location, we can determine the SNR at the sensors for every other location through the use of (4.12). We also assume that the accuracy of the each sensor's DOA estimate of a source is determined only by the SNR of that source's signal at that sensor. Our previous work has shown this to be a valid assumption [17, 18].

Following the DOA estimation method of [18], we performed simulations to characterize the DOA estimation error, using sensors consisting of 4-element circular microphone arrays with a radius of 2 cm. We assumed an anechoic environment and simulated various SNR cases ranging from -5 dB to 20 dB. 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. Fig. 4.2 shows the standard deviations obtained when the DOA estimation error at each SNR was fitted with a Gaussian distribution. The fitted curve in Fig. 4.2 is given by

$$f(x) = 1.979e^{-0.2815x} + 1.884. \tag{4.13}$$

It is of note that the standard deviation does not tend to zero at high SNR, which is due to the low number of microphones in each array and the inherent limitations of the real-time method of [17, 18]. By specifying a reference SNR at the center of the cell, the SNR at each sensor can then be calculated through geometry and the use of (4.12), the DOA estimation standard deviation is then taken from the fitted curve of Fig. 4.2. It must be emphasized here that our framework results in a different SNR and, therefore, a different DOA estimation error standard deviation at each sensor.

In order to simulate missed detections, we define the *Minimum Angular Source Separation*

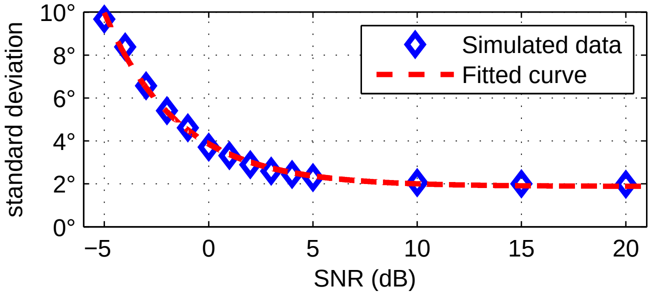


Figure 4.2: Modeling the effect of SNR on DOA estimation error standard deviation for a 4-element circular microphone array with 2 cm radius.

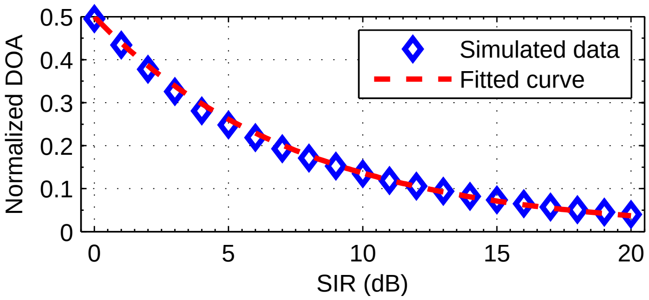


Figure 4.3: Modeling the effect of MASS and SIR on DOA estimation error for a 4-element circular microphone array with 2 cm radius.

(MASS) of a sensor. If the angular distance between two sources is less than the MASS, then the sensor will only detect one source. Also, it was important to model the effect on DOA estimation when two sources were within the MASS of a sensor. We performed a simulation study where two sources were set at various separations of up to 20°—below the MASS of the method of [18]—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—whilst preserving their angular separation—to avoid any orientation biasing effects. In all simulations only one source was detected and Fig. 4.3 shows the results of these simulations, where the DOA offset has been normalized by the separation between the sources. The fitted curve of Fig. 4.3 is given by

$$f(x) = 0.5e^{-0.12987x}. \tag{4.14}$$

It is clear that the detected source's DOA is estimated exactly in the middle of the true DOAs when the sources have equal energy, and moves gradually towards the dominant source as

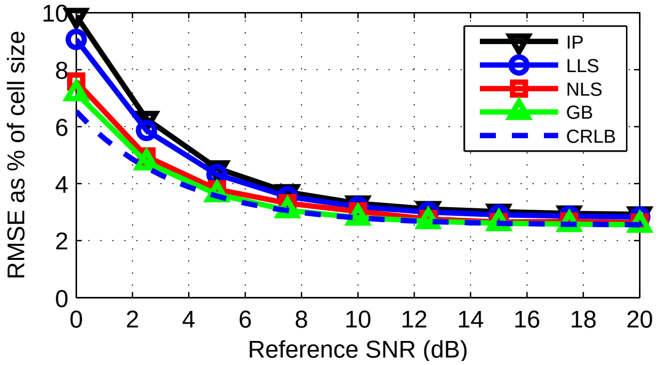


Figure 4.4: Location estimation error as a percentage of cell size $V = 4$ meters for a single source in a square 4-node cell, for various values of SNR measured at the center of the cell. IP: Intersection Point Method, LLS: Linear Least Squares, NLS: Non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

the weaker source decreases in energy. We used the fitted curve of Fig. 4.3 in all simulations involving more than one source.

4.4.2 Simulation Results

In all simulations, the sources were located anywhere within the cell with independent uniform probability and the error measurement used was the Root Mean Square Error (RMSE) between the estimated locations and the true source locations. Fig. 4.4 presents the results of our simulations of a single source. We considered the NLS estimator of (3.13) of Section 3.3.1, the LLS estimator of (3.21) of Section 3.3.2, the intersection point (IP) method of Section 3.3.3, the proposed iterative GB method and the bound of Section 3.3.4. It is clear that all the methods perform close to the bound, with the NLS and GB methods being the closest. However, as we will show later (Section 4.4.3) the GB method is significantly more efficient in terms of computation time. For the IP method $\gamma_{||} = 20°$, and the iterative GB method used an initial and final grid with grid point spacing of 12.5% and 0.25% of the sensor spacing $V = 4$ meters, respectively.

The multiple source localization performance of the P-NLS, IP, and GB methods, as well as the bound of Section 3.4.4 for two and three sources are displayed in Figures 4.5 & 4.6, respectively. As the P-NLS method overestimates the number of sources—estimating a number of ghost sources too—both the P-NLS and GB methods used the brute force approach of Section 4.3.1 for the final source location selection. These results are for the idealized case of 0° MASS, nonetheless, it is very encouraging to see how close the performance of the GB method gets to the lower bound. However, it is evident that the performance of the IP

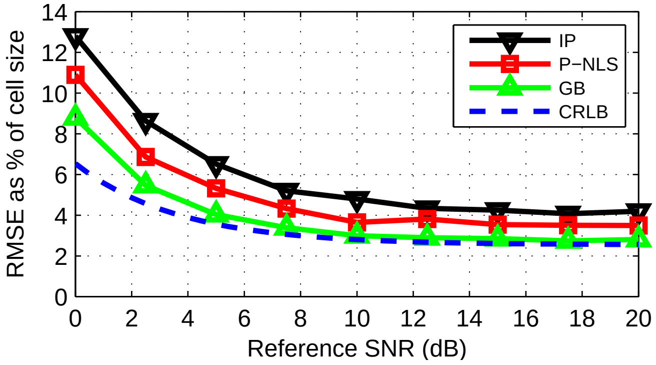


Figure 4.5: Location estimation error as a percentage of cell size $V = 4$ meters for two sources in a square 4-node cell with a MASS of 0°, for various values of SNR measured at the center of the cell. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

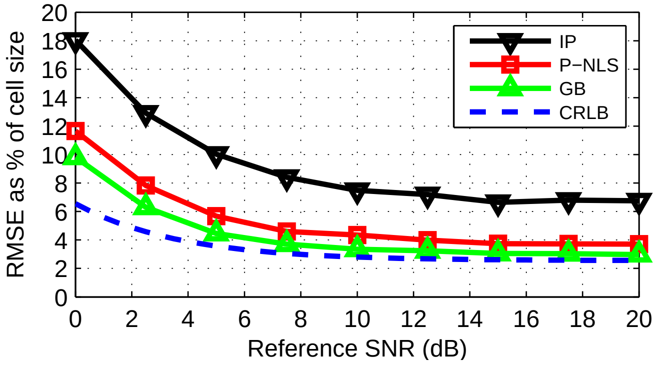


Figure 4.6: Location estimation error as a percentage of cell size $V = 4$ meters for three sources in a square 4-node cell with a MASS of 0°, for various values of SNR measured at the center of the cell. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

method degrades with three sources.

Any realistic sensors and DOA estimation algorithm will have a non-zero MASS, and the performance of all localization algorithms is expected to degrade significantly as the MASS increases. This is due to the fact that the accuracy of the algorithms degrades as C_{K} decreases, and an increasing MASS directly decreases C_{K}, especially as the number of sources increases. Another way to think of this is that as the MASS increases, the accuracy of the DOA estimates from each sensor is much more likely to degrade significantly, due to the "merging" effect illustrated in Fig. 4.3. In the extreme case, C_{K} will be zero—i.e., no sensors will detect the true number of sources—and the localization algorithm will underestimate the number of

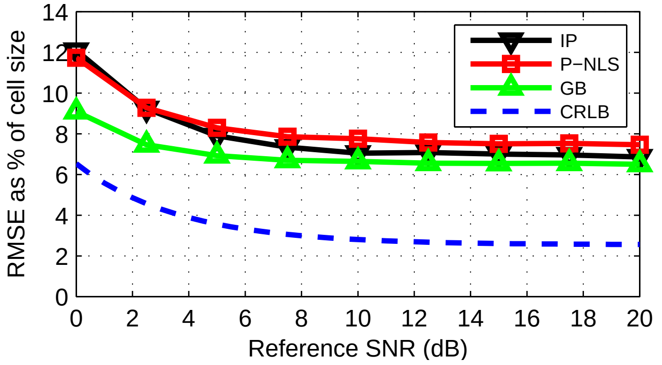


Figure 4.7: Location estimation error as a percentage of cell size $V = 4$ meters for two sources in a square 4-node cell with a MASS of 20°, for various values of SNR measured at the center of the cell. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

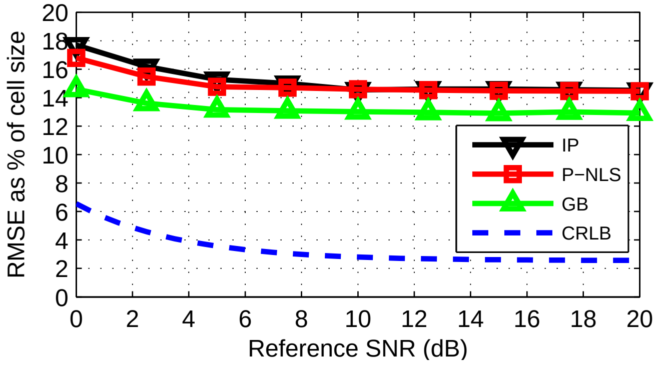


Figure 4.8: Location estimation error as a percentage of cell size $V = 4$ meters for three sources in a square 4-node cell with a MASS of 20°, for various values of SNR measured at the center of the cell. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

source locations. A more realistic case of 20° MASS is presented in Figures 4.7 & 4.8, and the degrading effect of the increased MASS is clear, particularly for the three source case. Note again, that the GB method consistently performs the best.

All the previous results have considered the DOA estimation error at the sensors to be modeled as in Fig. 4.2. In Fig. 4.9 & 4.10 we consider the location error for two sources with increased DOA estimation error when the reference SNR is 20 dB. This is modeled by taking the result of Fig. 4.2 and adding an additional Gaussian noise term with a zero-mean and standard deviation of 1°–10° at each sensor node. Again, in the 0° MASS case, the methods show a reasonable agreement with the lower bound, and as the MASS moves to 20°, the

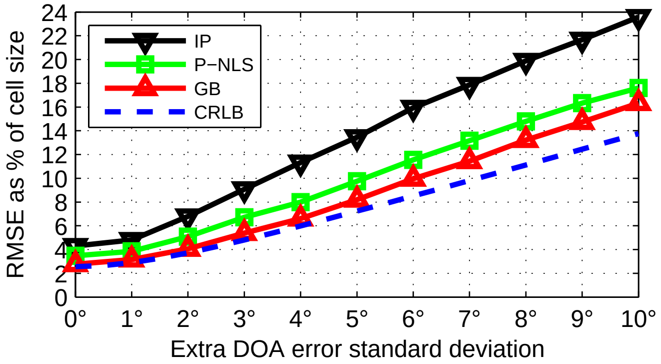


Figure 4.9: Location estimation error as a percentage of cell size $V = 4$ meters for two sources in a square 4-node cell with a MASS of 0° and signals having 20 dB SNR at the center of the cell, for various values of extra DOA error standard deviation. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

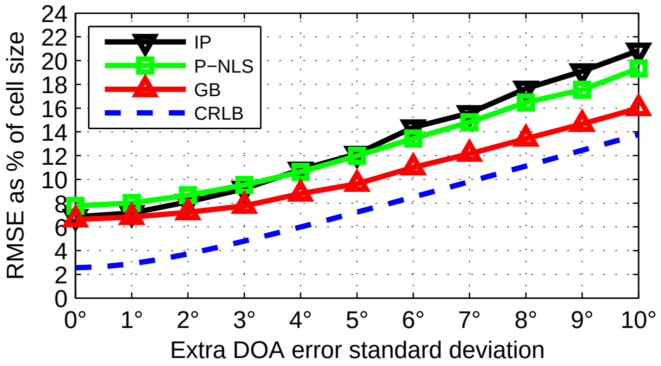
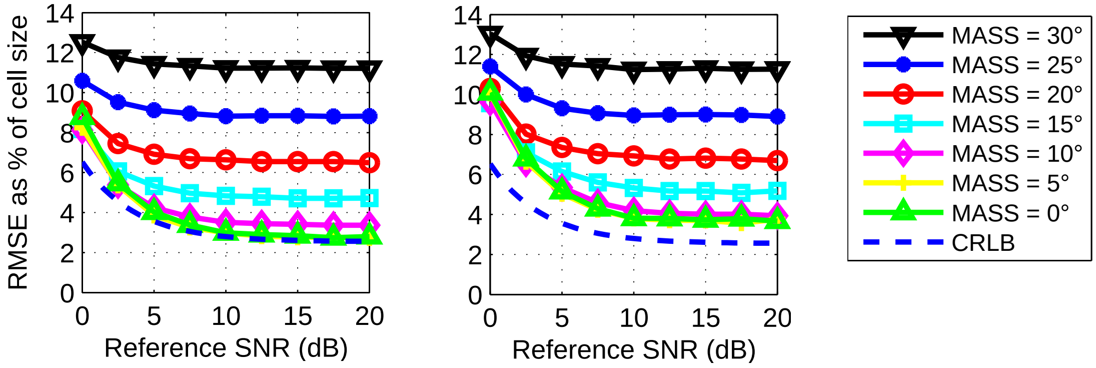


Figure 4.10: Location estimation error as a percentage of cell size $V = 4$ meters for two sources in a square 4-node cell with a MASS of 20° and signals having 20 dB SNR at the center of the cell, for various values of extra DOA error standard deviation. IP: Intersection point method, P-NLS: Position non-linear least squares, GB: Grid-based method, CRLB: Cramer-Rao Lower Bound.

performance of all the methods suffer. Once again, the proposed GB method performs the best with the added DOA estimation error.

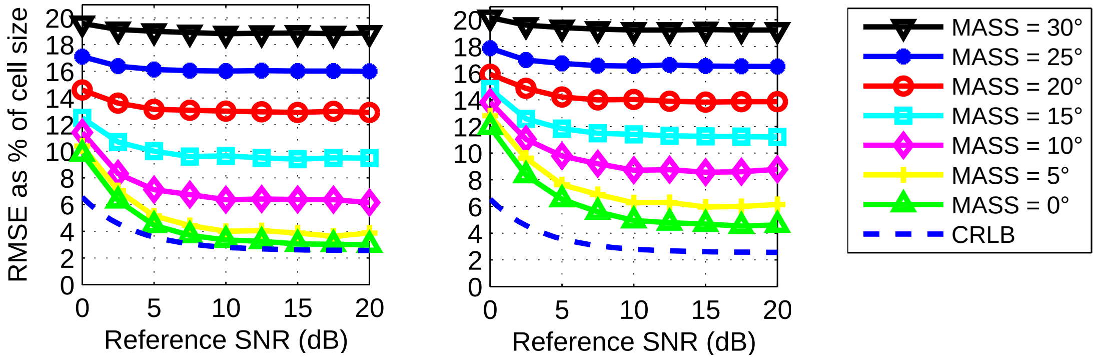
With the sequential approach of Section 4.3.2 we presented a solution to the high complexity brute force approach of 4.3.1, whilst acknowledging that its performance may be worse than the brute force approach. Figures 4.11 & 4.12 illustrate the difference in performance for the two approaches with the GB method for two and three sources, respectively. It is clear that very little performance is lost using the sequential approach particularly at the higher—and more realistic—values of MASS. The loss in performance is higher at low values



(a) Brute force

(b) Sequential

Figure 4.11: Location estimation error as a percentage of cell size $V = 4$ meters for two sources in a square 4-node cell using the grid-based method and the final step approaches of Sections 4.3.1 & 4.3.2 with various values of MASS and SNR measured at the center of the cell.



(a) Brute force

(b) Sequential

Figure 4.12: Location estimation error as a percentage of cell size $V = 4$ meters for three sources in a square 4-node cell using the grid-based method and the final step approaches of Sections 4.3.1 & 4.3.2 with various values of MASS and SNR measured at the center of the cell.

of MASS, and for the three source case. Although it is not shown here, because the P-NLS method must use either the brute force or the sequential approach, it is expected to suffer a very similar performance loss to that of the GB method. Figures 4.11 & 4.12 also illustrate the effect of MASS on the RMSE, highlighting the importance that the DOA estimation used has a low MASS.

Table 4.1: Mean execution times in milliseconds for localization methods for one set of DOA estimations.

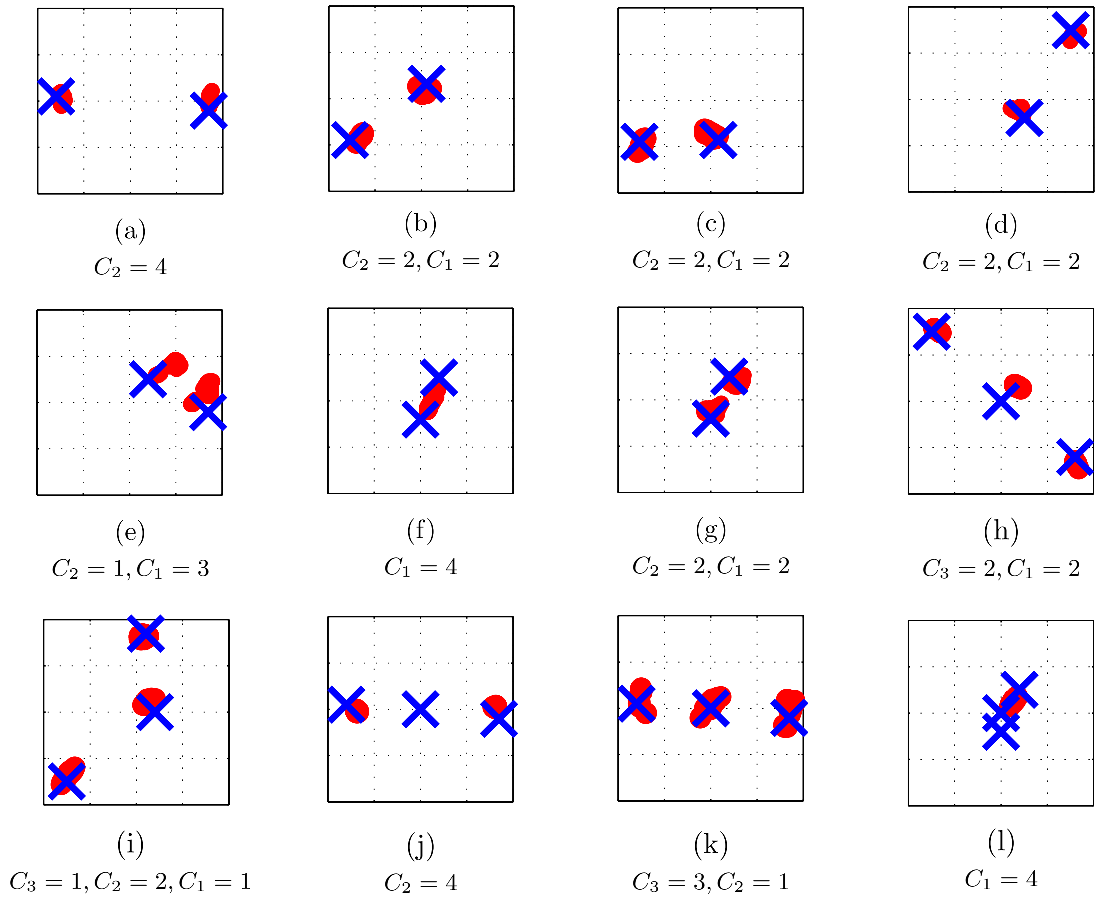
		MASS = 0°		MASS = 20°	
	one	two	three	two	three
Method	source	sources	sources	sources	sources
LLS	0.12	–	–	–	–
IP	0.69	6.89	44.49	5.31	16.16
GB (& BF)	1.72	36.03	2961.57	19.18	214.34
GB (& Seq.)	1.72	29.39	162.79	16.79	26.69
P-NLS (& BF)	18.88	381.95	5033.43	205.12	509.59
P-NLS (& Seq.)	18.88	375.32	2238.82	202.72	322.08

4.4.3 Complexity

All the localization algorithms under evaluation were implemented in Matlab on a Windows laptop with a Core i5 CPU running at 2.53 GHz with 4GB RAM, and their mean execution times are presented in Table 4.1. Note that while the absolute execution times may be highly dependent on the machine, we are only interested here in the relative times between the methods. In the one source case the LLS method is clearly the fastest, while the IP method is the fastest in the multiple source cases. The (P-)NLS methods are clearly the slowest methods; this is due to non-linear optimization they require. Table 4.1 also highlights the dramatic reduction in complexity when using the sequential rather than the brute force approach for the data association problem, particularly in the three source case. These results, together with those of Section 4.4.2, strongly suggest that the GB method with the sequential approach is the best choice given its accuracy and moderate complexity. To further verify its suitability, we implemented the GB method with the sequential approach in C++ and measured that it only consumed 25% of the available processing time, making it an excellent candidate for a real-time system.

4.4.4 Results of Real Measurements

We also performed some real recordings of omnidirectional acoustic sources in a 4-node square cell with sides 4 meters long. The sensors on the nodes were circular 4-element microphone arrays with a radius of 2 cm, and the DOA estimation was performed by our real-time system of [17, 18]. The sources were recorded speech played back simultaneously through loudspeakers at different locations, and their SNR at the center of the cell was measured to be about 10 dB. For playback we used a loudspeaker facing towards the sky in order to simulate the directivity pattern of an omnidirectional source. Although a 4×4 metre square is not a particularly large area, since we measure our reference SNR at the center of the



(a) $C_2 = 4$

(b) $C_2 = 2, C_1 = 2$

(c) $C_2 = 2, C_1 = 2$

(d) $C_2 = 2, C_1 = 2$

(e) $C_2 = 1, C_1 = 3$

(f) $C_1 = 4$

(g) $C_2 = 2, C_1 = 2$

(h) $C_3 = 2, C_1 = 2$

(i) $C_3 = 1, C_2 = 2, C_1 = 1$

(j) $C_2 = 4$

(k) $C_3 = 3, C_2 = 1$

(l) $C_1 = 4$

Figure 4.13: Location estimates (the red clouds) using the proposed grid-based method in a square 4-node cell, for real recordings of two [(a)–(g)] or three [(h)–(l)] simultaneous sources (the blue X's).

cell, these results should be scalable to larger cells. Fig. 4.13 shows the location estimates from the real recordings using the proposed grid-based method for different layouts of two and three sources. The red dots show the cloud of estimates over about 5 seconds, and show quite accurate localization. The pairs (f) & (g) and (j) & (k) warrant further discussion. All of the plots *except* (g) and (k) used the standard parameter set of [17,18] which has a MASS of around 20°, and it is clear that in (f) and (j) the source locations are underestimated. By modifying some of the parameters of the DOA estimation, were we able to decrease the system's MASS so that all the sources in (g) and (k) could be localized, albeit with a greater variance in the estimates.

It should be noted that these recordings took place outdoors, and as such did not have many reflections, but there was a significant level of distant noise sources, such as cars and dogs barking. Furthermore, the orientations of the sensors were not finely calibrated, and the DOA estimates likely have unintended offsets of a few degrees. Thus the conditions were far from ideal, making the results of our proposed localization method even more encouraging.

4.5 Conclusions

In this chapter we presented the grid-based method for multiple source localization in a wireless acoustic sensor network using DOA estimates. For the single source case, our grid-based method is a computationally efficient non-linear least squares estimator. For multiple sources, the grid-based method estimates a location for each possible DOA combination from the sensors and then decides which locations correspond to the locations of the sources using the brute-force or sequential approaches described in Section 4.3. Our performance evaluation using both simulated and real recorded signals showed that our method outperforms other state-of-the-art methods both in accuracy and computational complexity. It also revealed that a determinant factor that degrades performance is the data-association problem in scenarios with missed detections. Thus, in the next chapter we will focus on the data-association problem and propose a solution that can find the correct association of DOAs from the nodes to the sources prior to the location estimation task and accurately localize multiple sources.

Chapter 5
The data-association problem

5.1 Introduction

As mentioned in Chapter 3 the major difficulty when localizing multiple sources using DOA estimates arises due to the data-association problem. The problem occurs because the central node that receives the multiple DOA estimates (one for each detected source) from the sensors cannot know to which source they belong. The correct association of DOAs across the sensors that corresponds to the same source must be found, otherwise location estimation will result in "ghost sources", i.e., location estimates that do not correspond to the locations of the real sources.

The problem becomes even more challenging when considering more realistic scenarios where missed detections occur and thus some sensors may not detect all the sources, underestimating their number. In this scenario the central node not only is not aware of the correct association of DOAs to the sources, but the number of DOA estimates across the sensors may also vary as a result of missed detections.

The Grid-based method that we developed in Chapter 4 for localizing multiple sound sources accounts for the data-association problem by applying a localization step to estimate a candidate location for every possible DOA combination from the sensors. It then decides which of the candidate location estimates correspond to real sources based on the location estimates and their corresponding DOA combination using the brute-force (Section 4.3.1) or the sequential (Section 4.3.2) approach.

However, when the number of sources is increased, a performance degradation is evident because the method cannot correctly identify the locations that correspond to the real-sources (data-association problem), but also as a result of increased missed detections. This performance degradation is evident in all the results presented in Chapter 4, where missed detections have been modeled using the Minimum Angular Source Separation (MASS) that defines the minimum angular separation that two sources must have so as to be both detected by the DOA estimation method. It can be observed that the methods' performance severely degrades when the MASS is increased.

In this chapter, we develop a novel approach for addressing the data-association problem, taking also into account the missed detections that can occur in a realistic scenario. The approach presented in this chapter has been published in [88,89]. The proposed method first estimates the correct association of DOAs from the sensors to the sources. When this association is identified, localization is achieved by applying any single-source location estimator to the resulted DOA associations. The method utilizes additional information—apart from the DOA estimates—transmitted by the sensors. In this spirit, each sensor estimates and transmits features associated with each source it detects. For the same source, such features must be "similar" across the different sensors. We denote such features as the *association features*. We propose the use of features that describe how the frequency components of the captured signals at each node are distributed to the sources and we show that these features are robust to missed detections, reverberation, noise, and moving sources.

The association of DOAs to the sources is found by comparing the corresponding features and separating them into groups according to their similarity. As each feature is connected to a source's DOA, the grouping of the features reveals the association of DOAs from the sensors to the sources. For this grouping procedure, we propose a greedy association algorithm which can work with an arbitrary number of sensors and sources, results in high association accuracy and is computationally efficient.

As the features need to be transmitted to the fusion center for the association procedure to be performed, we also study how to reduce the amount of information that needs to be transmitted. We propose a scheme that reduces the bitrate requirements of our method up to 88% without affecting its accuracy. Our experimental evaluation also reveals how often missed detections occur in practice, which highlights once more the need for the data-association and localization method to take into account missed detections.

The remainder of this chapter is organized as follows: Section 5.2 describes the proposed feature extraction procedure which is carried out in each node (microphone array) individually. The features are transmitted to the fusion center which performs the data-association based on the algorithm described in Section 5.3. Section 5.4 presents an approach to reduce the bitrate requirements of the proposed methodology. Finally, Section 5.5 presents the performance evaluation of our method.

5.2 In-node feature extraction

The feature extraction is based on the assumption that each time-frequency bin belongs to at most one source. The assumption that only one source is dominant in each time-frequency bin is known as the W-disjoint orthogonality (WDO) assumption and has been shown to be valid especially for speech signals [83,84]. The feature computation is based on the estimation of narrowband DOA estimates in each time-frequency bin and the assignment of each bin to a source based on the corresponding estimated DOA. The features are computed for each time-frame and represent the number of times each frequency bin was assigned to a source in

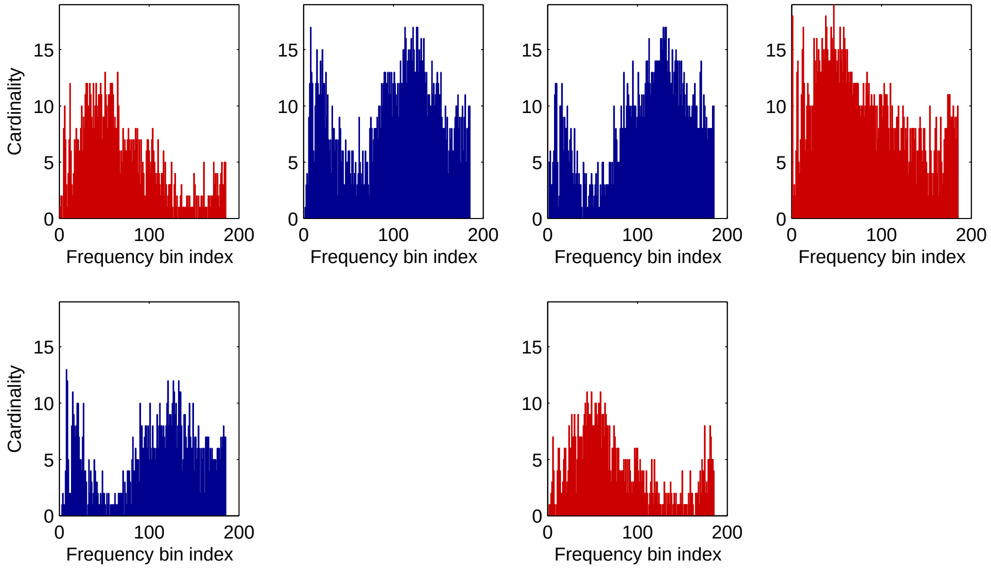


Figure 5.1: Example of association features for a WASN with $M = 4$ arrays (columns) and $K = 2$ sound sources (rows). The colors indicate the association features that correspond to the same source in different arrays and are most similar to each other. In this example, array 2 and array 4 have exhibited missed detections thus detecting and estimating the corresponding association feature for only one source.

the last B frames, where B refers to the *history length* or *block size*. Each feature is associated to a source and provides an estimate of the distribution of the frequency components to that source.

The microphone array signals at each array m are transformed into the Short-Time Fourier Transform (STFT) domain, resulting in the signals $X_{m,i}(\tau, \ell)$ where i is the microphone index and τ, ℓ denote the time-frame and frequency bin index respectively. In the sequel, we omit m and τ as the procedure is repeated in each array for each time-frame. We also denote as L the set of frequency bins ℓ up to a maximum frequency $\ell_{\max}$. We set $\ell_{\max}$ to the spatial-aliasing cutoff frequency which depends on the array geometry and describes the frequency up to which reliable DOA estimates can be found.

In each frequency $\ell \in L$ we estimate a DOA, resulting in the narrowband DOA estimates $\phi(L)$. For the narrowband DOA estimation any method available in the literature can be utilized. Also, a broadband DOA estimation method estimates the number of detected sources $\hat{K}$ and their corresponding DOAs $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_{\hat{K}}\}$. This can be achieved with an arbitrary broadband DOA estimation method, e.g., by processing the narrowband DOA

Algorithm 1 Feature computation at the mth node

Input: Frame of microphone array signals $\mathbf{X}_{\text{fr}}(\ell)$ in the frequency domain, History length B, User-defined threshold ϵ

Output: Association Features $F_{m,k}(\ell)$ for each detected source k

```
for each frequency bin $\ell$ in $L$ do
  $\phi(\ell)$ = Narrowband_DOA_Estimation($\mathbf{X}_{fr}(\ell)$)
end for
$(\hat{\boldsymbol{\theta}}, \hat{K})$ = Broadband_DOA_Estimation($\mathbf{X}_{fr}$)
$F_{m,k}(\ell) = \mathbf{0}, k = 1, \ldots, \hat{K}$
for each frame $\tau'$ between the current frame $\tau$ and $B$ previous frames do
  for each frequency bin $\ell$ in $L$ do
    $k \leftarrow \arg\min_p \left( A(\phi_{\tau'}(\ell), \hat{\theta}_p \right)$
    if $A\left(\phi_{\tau'}(\ell), \hat{\theta}_k\right) < \epsilon$ then
      $F_{m,k}(\ell) \leftarrow F_{m,k}(\ell) + 1$
    end if
  end for
end for
```

estimates with a matching-pursuit algorithm as proposed in [16, 18].

Then, the frequencies in L are assigned to the detected sources. The assignment is based on the DOAs of the sources $\hat{\boldsymbol{\theta}}$ at the current frame and the narrowband DOA estimates in each frequency bin $\phi_{\tau'}(\ell)$ for each frame τ' between the current and B previous frames. A frequency bin $\ell \in L$ is assigned to source p (with corresponding direction $\hat{\theta}_p$) if the following two conditions are met:

$$A(\phi_{\tau'}(\ell), \hat{\theta}_p) < A(\phi_{\tau'}(\ell), \hat{\theta}_q), \quad \forall q \neq p, \tag{5.1}$$

$$A(\phi_{\tau'}(\ell), \hat{\theta}_p) < \epsilon, \tag{5.2}$$

where $A(X, Y)$ denotes an angular distance function that returns the difference between X and Y in the range of $[0, \pi]$. In other words, Eqs. (5.1) and (5.2) suggest that a given frequency bin is assigned to the source whose DOA is closest to the estimated DOA at that bin, as long as that distance does not exceed a pre-defined threshold ϵ. When Eq. (5.2) is not satisfied, the given frequency bin is rejected and not assigned to any of the sources.

Since the assignment is carried out for the frequency bins for the current and B previous frames, a histogram can be formed for each detected source that counts how many times each frequency bin was assigned to that source. Note that since B frames are considered, a frequency bin can be assigned to a source up to B times. These histograms constitute the proposed association features which are transmitted to the fusion center together with the estimated DOAs for the detected sources. The proposed feature extraction procedure is presented in Algorithm 1.

Since all the arrays receive the same signals—albeit with relative phase differences—the histograms across the arrays that belong to the same source are expected to be similar. As each histogram is associated with a source DOA, the grouping of the histograms in K groups, based on their similarity, is expected to reveal the association of DOAs from the arrays to the K sources.

An example of these association features is shown in Fig. 5.1 for a WASN of 4 microphone arrays with two active sound sources. The colors indicate the histograms that correspond to the same source at the different arrays. The association features (i.e., histograms) that correspond to the same source are expected to be "similar". In this example, two arrays have exhibited missed detections thus being able to detect only one source and thus estimating a single association feature.

5.3 Data-association algorithm at the fusion center

Given the estimated association features, the goal of the association algorithm, which is carried out at the fusion center, is to group them according to their similarity in K groups. We remind the reader that K is the number of sources which is assumed to be known. As each feature is associated with a DOA estimate, the grouping of the features will reveal the association of DOAs from the arrays to the sources, which is indicated by the different colors in the example of Fig. 5.1. In this section, we formally define the data-association problem as an assignment problem of the association features to K groups.

Let $\boldsymbol{F}$ denote the set of all association features and $\boldsymbol{F}_m$ denote the set of features $F_{m,k}$, $k = 1, \ldots, \hat{K}_m$ for all detected sources from the mth array. The problem is to find a set $\boldsymbol{G}$ that contains K groups of features, denoted as G_i, $i = 1, \ldots, K$, $G_i \in \boldsymbol{G}$, such that:

(a) features from the same array cannot be assigned to the same group,

(b) each feature must be assigned to exactly one group, and

(c) all groups contain features that are "similar" to each other.

We call the set $\boldsymbol{G}$ an assignment of features to groups. As each feature corresponds to a source DOA, the resulting K groups provide the association of DOAs across the arrays for the K sources. We proceed by proposing and defining a way to measure the quality of an assignment.

Let D be a function measuring the dissimilarity of two features, taking values in $[0, 1]$. We define the *score* of each group G_i as the maximum pairwise dissimilarity of its contained features:

$$Score(G_i) = \max_{p,q} D(F_p^i, F_q^i), \tag{5.3}$$

where F_p^i denotes the pth feature of group G_i.

We define the overall score of an assignment $\boldsymbol{G}$ as the maximum score among the scores

Algorithm 2 Association Algorithm

```
Input: Features $\mathbf{F}$ , Number of Sources $K$
Output: Assignment $\mathbf{G}$
  $\mathbf{S} \leftarrow \bigcup_i CreateInitialAssignment(\mathbf{F}_i, K)$ (Fig. 5.2(a))
  while $|\mathbf{S}| > 1$ do (Fig. 5.2(b))
    $(i, j) \leftarrow \arg\min_{i,j} Score(Merge(\mathbf{S}_i, \mathbf{S}_j))$
    $\mathbf{S} \leftarrow \mathbf{S} \setminus (\mathbf{S}_i \cup \mathbf{S}_j) \cup Merge(\mathbf{S}_i, \mathbf{S}_j)$
  end while
  $\mathbf{G} \leftarrow GetFinalAssignment(\mathbf{S})$ (Fig. 5.2(c))
  while $\min_{p,q} Score(\mathbf{G}[F_p^i \leftrightarrow F_q^j]) < Score(\mathbf{G})$ do
    $\mathbf{G} \leftarrow \mathbf{G}[F_p^i \leftrightarrow F_q^j]$
  end while
```

of its contained groups $G_i \in \boldsymbol{G}$:

$$Score(\boldsymbol{G}) = \max_i Score(G_i). \tag{5.4}$$

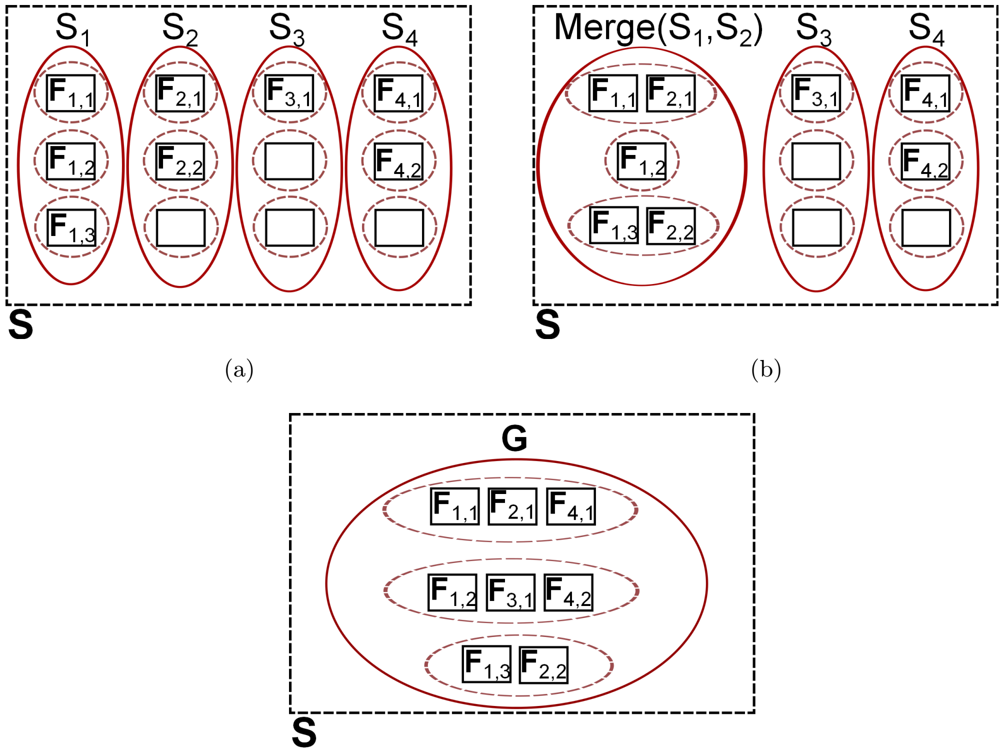
Our goal is to find an assignment that minimizes (5.4), while satisfying constraints (a) and (b) mentioned above. Thus, the solution to the data-association problem can be formally defined as:

$$\arg\min_{\boldsymbol{G}} Score(\boldsymbol{G}). \tag{5.5}$$

In case two assignments result in the same score, we sort the scores of their groups in descending order and compare their maximum non equal score. The motivation behind this formulation of the data-association problem is that we want to find an assignment where *all groups* contain features that are as similar as possible to each other, since the contained features in each group correspond to the same source.

The next step is to deduce an algorithm that can efficiently solve (5.5). A straightforward approach would be to exhaustively enumerate all possible assignments and choose the one that satisfies (5.5). Although it can guarantee to find the assignment with the minimum score, such a naive brute-force approach cannot be realized in practice due to its prohibitively large computational requirements, as the number of possible assignments can grow as $(K!)^M$. Next, we derive a greedy algorithm that can efficiently solve (5.5). The algorithm—shown in Algorithm 2—does not necessarily identify the optimal solution, but it is simple, fast, and as our experimental results indicate, it finds good solutions in practice.

First, for each set $\boldsymbol{F}_m$, $m = 1, \ldots, M$ we create an assignment $\boldsymbol{S}_m$. The assignment contains K groups, where each group contains a single feature. If the array has exhibited missed detections, thus having estimated less than K features, some groups are left empty. This procedure is illustrated in Fig. 5.2(a). The algorithm then tries to greedily merge those



(a)

(b)

(c)

Figure 5.2: Example of the association algorithm for $M = 4$ arrays and $K = 3$ sources. (a) First, an assignment of features to K groups is created for the set of features $\mathbf{F}_m$ for each array. The empty boxes represent the empty groups, as the corresponding arrays have detected less than K sources. (b) The algorithm finds the assignment (array 1 and 2 in this example) that, when merged, produce the best score according to (5.5) and merges them. (c) The merging operations stop when a single assignment $\mathbf{G}$ remains.

assignments $\boldsymbol{S}_m$ until only one remains. The merging is done by considering all possible ways to merge two assignments and selecting the one which produces the best possible score according to (5.5) (Fig. 5.2(b)). The possible ways to merge the assignments equals $K!$. This is known as the *Linear Bottleneck Assignment Problem* [90], which can be solved efficiently in polynomial time. However, when K is relatively small, a brute-force approach is often faster. The algorithm finishes when only a single assignment, $\boldsymbol{G}$, remains (Fig. 5.2(c)).

Then, in order to further refine the estimated assignment, we perform a second greedy step. In this step, we select two features F_p^i and F_q^j from different groups i and j and try to swap them in order to further reduce the score of $\mathbf{G}$; for brevity, we use $\mathbf{G}[F_p^i \leftrightarrow F_q^j]$ to refer to the new assignment in which F_p^i and F_q^j are interchanged. We also allow one of them to be empty or, in other words, to move a feature from one group to another. The algorithm terminates if no such pair exists. As our results in Section 5.5.3 indicate, this second greedy step was found to result in significant performance gain.

5.4 Reducing transmission requirements

In terms of transmission requirements, each array must transmit the histogram (association feature) for every source it detects to the fusion center. In this section, we quantify the transmission requirements of our method and propose how to reduce the amount of information that needs to be transmitted.

The number of bins in a histogram equals the number of frequency bins that are processed. As discussed in Section 5.2, the processing for the extraction of the histograms is performed for a frequency range up to frequency $\ell_{\max}$. Let N_ℓ denote the number of frequency bins available for processing, i.e., the number of frequency bins up to frequency $\ell_{\max}$. Given a history length of B frames, the maximum cardinality of a given bin in the histogram is B. Thus the number of bits required to transmit a histogram is $\lceil N_\ell \log_2(B) \rceil$.

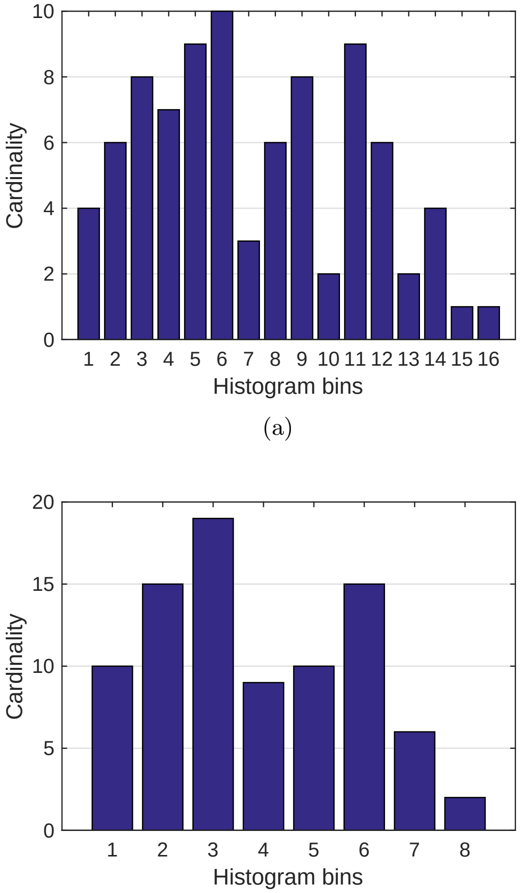
We propose to reduce the transmission requirements by reducing the number of bins in the histograms by performing a decimation process as follows: Let $h : \mathcal{A} \to \mathcal{B}$ be a function describing a histogram where its domain corresponds to the frequency bin indices, i.e., $\mathcal{A} = \{1, \ldots, N_\ell\}$ and its range corresponds to the cardinality of each bin, i.e., $\mathcal{B} = \{0, 1, \ldots, B\}$. A decimated version of h by a factor of d can be formed by grouping each consecutive d bins and summing their cardinalities according to:

$$h'(x) = \sum_{k=1}^{d} h\left(d(x-1) + k\right), \tag{5.6}$$

where $h' : \mathcal{C} \to \mathcal{D}$ with $\mathcal{C} = \{1, \ldots, \lceil \frac{N_\ell}{d} \rceil\}$ and $\mathcal{D} = \{1, \ldots, d \cdot B\}$. After the decimation process the domain of the new histogram has been shrunk by a factor of d and the range of possible values for each bin is now $d \cdot B$. The number of bits N_b required to transmit the decimated histogram depends on the history length B, the number of bins N_ℓ, and the decimation factor d through:

$$N_b = \lceil \frac{N_\ell}{d} \log_2(d \cdot B) \rceil. \tag{5.7}$$

The initial histogram, without any decimation, corresponds to the case where $d = 1$. An example of a histogram and its decimated version by a factor of $d = 2$ is shown in Fig. 5.3. In this example, the initial histogram has $N_\ell = 16$ bins and its maximum cardinality is 10, thus requiring $\lceil 16 \log_2(10) \rceil = 54$ bits for transmission. After decimation by $d = 2$ the number of bins have reduced to $\frac{N_\ell}{d} = 8$ and the maximum cardinality is now 20, thus requiring 35 bits. As we will show in Section 5.5.5, we can apply a decimation process by a factor of 16 to our histograms, thus significantly reducing the amount of information that needs to be



(a)

(b)

Figure 5.3: Example of a histogram with $N_\ell = 16$ bins (a) before and (b) after a decimation process by a decimation factor of $d = 2$.

transmitted in the network, without degradation in performance.

5.5 Evaluation

To evaluate the proposed method we performed simulations on a square cell of a WASN with dimensions of $V = 4$ meters with $M = 4$ nodes which were configured as shown in Fig. 5.4. Each node was an 8-element uniform circular microphone array with 5 cm radius. The sound sources were speech recordings of 2 seconds duration, sampled at 44.1 kHz, and had equal power when placed at the center of the cell. The signal-to-noise ratio (SNR) was measured as the ratio of the power of each source signal when located at the center of the cell to the power of the noise signal. To simulate different SNR values we added white Gaussian noise at each

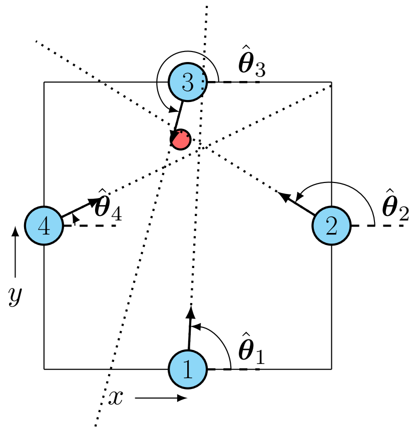

3
$\hat{\boldsymbol{\theta}}_3$
4
$\hat{\boldsymbol{\theta}}_4$
2
$\hat{\boldsymbol{\theta}}_2$
y
1
$\hat{\boldsymbol{\theta}}_1$
x

Figure 5.4: WASN cell used for the experimental evaluation. The network consists of $M = 4$ microphone array nodes (blue circles, numbered 1–4).

microphone, uncorrelated with the source signals and the noise at the other microphones. Note that this framework results in different SNR at each array depending on how close the source is to the arrays.

We simulated a room of dimensions of $10 \times 10 \times 3$ meters using the Image-Source method [91] and produced signals of omnidirectional sources at different reverberation conditions. The WASN cell was placed at the middle of the room. Both the nodes and the sources were placed at 1.5 m. height. More specifically, the nodes were placed at $(5, 3, 1.5)$, $(7, 5, 1.5)$, $(5, 7, 1.5)$, and $(3, 5, 1.5)$. In terms of number of sources, we considered scenarios of two and three simultaneously active speakers. Each simulation was repeated 30 times and the sources were placed at different locations within the cell with independent uniform probability. For narrowband DOA estimation we used the method proposed in [4], which is designed for the uniform circular array geometry. For the estimation of the broadband DOAs of the sources in each time frame, we applied our previously proposed methodology of [16, 18]. Note that in our evaluation, we use circular microphone arrays and the DOA estimation methods employed are tailored for this specific array geometry. However, the proposed methodology is independent of the array geometry and the DOA estimation method in the sense that any DOA estimation method available in the literature can be employed to infer the narrowband and broadband DOA estimates.

For processing we used frames of 2048 samples with 50% overlap. The FFT size was 2048. We set $\ell_{\max}$ to 4 kHz which is the spatial aliasing cutoff frequency for our given array geometry. The threshold ϵ for the frequency assignment in Eq. (5.2) was set to 10° and we used a history length of $B = 21$ frames, which corresponds to 0.5 seconds. As a dissimilarity

Table 5.1: Experimental Parameters

parameter	notation	value
Room		10 × 10 × 3 m.
WASN cell		square
WASN side length	V	4 meters
Node type		8-element uniform circular array, 5 cm radius
Number of nodes	M	4
Framesize		2048 samples
Overlapping in time		1024 samples
FFT size		2048 samples
Sampling frequency	F_s	44.1 kHz
Highest frequency for processing	$\ell_{\max}$	4 kHz
Threshold for frequency assignment	ϵ	10°
History length (block size)	B	21 frames (0.5 sec.)
Decimation factor	d	1

measure in (5.3), we used the Pearson Correlation Coefficient distance which is defined as [92]:

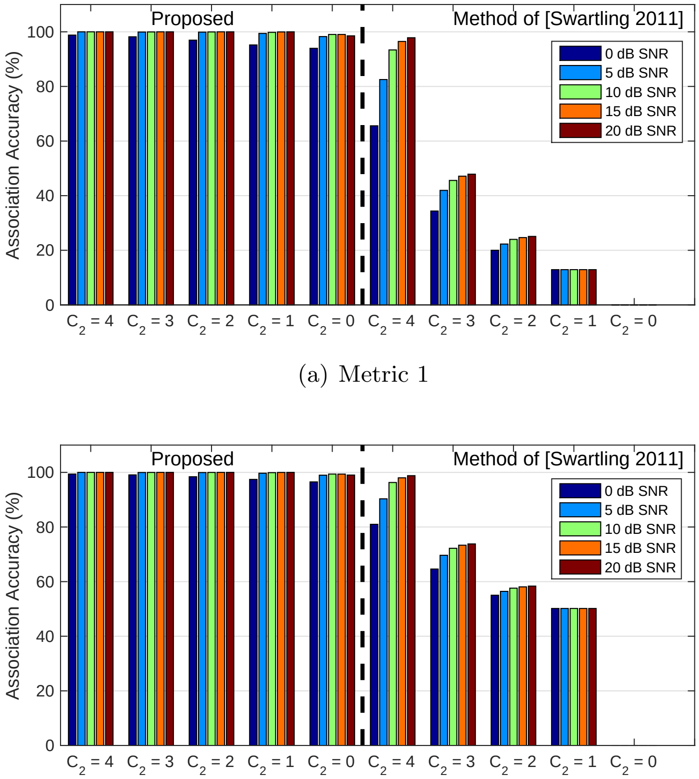
$$D(X, Y) = \frac{1 - r_{X,Y}}{2}, \tag{5.8}$$

where $r_{X,Y}$ is the Pearson correlation coefficient between X and Y. Eq. (5.8) takes values in the range $[0, 1]$. The parameters are summarized in Table 5.1 and are used throughout our experimental evaluation, unless stated otherwise.

5.5.1 Evaluation Metrics

To measure the association accuracy we utilize two metrics. The first is denoted as *Metric 1* and measures the percentage of time frames where a correct DOA association is found. We define a DOA association as correct when *all* DOAs are assigned to the correct source from *all* the arrays.

When an association is not correct, it means that some DOAs from some arrays are assigned to an erroneous source. However, although the association is erroneous there are still pairs of DOAs from different arrays that were associated correctly. As an example, let us assume that in Fig. 5.2(c) the DOA that corresponds to feature $F_{4,1}$ was erroneously assigned to the source that corresponds to the first group. While this association is erroneous—according to Metric 1—there are pairs of DOAs from arrays that are associated correctly,



(a) Metric 1

(b) Metric 2

Figure 5.5: Data-association accuracy for two sources in an anechoic environment for different values of SNR and C_2

.

such as the pairs $(F_{1,1}, F_{2,1})$, $(F_{1,2}, F_{3,1})$,$(F_{3,1}, F_{4,2})$, $(F_{1,2}, F_{4,2})$, and so on, while other pairs are associated erroneously, such as pairs $(F_{1,1}, F_{4,1})$, $(F_{2,1}, F_{4,1})$. Of course, the more these correct pairs of DOAs are, the less impact an erroneous pair will have to the data-association and thus to the localization error. To quantify the correct "parts" of a DOA association—that it can albeit be erroneous according to the definition of Metric 1—we use our second metric (denoted as *Metric 2*), which counts the percentage of correct pairwise associations between all pairs of arrays.

5.5.2 Robustness to missed detections

First, we evaluate the efficiency of our proposed association features and our proposed association algorithm in scenarios with missed detections. We assume that the DOAs of the sources in each time-frame, i.e., vectors $\hat{\boldsymbol{\theta}}$ at each array are known. We define C_s as the number of arrays that detected s sources, i.e., $C_2 = 3$ indicates that three arrays detected two sources. To simulate missed detections, we fix C_s and remove some DOAs from some arrays until the desired value of C_s is reached. The removed DOAs as well as the arrays that exhibit the missed detections are selected at random in every frame, under the constraint that each source must be detected by at least one array which is a necessary condition to find the DOA associations for all active sources.

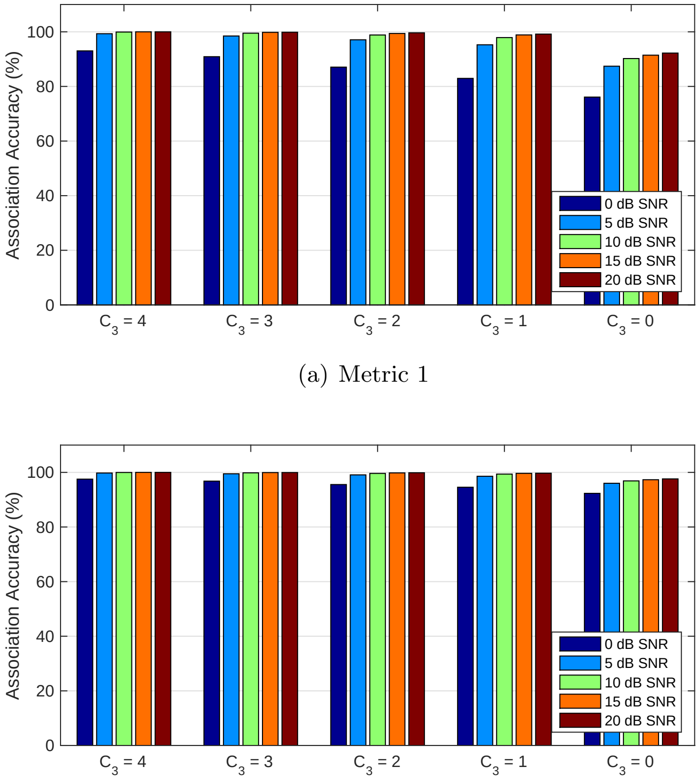
Fig. 5.5 depicts the data-association accuracy (using the two aforementioned metrics) for an anechoic scenario of two active sound sources for all possible values of C_2, i.e., the number of arrays that detected the two sources. For comparison, Fig. 5.5 also presents the results using the association features proposed in [81] (denoted as [Swartling 2011] in the figure). These features were modified to work with circular microphone arrays. For association, we applied our proposed association algorithm on the features extracted from [81], as the association algorithm proposed in [81] works only for the case of two arrays. From Fig. 5.5 it is evident that our approach is robust to missed detections, achieving more than 90% accuracy for all SNR cases and all values of C_2.

On the other hand, the association features proposed in [81] are less robust to noise and missed detections. A severe performance degradation is evident, especially when missed detections are present ($C_2 < 4$). A key reason for that is the fact that the association features were not designed to handle missed detections: when a source is not detected, the method of [81] erroneously assigns its frequencies to the other sources, thus degrading the association performance. Our proposed method avoids such erroneous assignments though the use of Eq. (5.2). It is noteworthy that our proposed approach can accurately find the correct association of DOAs to the sources, even in the extreme case where all arrays detected only one source, i.e., $C_2 = 0$. Finally, the features of [81] cannot handle the case where $C_2 = 0$ (the corresponding area in Fig. 5.5 is left blank). In this case, the association features of [81] cannot provide any useful information in order to estimate the correct association of DOAs to the sources.

Fig. 5.6 depicts the association accuracy, using Metric 1 and 2, for an anechoic scenario of three active sound sources and different values of SNR and C_3, i.e., the number of arrays that detected three sources. For each value of C_3 the figure presents the mean association accuracy over all possible combinations of C_2 and C_1. Again, the robustness of the proposed approach to missed detections is evident: our method achieves high accuracy for all SNR values even in the case where missed detections are so prominent that none of the arrays detected three sources, i.e., $C_3 = 0$.

5.5.3 Data association algorithm

We now demonstrate the effectiveness of our data association approach to more realistic scenarios with reverberation, where the DOAs of the sources in each time frame are estimated using the method of [16, 18]. The data-association accuracy for two and three active sound



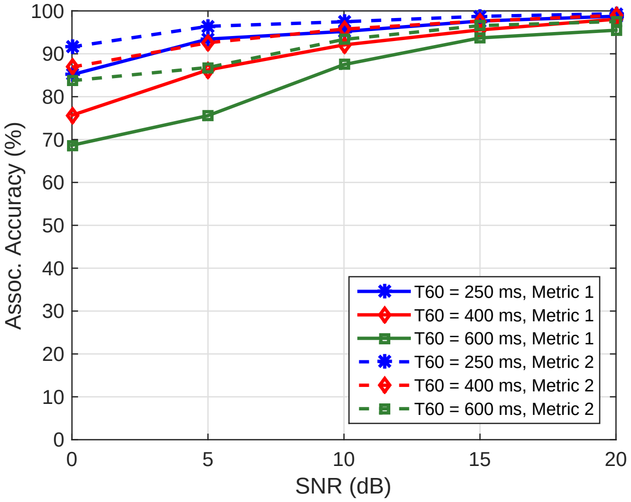
(a) Metric 1

(b) Metric 2

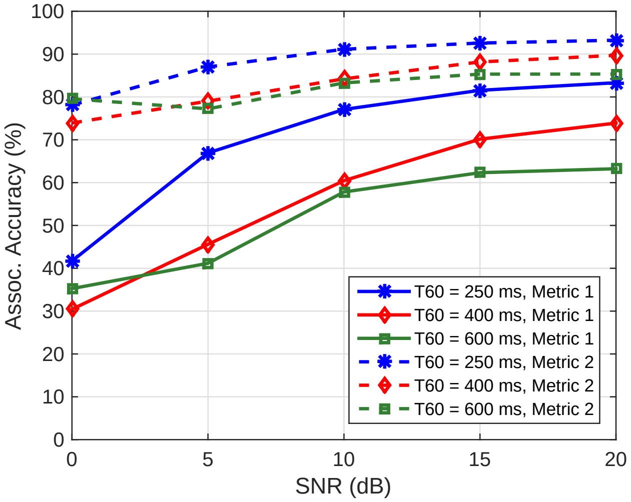
Figure 5.6: Data association accuracy of the proposed method for three sources in an anechoic environment for different values of SNR and C_3.

sources in scenarios with reverberation time $T_{60} = 250$, 400, and 600 ms is shown in Fig. 5.7. It can be observed that, while the association accuracy (Metric 1) decreases with increasing reverberation time, most of the DOAs between pairs of arrays (Metric 2) are still associated correctly both for the two and three sources case. This indicates that while association errors in frames occur more often, most DOA pairs are assigned to the correct source.

In contrast to the results in Section 5.5.2, the values of C_2 and C_3 now vary in each time frame, as the DOAs of the sound sources are now estimated. Thus, C_2 and C_3 depend on how many sources the DOA estimation method was able to detect at each array at each time frame. To quantify how often missed detections occur, we counted in how many frames each value of C_2 and C_3 occurs. We observed that for the two sources case, approximately in only 12% of the frames all four arrays detected two sources, in 21% of the frames $C_2 = 3$, in 33% of the frames $C_2 = 2$, in 19% of the frames $C_2 = 1$, and in 15% of the frames all arrays detected only one source, i.e., $C_2 = 0$. The problem of missed detections becomes even more evident in



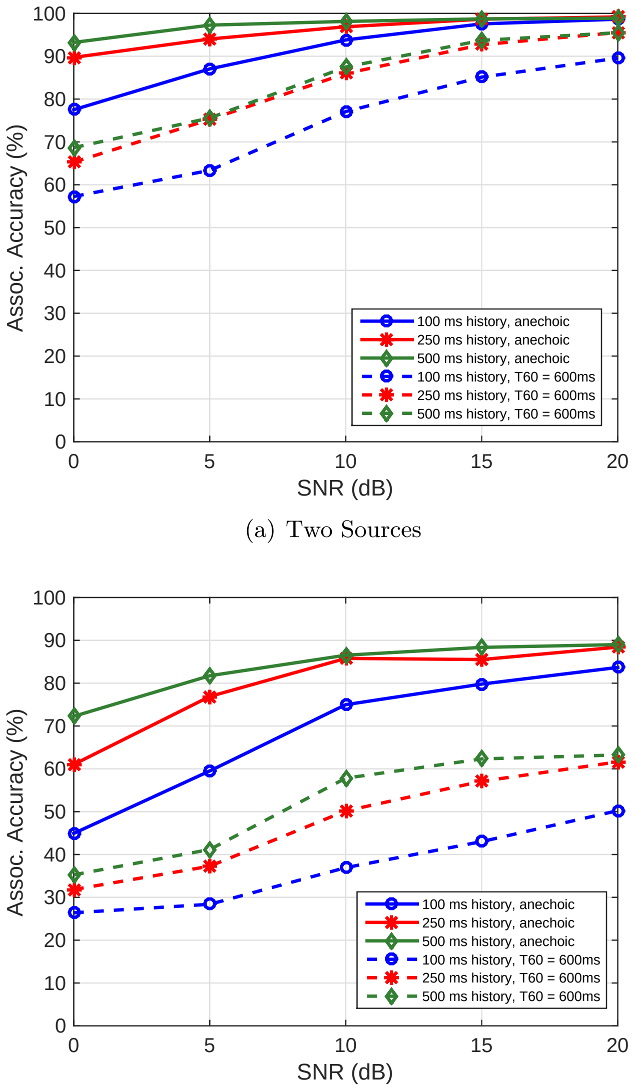
(a)



(b)

Figure 5.7: Data-association accuracy of the proposed method for different SNR values and reverberation conditions for (a) two and (b) three active sound sources.

the three sources case where in approximately 62% of the frames none of the arrays detected three sources ($C_3 = 0$), in 34% of the frames only one array detected three sources ($C_3 = 1$), and in only 5% of the frames the value of C_3 is greater than one. These numbers not only reveal once again the robustness of our method to missed detections, but also highlight the importance of the association method to take into account missed detections, as they occur very often in practice.



(a) Two Sources

(b) Three Sources

Figure 5.8: Data association accuracy of the proposed method, using Metric 1, for different values of the history length in anechoic and reverberant conditions.

Moreover, in Fig. 5.8 we demonstrate how the history length B can affect the association accuracy. We consider two and three active sound sources in anechoic conditions and in a scenario with reverberation time $T_{60} = 600$ ms, and we plot the association accuracy (using Metric 1), versus SNR for history lengths of 0.1, 0.25, and 0.5 seconds. In the two sources case, the performance is improved when increasing the history length from 0.1 seconds to 0.25 and 0.5 seconds, especially in the reverberant scenario. However, the performance when using a history length of 0.25 and 0.5 seconds is very similar. In general, there is an obvious performance improvement—especially for the three sources case—as the history length increases, both in anechoic and reverberant conditions, showing that increasing the history length makes the association features more robust to noise and reverberation. However, increasing the history length also increases the latency of the system, in turn decreasing responsiveness.

Finally, we evaluate the ability of our proposed association algorithm to find the optimal solution, according to the problem we defined in Section 5.3. As a greedy algorithm, our proposed data-association algorithm is not guaranteed to find the optimal solution to the problem defined by Eq. (5.5). As described in Section 5.3, the optimal solution can be guaranteed by exhaustively testing all possible assignments and choosing the one with the minimum score according to (5.5). We call this approach the brute-force version of the association algorithm. This version is impractical as the number of assignments under test is prohibitively large when the number of sources and nodes increases.

To demonstrate the ability of our greedy association algorithm to solve (5.5), we compare the solutions it provides with the optimal solution derived by the brute force version of the algorithm where all possible assignments are examined. Also, in order to quantify the performance gain of each of the two steps of our data-association algorithm, we include in the comparison another version of our algorithm where only the first greedy step is performed. The comparison was made on scenarios of two and three active sound sources, on all 30 source configurations, different SNR and reverberation conditions (SNR ranging from 0 dB to 20 dB with a step of 5 dB, and reverberation times of $T_{60} = 250, 400, 600$ ms). We observed that our greedy algorithm was able to find the optimal solution in 97.22% of the time frames for the two sources and in 84.31% of the time frames for the three sources cases. Moreover, even in the cases where the optimal solution cannot be found, the association accuracy results in Fig. 5.7 reveal that our greedy algorithm can still find good solutions. In contrast, the version of our algorithm that performs only the first greedy step was able to find the optimal solution in only 66.43% of the frames for the two sources case and in only 28.52% of the time frames for the three sources case. Thus, a significant performance gain is achieved due to the second greedy step of the proposed data-association algorithm.

To quantify the performance gain of our greedy association algorithm in terms of computation time, Table 5.2 compares the mean execution times of the greedy and brute force versions. The execution times were measured in MATLAB on a Windows desktop PC with a Core i7 CPU running at 3.4 GHz with 16 GB RAM. Note that while the absolute execution times may be highly dependent on the machine and the programming language, we are only interested here in the relative times between the two versions of the algorithm. It can be observed that, while the brute force version is computationally more efficient for the case of two active sound sources, it becomes impractical for the case of three active sound sources due to the high number of possible assignments that it needs to test. On the other hand, the

Table 5.2: Mean and standard deviation (in milliseconds) for the execution time of the greedy and brute force version of our data-association algorithm for two and three active sound sources

Two Sources

	Greedy	Brute Force
Mean	3.37 ms	1.93 ms
Std	0.69 ms	0.43 ms

Three Sources

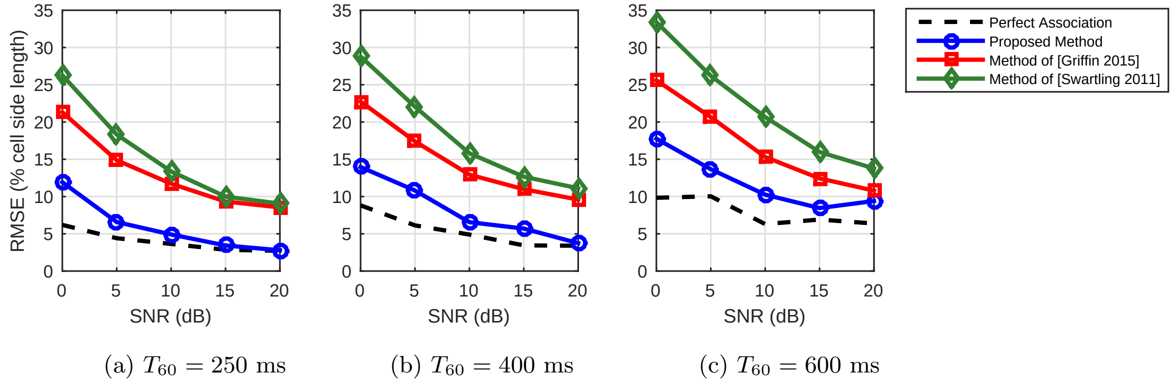
	Greedy	Brute Force
Mean	7.58 ms	181.90 ms
Std	2.38 ms	28.97 ms

execution time of the greedy version in the three sources case remains in the same order of magnitude compared to the two sources one.

5.5.4 Location estimation accuracy

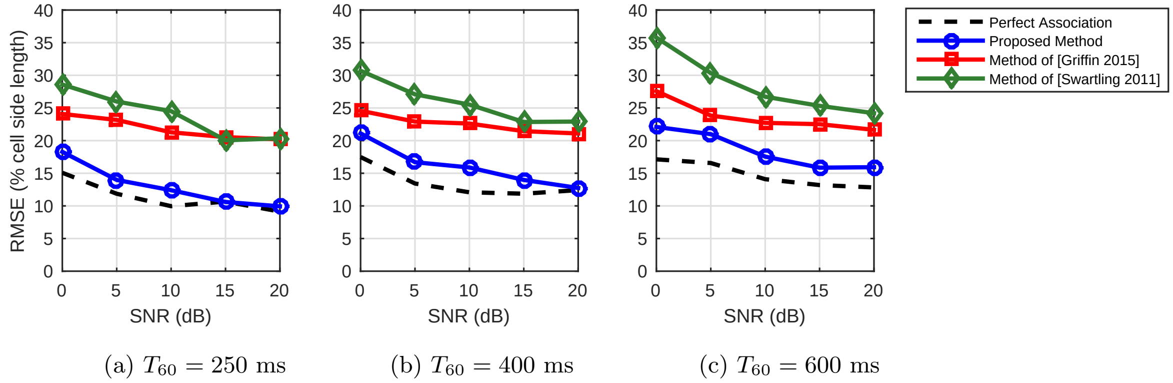
When the correct association of DOAs to the sources is found, the locations of the sound sources can be estimated by applying a single source location estimator to the corresponding DOA associations. In this section, we evaluate the localization accuracy of our proposed approach and compare it with the use of the association features extracted from [81] (denoted as [Swartling [2011]). For comparison, we also include the localization performance of our multiple source grid-based estimator, which we proposed in [85] and presented in Chapter 4 (denoted as [Griffin 2015]). We remind the reader that this estimator infers a location for every possible DOA combination from the arrays and on a second step decides which locations correspond the true sources' locations, using no additional information apart from the DOA estimates.

In order to localize the sound sources using the proposed method and the method of [81], we apply our previously proposed single-source grid-based location estimator [85], presented also in Chapter 4, on the estimated DOA associations. To measure the localization performance we use the root-mean square error (RMSE) over all sources, all 30 different source configurations and over all frames where each source was detected by at least two arrays, which is a necessary condition to infer a location estimate for all sources. These frames represent the 90% and 63% of all frames under test for the two and three sources case respectively. As the use of association features of [81] cannot provide a DOA association in some cases (for example when $C_2 = 0$ for the two sources case and $C_3 = 0$ for the three sources case) we



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Figure 5.9: Localization error as a percentage of cell side length $V = 4$ meters for two active sound sources and different reverberation scenarios.



(a) $T_{60} = 250$ ms (b) $T_{60} = 400$ ms (c) $T_{60} = 600$ ms

Figure 5.10: Localization error as a percentage of cell side length $V = 4$ meters for three active sound sources and different reverberation scenarios.

consider for this method only the frames where a DOA association can be estimated. These frames represent the 83% and 36% of the total frames under test for the two and three sources case respectively. These numbers highlight again the advantage of our proposed method in terms of its ability to find a DOA association even in scenarios with severe missed detections.

Figures 5.9 and 5.10 depict the location error for various reverberation conditions and for scenarios with two and three simultaneously active sound sources. The location error when using the estimated DOAs but assuming that the correct association of DOAs to the sources is known (denoted as Perfect Association) is also included to represent the best-case scenario. As expected the localization performance degrades with increasing reverberation time. The performance of the best-case scenario with perfect associations also degrades as the DOA estimates suffer from larger noise due to the high reverberation conditions. In general, it can be observed that the proposed method always achieves the best localization performance, providing location estimates very close to the best-case especially for the higher SNR values and for both two and three active sound sources. The other two methods always perform worse than the proposed one, and their performance degradation is more evident in the three sources case.

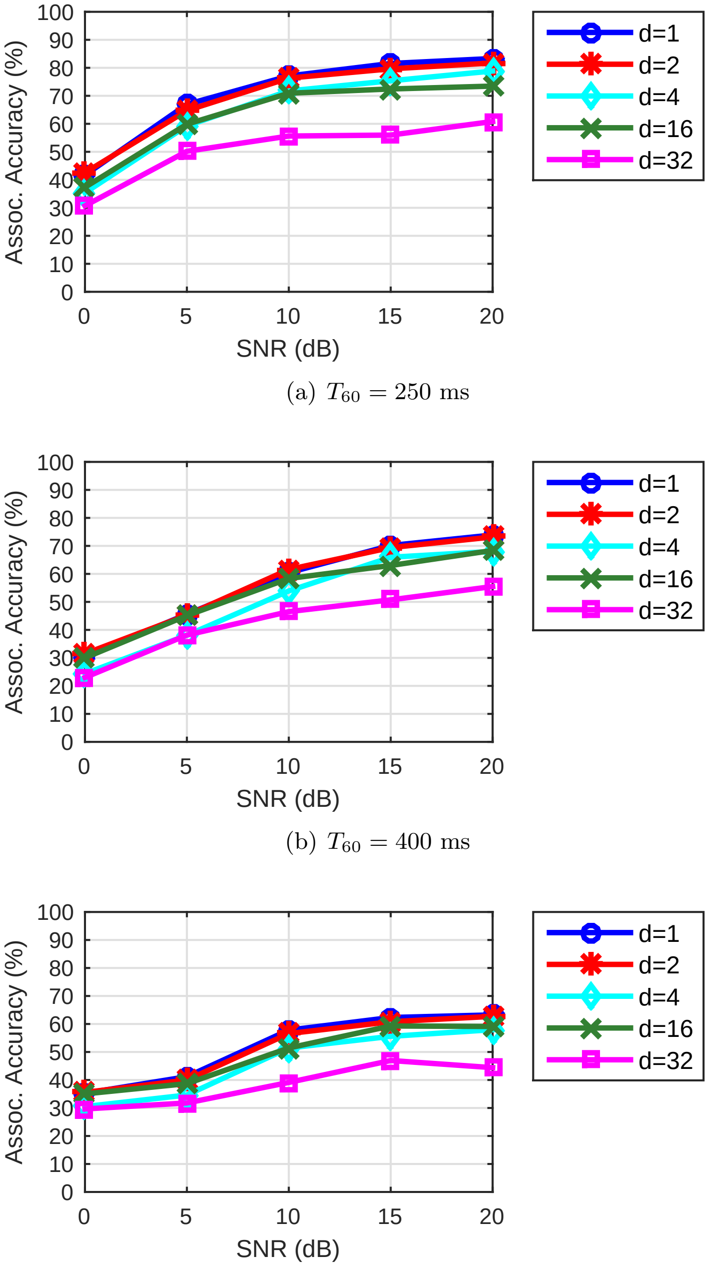
5.5.5 Reduction in transmission requirements

In this section, we evaluate the performance of our proposed decimation process applied to the association features in order to reduce the amount of information that needs to be transmitted by the nodes. We examine the effect of the decimation factor on the data-association and localization accuracy and investigate how much we can reduce the transmitted information without affecting performance.

Fig. 5.11 depicts the association accuracy using Metric 1 for a scenario of three active sound sources for different decimation factors, namely $d = 1$ (i.e., no decimation), $d = 2$, $d = 4$, $d = 16$, and $d = 32$ and for different reverberation conditions. The corresponding localization error, when the single-source grid-based method [85] is applied on the estimated DOA associations is shown in Fig. 5.12. It can be observed that the performance for all decimation factors up to $d = 16$ is very similar to the case where no decimation is applied (i.e., $d = 1$). The association and localization performance exhibits higher degradation when a decimation factor of $d = 32$ is used.

To quantify the gain in terms of reduction in information that must be transmitted, we can use Eq. (5.7) to calculate how many bits are required to transmit an association feature (i.e., histogram) for every value of d: 813, 499, 296, and 97 bits for $d = 1$ (i.e., no decimation), $d = 2$, $d = 4$, and $d = 16$, respectively. These numbers are obtained by substituting into Eq. (5.7) the block size $B = 21$, and the number of bins $N_\ell = 185$, which results from an FFT size of 2048 samples, a sampling frequency of 44.1 kHz and a maximum frequency for processing $\ell_{\max} = 4$ kHz (see Table 5.1). The results suggest that, by applying a decimation process by a factor of $d = 16$ to the estimated association features, we can reduce the amount of information that is required to transmit a histogram by almost 88%, with minor losses in the association and location estimation performance.

Finally, for the tested scenario of three simultaneously active sound sources, Table 5.3 depicts the worst-case and average bitrate requirements for a node. The worst-case corresponds to the case where all arrays detected all sources, thus requiring to transmit the maximum possible number of association features, which in the case of three sources is three. However, in a realistic situation missed detections will occur and thus the nodes will rarely need to transmit three association features. This corresponds to the average case in Table 5.3 which depicts the average bitrate over all 30 different tested source configurations, all SNRs, and reverberation times.



(a) $T_{60} = 250$ ms

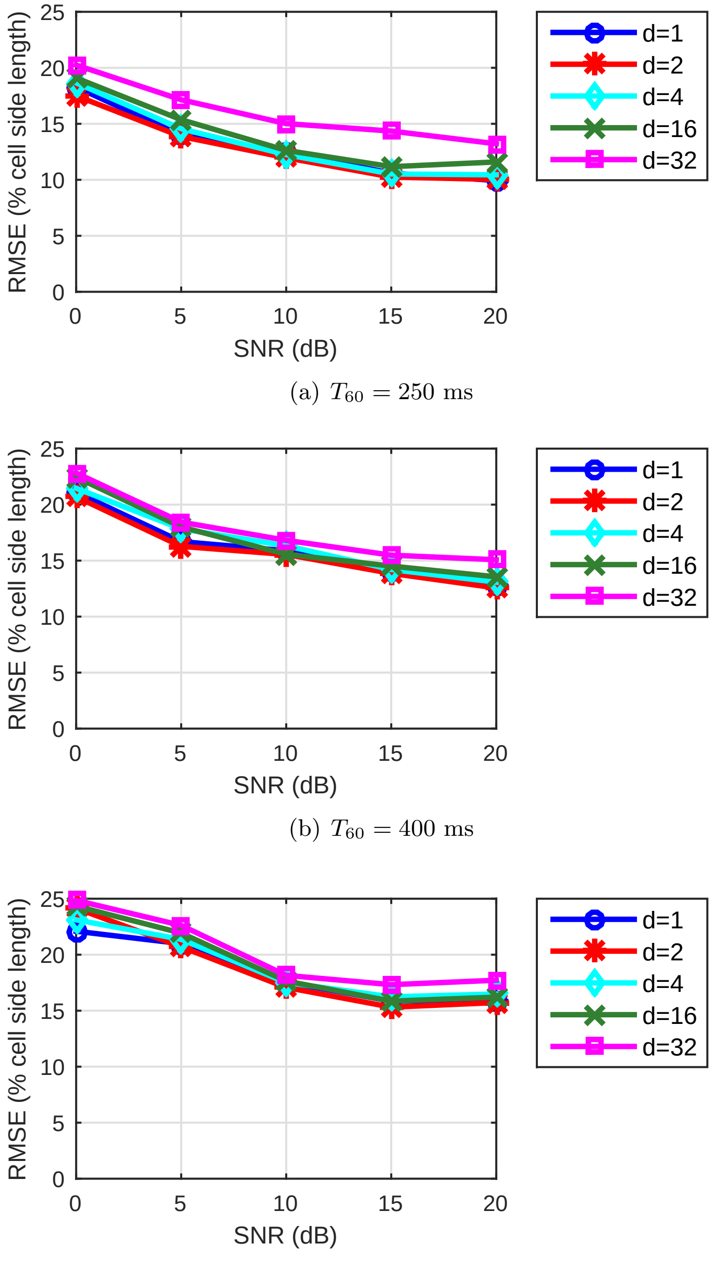
(b) $T_{60} = 400$ ms

(c) $T_{60} = 600$ ms

Figure 5.11: Association accuracy using Metric 1 for three active sound sources, different reverberation conditions, and different values of the decimation factor in the association features.

5.5.6 Moving sources

Finally, we demonstrate our method's ability to perform data-association and accurate location estimation in scenarios with moving sources. Fig. 5.13 depicts the location estimates for



(a) $T_{60} = 250$ ms

(b) $T_{60} = 400$ ms

(c) $T_{60} = 600$ ms

Figure 5.12: Localization error as a percentage of cell side length $V = 4$ meters for three active sound sources, different reverberation conditions, and different values of the decimation factor in the association features.

all time frames, when the single-source grid-based method [85] is applied to the estimated DOA associations for a simulated scenario with one moving and one static source (*Moving1*)

Table 5.3: Worst-case and average transmission requirements, (in Kbits per second) for a node for three active sources.

	Worst-case bitrate	Average bitrate
no decimation	105 Kbps	63 Kbps
$d = 2$	64 Kbps	39 Kbps
$d = 4$	38 Kbps	23 Kbps
$d = 16$	12 Kbps	8 Kbps

Table 5.4: Localization error as a percentage of the cell side length $V = 4$ meters and data-association accuracy using Metric 1, for moving sources for different values of the history length.

	Moving1		*Moving2*	
	Metric 1	RMSE	Metric 1	RMSE
100 ms history	84%	7.02%	82%	9.80%
250 ms history	88%	5.17%	86%	8.25%
500 ms history	87%	5.16%	84%	7.94%

at $T_{60} = 400$ ms reverberation time. The sources were 5 seconds in duration. The WASN setup and simulation parameters were the same as shown in Table 5.1. A decimation process by a factor of $d = 16$ was applied to the association features to reduce bitrate needs. The static source was located at $(5, 4)$ meters and the moving one starts from point $(6.5, 6)$ meters and moves on a straight line to point $(3.5, 6)$ meters. Finally, Fig. 5.14 depicts the location estimates for a scenario of two moving sound sources (*Moving2*) at $T_{60} = 400$ ms reverberation time. The first source starts again from point $(6.5, 6)$ meters and moves on a straight line to point $(3.5, 6)$ meters, while the second one starts from point $(3.5, 4)$ meters and moves on a straight line to point $(6.5, 4)$ meters. It can be observed that in both cases, the method produces accurate and smooth location estimates, indicating its ability to localize moving sources.

To evaluate the effect of the history length in the case of moving sources, Table 5.4 presents the location error and the DOA association accuracy, using Metric 1, for the two aforementioned scenarios (*Moving1* and *Moving2*) for different values of the history length. The results—which are in accordance with the ones in Fig. 5.8 for the two sources case—show that the 0.1 second history exhibits the worst performance, while the performance for the 0.25 and 0.5 seconds history is very similar, with the 0.25 seconds history exhibiting slightly higher DOA association accuracy but also slightly higher location error.

Finally, we evaluate the effect of the moving source's velocity to the data-association and

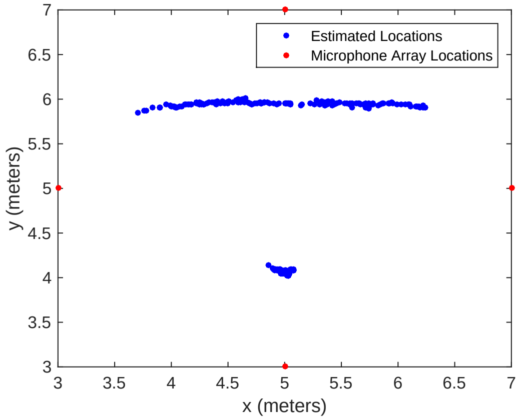


Figure 5.13: Location estimates of the proposed method for all time frames for a scenario with one moving and one static sound source at $T_{60} = 400$ ms reverberation time.

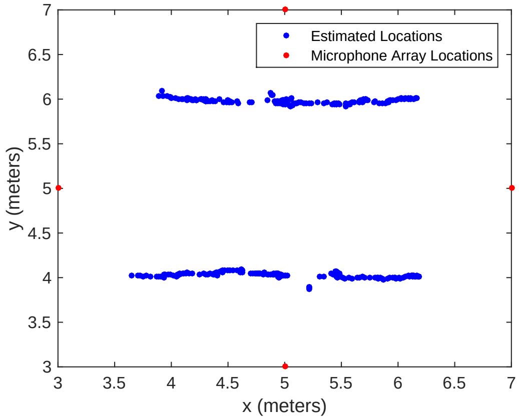
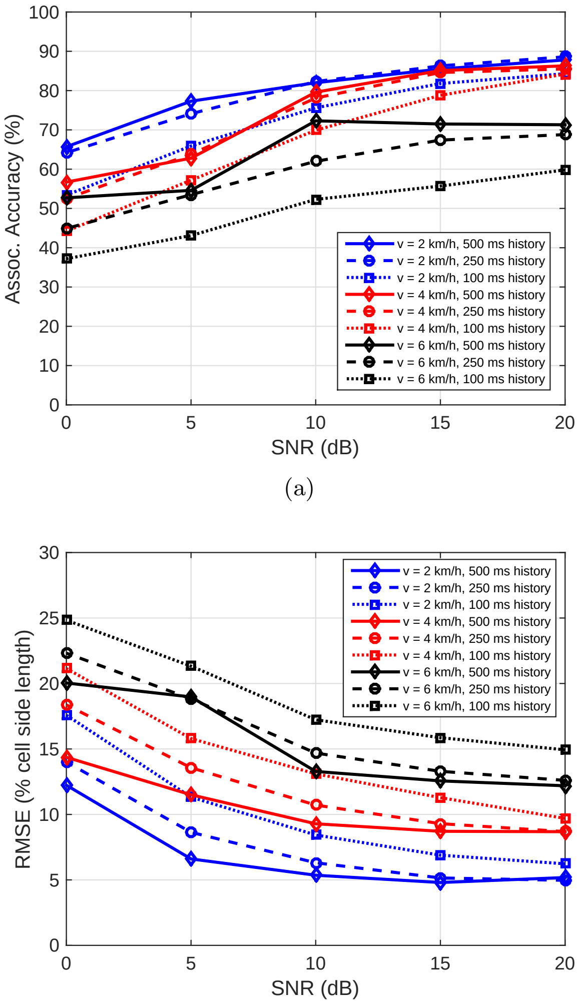


Figure 5.14: Location estimates of the proposed method for all time frames for a scenario with two moving sound sources at $T_{60} = 400$ ms reverberation time.

localization accuracy. We consider a scenario of one static source at $(5, 4)$ meters and one moving source, at $T_{60} = 400$ ms reverberation time. The moving source moves on a straight line from point $(6.5, 6)$ to point $(3.5, 6)$ and then back to point $(6.5, 6)$. We simulated different velocities for the moving source, from slow to fast walking speeds, namely $v = 2$ km/h, $v = 4$ km/h, and $v = 6$ km/h and applied the proposed methodology to estimate the association of DOAs from the arrays to the sources and the final sources' locations. The WASN setup and simulation parameters were the same as shown in Table 5.1 and a decimation process by a factor of $d = 16$ was applied. The experiment was repeated 10 times. Fig. 5.15 shows the data-association accuracy using Metric 1 and the location estimation error for various SNR levels and various values for the history length. As expected, the performance degrades when the velocity of the source increases. Again, a history length of 0.5 seconds exhibits the best performance, while a history length of 0.1 seconds exhibits the worst performance. As in the previous experiments with two active sound sources, the performance when using a history length of 0.25 seconds is very similar to the one when using a history length of 0.5 seconds. It is generally evident that the speed of a moving source does not affect the choice of the history length. This can be explained by the fact that during the design of the association features, the narrowband DOA estimates in the B previous frames are compared to the broadband DOA estimates of the sources in the current frame. As a result, if the DOA of the source has significantly changed in the last B frames, the distance between the narrowband DOA and the broadband DOAs of the sources will be large and the corresponding frequencies will not be taken into account due to Eq. (5.2). Finally, it can be observed that the method provides satisfactory performance even in the case where the moving source moves as fast as 6 km/h, validating again its ability to localize moving sources.

5.6 Conclusions

In this chapter we considered the data-association problem that occurs when localizing multiple sources from their DOA estimates. We presented a solution that uses additional information transmitted by the sensors, apart from the DOA estimates. Each node estimates an association feature for each source it can detect. The association features describe how the frequency components of the captured signals are distributed to the sources. The correct association of DOAs from the nodes to the sources is found based on the similarity of the association features. To keep the amount of information that needs to be transmitted in the network at low levels our method also incorporates a scheme for reducing the bitrate needs up to 88% without affecting accuracy. Through simulations we validated the efficiency of our method to accurate solve the data-association problem and localize multiple sources in scenarios with missed detections, reverberation, noise, and moving sources. In the next chapter we will relax the assumption of known number of sources. We will assume that the number of sources is also unknown and we will describe our proposed method for source counting and location estimation.



(a)

(b)

Figure 5.15: (a) Data-association accuracy using Metric 1 and (b) localization error as a percentage of the cell side length $V = 4$ meters for a scenario of one static and one moving source with different velocities v and different values for the history length.

Chapter 6
Source localization and counting

6.1 Introduction

So far, we considered the multiple source localization problem in scenarios where the number of sources is known both at the sensors and at the fusion center. We developed approaches that operate in two steps: first, the direction of arrival for each detected source is estimated in the sensor node at each time instant. In the second step, the DOA estimates are transmitted to the fusion center to infer the sources' locations. This step requires addressing the data-association problem and applying DOA-based location estimators, as discussed in the previous chapters.

When the number of sources is unknown, the methods considered so far can perform source counting based solely on the number of sources (i.e., number of DOA estimates) detected by the sensors. However, when the number of true sources is not available at sensor level, each sensor may overestimate or underestimate the sources number. Thus, the number of sources in each sensor can vary and it would be very challenging for the fusion center to determine the correct number of sources based on the individual estimates received by the sensors.

In this chapter, we consider the joint problem of sound source location estimation and counting in scenarios where the number of sources is unknown both to the individual sensors and the fusion center. We develop a hybrid method to estimate both the number of sources and their locations. In our method, which was published in [93], the sensors estimate and transmit a direction of arrival for every frequency component of the captured signals. The fusion center then estimates a location for each frequency, using the DOA estimates from the arrays at that frequency. To estimate the final sources' locations and their number, we use these per-frequency location estimates and their corresponding two-dimensional histogram that describes the number of times each location is estimated and is thus a probability indicator about the sources' final locations. We assume that the per-frequency location estimates have been generated by a Gaussian mixture where the number of Gaussian components is also an unknown parameter. Through statistical modeling using the Bayesian K-means clustering algorithm [94] we estimate the number of Gaussian components as well as their means. The sources' number is given by the estimated number of Gaussians, while the sources' locations are given by their means. We also tested the applicability of our proposed method in a real WASN. To do that, we implemented our method in C++ and we tested it on a real two-node WASN where the nodes consist of our digital MEMS microphone arrays, constructed at FORTH-ICS. We validated that our method can run in real-time and produce satisfactory results.

The remainder of this chapter is organized as follows: Section 6.2 describes the processing which is carried out at each individual sensors for the estimation of the narrowband DOA estimates. Section 6.3 presents the proposed approach for location estimation and counting based on the narrowband DOA estimates transmitted by the sensors and Section 6.4 presents the performance evaluation of the method. Section 6.5 discusses the practical challenges when applying our method in real-life scenarios. Such challenges include the positioning of the sensors and the directivity patterns of real speakers. We incorporate our method with a DOA estimation method that takes into account reflections occurred when the sensors are placed near walls and we evaluate our method in a dataset of real recordings with real speakers that we collected. Finally, Section 6.6 discusses the real-time implementation of our method in a real two-node WASN.

6.2 In-node processing

The signals received at the ith microphone of each (say the mth) array are first transformed into the Short-Time Fourier Transform domain, resulting in the signals $X_{m,i}(\tau, \ell)$ where τ and ℓ denote the time frame and frequency bin index, respectively. We denote as (τ, L) the set of frequencies ℓ for frame τ up to a maximum frequency $\ell_{\max}$. In the remainder, we omit τ, as the procedure is repeated in each frame. Each array m estimates a DOA in each frequency $\ell \in L$ resulting in the DOA estimates $\phi_m(L)$. For DOA estimation in each frequency we use the method of [4]. However, note that our proposed method for location estimation and counting is not restricted to a specific DOA estimation method or array geometry. The per frequency DOA estimates in $\phi_m(L)$ are then transmitted to the fusion center, which performs the localization and counting.

6.3 Processing at the fusion center

The fusion center estimates the number of active sources and their corresponding locations based on the DOA estimates in $\phi_m(L)$.

6.3.1 Per-frequency location estimation

First, a location is estimated for each frequency, based on the transmitted DOA estimates from the arrays at that frequency. For location estimation, we use the single-source version of the grid-based method presented in Chapter 4, which is a computationally efficient non-linear least squares estimator with high accuracy [85, 86]. We create a block of location estimates that contains the estimates of the current frame and B previous frames—also referred to as history length. Assuming that the signals are sufficiently sparse so that at most one source is dominant at each time-frequency bin [95], we expect that the location estimates will form K clusters around the true sources' locations.

6.3.2 Outlier rejection

To remove erroneous estimates occurred due to noise and/or reverberation, we construct a two-dimensional histogram from the set of location estimates obtained from the previous step. We smooth the histogram by applying an averaging filter with a rectangular window $w(\cdot, \cdot)$ of length h_X and h_Y in the $x-$ and $y-$dimension, respectively. Erroneous estimates are expected to be of low cardinality in the smoothed histogram. Thus, we remove any location estimates whose cardinality is less than q times the maximum cardinality of the histogram, where $q \in [0, 1]$ is a pre-defined constant. The effect of the outlier rejection is shown in Figure 6.1 for a case of two active sources at 20 dB signal-to-noise ratio.

6.3.3 Clustering

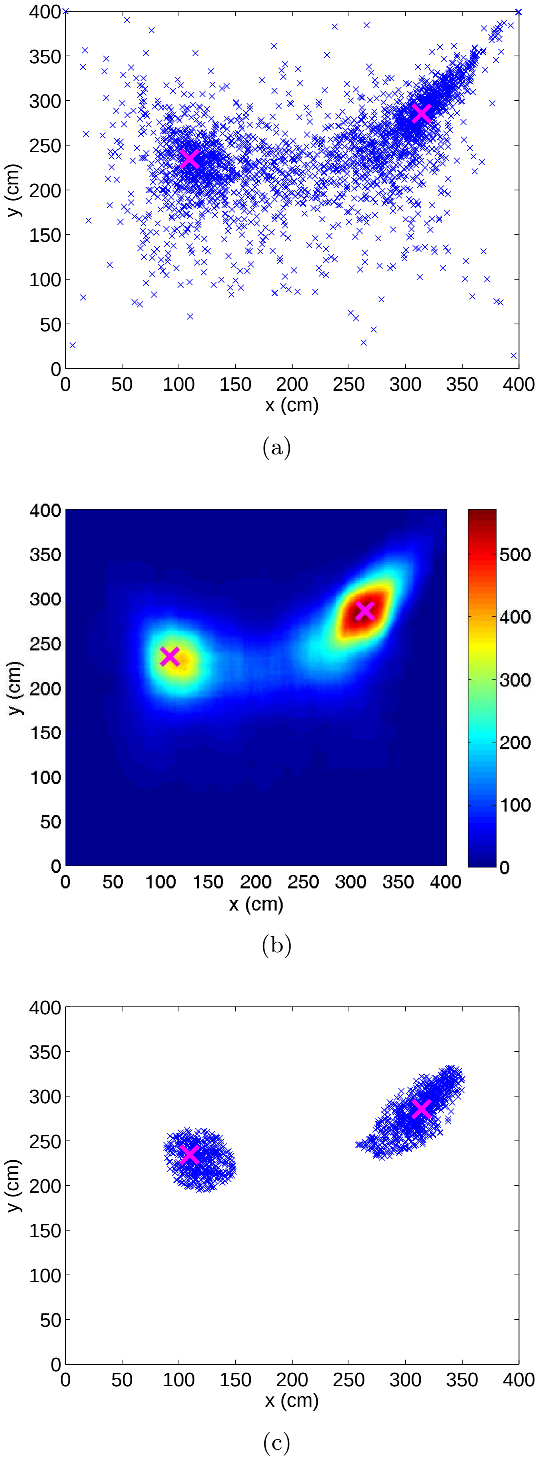
The location estimates that remained are used for localization and counting. To do this, we employ the Bayesian K-means clustering algorithm proposed in [94]. The algorithm estimates the number of clusters and their centroids, which in our case correspond to the number of active sound sources and their locations, respectively. In the following, we briefly describe the Bayesian K-means approach.

Let $\boldsymbol{p}_n$, $n = 1, \dots, N$ be the location estimates in the $D = 2$ dimensions, after the outlier rejection step. The algorithm assumes that the data have been generated by a mixture of C Gaussians:

$$p(\boldsymbol{p}_n | \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{c=1}^{C} \alpha_c \mathcal{N}(\boldsymbol{p}_n | \boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c) \tag{6.1}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_C\}$ are the mixing coefficients, $\mathcal{N}(\cdot)$ is the normal distribution, and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_1, \cdots, \boldsymbol{\Lambda}_C\}$ are the set of means and precision (inverse covariance) matrices of the Gaussians. We also define a C-dimensional binary cluster assignment variable $\boldsymbol{z}_n$ so that $z_{nc} = 1$ if the nth location estimate is assigned to cluster c (i.e., generated from the c-th Gaussian component) and $z_{nj} = 0$ for $j \neq c$.

In a Bayesian treatment of the mixture model, we place conjugate priors on the unknown



(a)

(b)

(c)

Figure 6.1: The effect of the outlier rejection process. (a) The per-frequency location estimates obtained using a history length of 1 sec. (b) The corresponding smoothed histogram obtained by filtering with a rectangular window of length $h_X = h_Y = 50$ cm in the $x-$ and $y-$ dimension. (c) The location estimates that remained after the outlier rejection process with $q = 0.35$. The X's represent the sources' true locations.

parameters:

$$p(\boldsymbol{\alpha}) = \mathcal{D}(\boldsymbol{\alpha}, \phi_0), \quad p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}_{\boldsymbol{0}}, \xi_0\boldsymbol{\Lambda})\ \mathcal{W}(\boldsymbol{\Lambda}|\eta_0, \boldsymbol{B}_0) \tag{6.2}$$

where $\mathcal{D}(\cdot)$ is the Dirichlet distribution and $\mathcal{W}(\cdot)$ is the Wishart distribution. The priors depend on the non-random hyper-parameters $\{\boldsymbol{m}_0, \boldsymbol{B}_0, \phi_0, \xi_0, \eta_0\}$.

The Bayesian K-means objective is to minimize the following function, jointly over assignment variables $\boldsymbol{z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$ and the number of clusters C:

$$\begin{aligned}\mathcal{F}(\boldsymbol{z}, C) = \sum_{c=1}^{C} \Bigg[& \frac{DN_c}{2}\text{log}\pi + \frac{1}{2}\text{log}\frac{\xi_c}{\xi_0} + \frac{\eta_c}{2}\text{log}|\boldsymbol{B}_c| - \text{log}\frac{\Gamma(\phi_c)}{\Gamma(\phi_0)} \\ & -\frac{\eta_0}{2}\text{log}|\boldsymbol{B}_0| - \text{log}\frac{\Gamma_D\left(\frac{\eta_c}{2}\right)}{\Gamma_D\left(\frac{\eta_0}{2}\right)} + \frac{1}{C}\text{log}\frac{\Gamma(N + C\phi_0)}{\Gamma(C\phi_0)}\Bigg]\end{aligned} \tag{6.3}$$

where the dependence on $\boldsymbol{z}$ is through the cluster-dependent quantities $\boldsymbol{B}_c, \xi_c, \eta_c, \phi_c$ that are described in the following, N_c denotes the number of locations that belong to cluster c, $\Gamma_D(x) = \pi^{\frac{D(D-1)}{4}} \prod_{i=1}^{D} \Gamma\left(x + \frac{1-i}{2}\right)$, $\Gamma(\cdot)$ is the Gamma function, and $|\cdot|$ denotes the determinant of a matrix.

Update rules: Given C clusters, the algorithm performs the cluster assignment by iteratively minimizing the cost function:

$$\mathcal{C} = \sum_{n=1}^{N}\sum_{c=1}^{C} z_{nc}\gamma_c(\boldsymbol{p}_n) \tag{6.4}$$

with

$$\begin{aligned}\gamma_c(\boldsymbol{p}_n) \quad = \quad & \frac{\eta_c}{2}(\boldsymbol{p}_n - \boldsymbol{m_c})^T\boldsymbol{B}_c^{-1}(\boldsymbol{p}_n - \boldsymbol{m}_c) + \frac{1}{2}\text{log}|\boldsymbol{B}_c| \\ & + \frac{D}{2\xi_c} - \frac{1}{2}\sum_{d=1}^{D}\Psi\left(\frac{\eta_c + 1 - d}{2}\right) - \Psi(\phi_c)\end{aligned} \tag{6.5}$$

where $\Psi(\cdot)$ is the “digamma” function and the cluster quantities $\{\boldsymbol{B}_c, \boldsymbol{m}_c, \xi_c, \eta_c, \phi_c\}$ are calculated as:

$$\begin{array}{lll} \boldsymbol{m}_c = \dfrac{N_c\bar{\boldsymbol{p}}_c + \xi_0\boldsymbol{m}_0}{\xi_c} & \eta_c = \eta_0 + N_c & \xi_c = \xi_0 + N_c \\ \boldsymbol{B}_c = \boldsymbol{B}_0 + N_c\boldsymbol{S}_c + \frac{N_c\xi_0}{\xi_c}(\bar{\boldsymbol{p}}_c - \boldsymbol{m}_0)(\bar{\boldsymbol{p}}_c - \boldsymbol{m}_0)^T & \phi_c = \phi_0 + N_c & \end{array} \tag{6.6}$$

where $\bar{\boldsymbol{p}}_c$ and $\boldsymbol{S}_c$ are the sample mean and sample covariance of cluster c, respectively. The algorithm—being similar to K-means—alternates between calculating the parameters $\{\boldsymbol{B}_c, \boldsymbol{m}_c, \xi_c, \eta_c, \phi_c\}$ and updating the cluster assignment according to:

$$z_{nc} = \begin{cases} 1 & \text{if } c = \arg\min_j \gamma_j(\boldsymbol{p}_n) \\ 0 & \text{otherwise} \end{cases} \tag{6.7}$$

Algorithm 3 Bayesian K-Means Clustering [94]

```
Input: Location estimates $\boldsymbol{p}_n, n = 1, \ldots, N$
Output: Number of clusters $C$, cluster centroids $\boldsymbol{m}_c, c = 1, \ldots, C$
  1. Initialization: Set $C = 1$, perform clustering using Eq. (6.4)-(6.7) until convergence,
  and evaluate Eq. (6.3) for this clustering assignment
  2. Split operations: Calculate the split score for each cluster and sort them in descending
  order of scores.
      (i) Split the cluster with the highest score into two with centroids $\boldsymbol{m} \pm \boldsymbol{d}$ where $\boldsymbol{m}$ is the
          centroid of the cluster to be split and $\boldsymbol{d} = \boldsymbol{s}\sqrt{\lambda}$ with $\boldsymbol{s}$ being the principal eigenvector
          of the sample covariance matrix $\boldsymbol{S}$ of the cluster to be split and $\lambda$ its corresponding
          eigenvalue.
     (ii) Perform clustering using Eq. (6.4)-(6.7) until convergence.
    (iii) Evaluate Eq. (6.3) for the new clustering assignment.
          a. If Eq. (6.8) holds, accept split and repeat STEP 2.
          b. Otherwise, reject split and go to STEP 2(i) to split the cluster with the next
             highest score.
  3. Merge operations: Calculate the merge score for each pair of clusters and sort them in
  descending order of scores.
      (i) Merge the data points of the pair of clusters with the highest score into one cluster.
     (ii) Perform clustering using Eq. (6.4)-(6.7) until convergence.
    (iii) Evaluate Eq. (6.3) for the new clustering assignment.
          a. If Eq. (6.8) holds, accept merge and go to STEP 2.
          b. Otherwise, reject merge and go to STEP 3(i) to merge the pair of clusters with
             the next highest score.
  4. Terminate when no more split/merge operations satisfy Eq. (6.8).
```

The clustering assignment updates converge when the cost in Eq. (6.4) is kept constant between iterations.

Split and merge procedures: To search over different number of clusters, Bayesian K-means introduces split and merge operations which are based on the work of [96]. For each cluster a split score is calculated, based on the Kullback-Leibler divergence between the local empirical probability density around that cluster and the Gaussian mixture model. Moreover, for each pair of clusters i and j a merge score is calculated based on the cosine distance of the $N-$dimensional vectors that contain the posterior probabilities of the data points for the ith and jth Gaussian. The reader is referred to [94, 96] for more details on the merge and split scores.

The complete algorithm is described in Algorithm 3. After each split and merge procedure the objective function in (6.3) is evaluated so as to accept or reject the split/merge operation. To avoid overestimation on the number of sources caused when the algorithm tries to overfit the data with a complex model with many clusters, we accept a split/merge operation only when the difference in the objective function is greater than a predefined threshold. More formally if we denote as $\mathcal{F}^b(\boldsymbol{z}, C)$ and $\mathcal{F}^a(\boldsymbol{z}, C)$ the value of (6.3) before and after a split/merge operation, we accept the operation iff

$$\frac{\mathcal{F}^b(\boldsymbol{z}, C) - \mathcal{F}^a(\boldsymbol{z}, C)}{\mathcal{F}^b(\boldsymbol{z}, C)} > t_{\text{accept}} \tag{6.8}$$

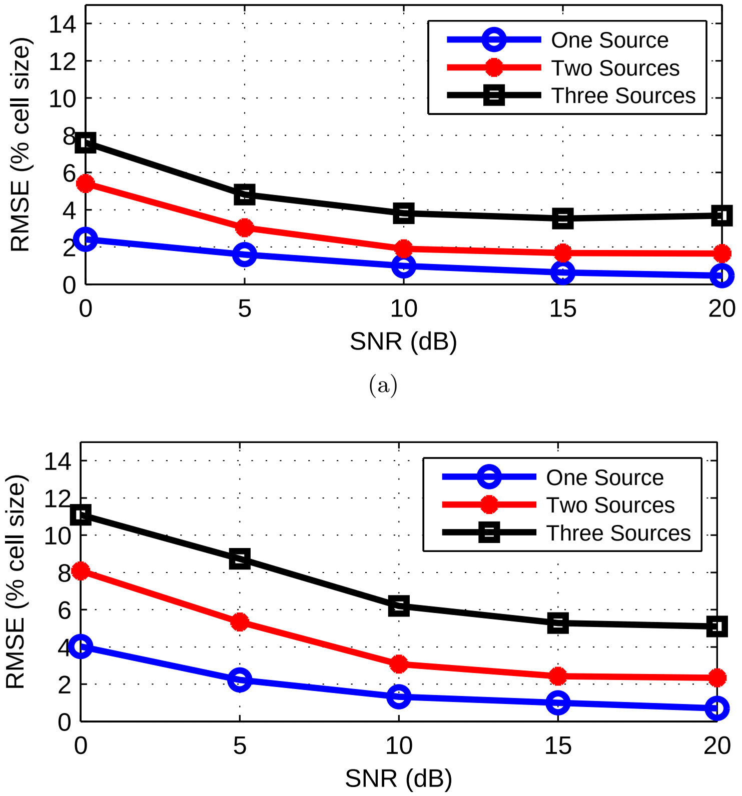
The algorithm terminates when no more split/merge operations satisfy (6.8), and outputs the number of clusters and their centroids, which denote the number of active sources and their locations.

6.4 Evaluation

6.4.1 Simulation Results

We performed simulations on a square cell of a WASN with dimensions of $V = 4$ m with four microphone arrays placed on the corners of the cell. Each array is a uniform circular array with $N = 8$ omnidirectional microphones and a radius $r = 0.05$ m. In each simulation, the sound sources were speech recordings of 3 seconds sampled at 44.1 kHz and had equal power when located at the center of the cell. The signal-to-noise ratio (SNR) was measured as the ratio of the power of each source signal when located at the center of the cell to the power of the noise signal. To simulate different SNR values we added white Gaussian noise at each microphone, uncorrelated with the source signals and the noise at the other microphones. Note that this framework results in different SNR at each array depending on how close the source is to the arrays.

We used the Image-Source method [91] to simulate a room of dimensions $10 \times 10 \times 3$ meters and produce signals of omnidirectional sources at various reverberation times. The WASN cell was placed in the middle of the room with the arrays and the sources being at 1.5 m height. We considered scenarios of up to three simultaneous sources. Each simulation was repeated 30 times and the sources were placed within the cell with independent uniform probability at a distance of at least 1 m away from each other and at least 0.5 m away from the arrays. For processing, we used frames of 2048 samples with 50% overlap, windowed with a Hamming window. The FFT size was 2048 and $\ell_{\max} = 4$ kHz which is the spatial-aliasing frequency for the given array geometry. For the block processing we used a history length of 1 second which corresponds to $B = 43$ frames and the rectangular window for the smoothing of the histogram was of length $h_X = h_Y = 50$ cm. For outlier rejection we set $q = 0.35$, thus removing location estimates whose cardinality in the smoothed histogram is less than 35% the maximum cardinality. The histogram bin size was 1 cm^2. For the Bayesian K-means, the hyper-parameters were set to $\phi_0 = 2$, $\xi_0 = 0.1$, $\eta_0 = D = 2$, $\boldsymbol{m}_0 = \bar{\boldsymbol{p}}$, and $\boldsymbol{B}_0 = 2d_0^2\boldsymbol{S}/\text{trace}(\boldsymbol{S})$, where $\text{trace}(\cdot)$ denotes the trace of a matrix, $\bar{\boldsymbol{p}}$ and $\boldsymbol{S}$ are the sample mean and sample covariance matrix of the data, and d_0 is determined by computing the



(a)

(b)

Figure 6.2: Location error as a percentage of the cell size $V = 4$ m for various number of active sound sources and reverberation time (a) $T_{60} = 250$ ms and (b) $T_{60} = 400$ ms

closest distance for 10% of the location estimates and averaging between the 3 closest pairs. Finally, we set t_{accept} to 0.01.

Table 6.1 depicts the source counting success rates as the percentage of frames for all 30 different source configurations—excluding the first $B - 1$ frames for the history block initialization—where the correct number of sources was found for one, two, and three simultaneously active sources for different SNR levels and reverberation time of $T_{60} = 250$ ms and $T_{60} = 400$ ms. The method almost always identifies the correct number of sources in the single source scenario. It also yields accurate source counting performance for two and three sources, especially at the higher SNR cases for both reverberation conditions. Figure 6.2 shows the corresponding root-mean square error (RMSE) as a percentage of the cell size V, over all sources, all 30 different source configurations for all frames where the correct number of sources was detected. It is evident that the proposed method achieves quite accurate localization for all cases especially at higher SNR levels.

Table 6.1: Source counting success rates for various active sources in different SNR and reverberation conditions.

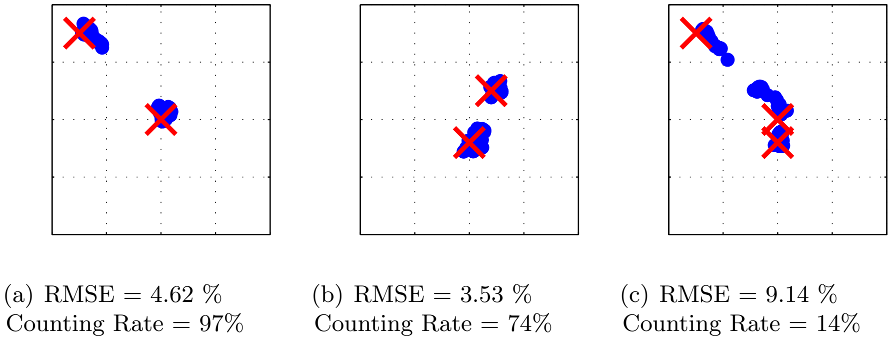
	$T_{60} = 250$ ms			$T_{60} = 400$ ms		
SNR	one source	two sources	three sources	one source	two sources	three sources
0 dB	98%	83%	51%	91%	68%	28%
5 dB	99%	89%	67%	97%	77%	46%
10 dB	98%	95%	88%	97%	93%	67%
15 dB	99%	98%	92%	96%	96%	80%
20 dB	98%	99%	94%	97%	97%	85%

6.4.2 Results of Real Measurements

We also performed some real recordings of omnidirectional acoustic sources in a 4-node square cell with sides $V = 4$ meters long. The nodes were 4-element circular microphone arrays of 2 cm radius. The sources were recorded speech signals of approximately 5 seconds duration, played back through loudspeakers at different locations and their SNR at the center of the cell was measured to be about 10 dB. The parameter setting for our method was the same as reported in Section 6.4.1. Figure 6.3 shows the results for different locations of two (Figure 6.3(a) & (b)) and three (Figure 6.3(c)) active sources. It can be seen that for the two source cases the method results in accurate counting and localization. For the three source case, the counting success rate is low, which is however also due to the fact that the two sources in the middle of the cell are located too close together. It should also be highlighted that these recordings took place outdoors, and while they may not have many reflections, there was a significant level of distant noise sources, such as dogs barking and cars passing by. Moreover, the locations and orientations of the arrays were not finely calibrated and had unintended offsets of a few centimetres and degrees. Thus, the conditions were far from ideal, making the results of our proposed localization and counting method even more encouraging.

6.5 Towards localization and counting in real-life conditions

In practical real-life scenarios several important issues arise that usually have a negative effect on the performance of any location estimation algorithm. One such important issue arises from limitations on the positioning of the microphone arrays in the room, which in real-life need to be placed close to walls in order not to pose restrictions on the activities of the speakers inside the room. The presence of such reflecting surfaces so close to the microphone arrays is generally known to have an adverse effect on the performance of the array, which can degrade the localization performance of the entire system. Also, the characteristics of real speakers in terms of their directivity pattern, spatial volume, and orientation are far more complicated than omnidirectional point sources which are usually used for simulations.



(a) RMSE = 4.62 %
Counting Rate = 97%

(b) RMSE = 3.53 %
Counting Rate = 74%

(c) RMSE = 9.14 %
Counting Rate = 14%

Figure 6.3: Location estimates (the blue clouds) throughout a 4-node square cell for real recordings of two and three active sources (the red X's). Each figure reports the location error (RMSE as % of the cell size $V = 4$ meters) and the source counting success rate.

In this section, we evaluate our source counting and localization method in a real environment with real speakers recorded by two microphone arrays which were placed near walls. To explicitly account for the performance degradation due to the positioning of the microphone arrays, we incorporate our method with our recently proposed method of [97] for DOA estimation which explicitly takes into account the early reflections that occur when the arrays are close to walls. Moreover, to investigate the challenges occurred in real-life conditions, we collected and present a dataset [1] of real recordings in a typical office room. We evaluate the performance of our system in this challenging dataset, which is publicly available in order to assist the research community move a step towards accurate localization in real-life scenarios.

In the remainder of this section, we first outline the "reflection-aware" model for DOA estimation in Section 6.5.1. Then, in Section 6.5.2 we describe the dataset of real recorded signals that we collected. Finally, we evaluate our source counting and location estimation method in this challenging dataset in Section 6.5.3.

6.5.1 "Reflection-aware" DOA estimation

Following [97], the DOA estimation accuracy for a circular array of N sensors and radius R, placed in front of a wall can be significantly improved by designing a propagation model which is aware of the earliest reflection introduced by the adjacent wall. Let us first review the typical propagation model, expressing the relative sound pressure at the mth microphone as a function of angular frequency ω and incident angle θ as

$$a_n(\omega, \theta) = e^{jkR\cos(\phi_n - \theta)}. \tag{6.9}$$

[1] Our dataset is publicly available at `https://github.com/spl-icsforth/WASN-Recordings-OfficeRoom`.

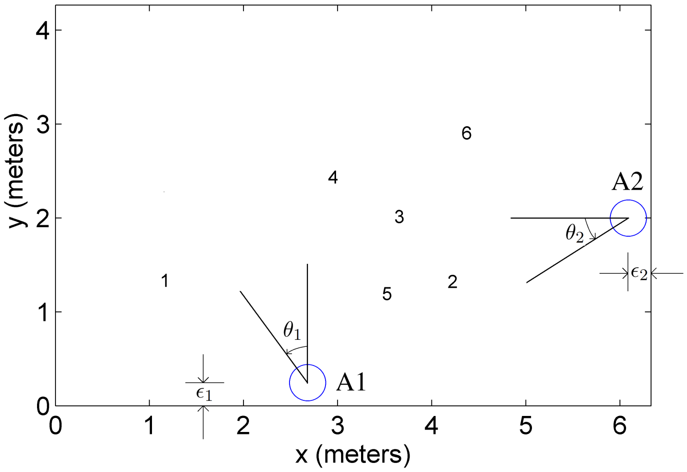


Figure 6.4: Recording setup used for gathering the real recorded dataset.

Here, $k = \omega/c$ is the wavenumber, c is the speed of sound and ϕ_n is the angle of the nth microphone which similar to θ, is defined with respect to the center of the circular disk. This typical propagation model accounts for the direct path of the sound only and ignores any distinct reflections that may occur.

A so-called half-space version of the propagation model associated to the same circular array can be designed with the model for the nth microphone defined as [97]

$$\hat{a}_n(\omega,\theta) = e^{jkR\cos(\phi_n-\theta)}e^{jk\epsilon\cos\theta} + he^{jkR\cos(\phi_n-\pi+\theta)}e^{-jk\epsilon\cos\theta}, \tag{6.10}$$

where ϵ is the distance of the array center from the adjacent wall, θ is the incident angle defined so that $\theta = 0^\circ$ is normal to the wall (Fig. 6.4), and $h \in [0,1]$ is the so-called Image Source Relative Gain (ISRG) which encodes the reflective properties of the wall. Assuming h to be real and constant with frequency is an affordable simplification, although in practice the wall reflectivity would be more accurately represented by a complex and frequency-varying ISRG [97]. Letting now $\hat{\mathbf{a}}(\omega,\theta) = [\hat{a}_1(\omega,\theta), \cdots, \hat{a}_N(\omega,\theta)]^T$ be the vector concatenating all the N terms from Eq. (6.10), the half-space steering vector is derived as

$$\mathbf{a}(\omega,\theta) = \hat{\mathbf{a}}(\omega,\theta)/\left\|\hat{\mathbf{a}}(\omega,\theta)\right\|_2, \tag{6.11}$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

Using any of the two propagation models, different approaches to DOA estimation can be employed. Here, we utilize a Minimum Variance Distortionless Response (MVDR) beamformer [5]. Performed in a time-frequency basis, with τ denoting the time-frame index, we find the narrowband DOA where the MVDR beamformer response is maximized by searching across the entire range of potential directions from 0 to 360 degrees.

6.5.2 Dataset of real recordings

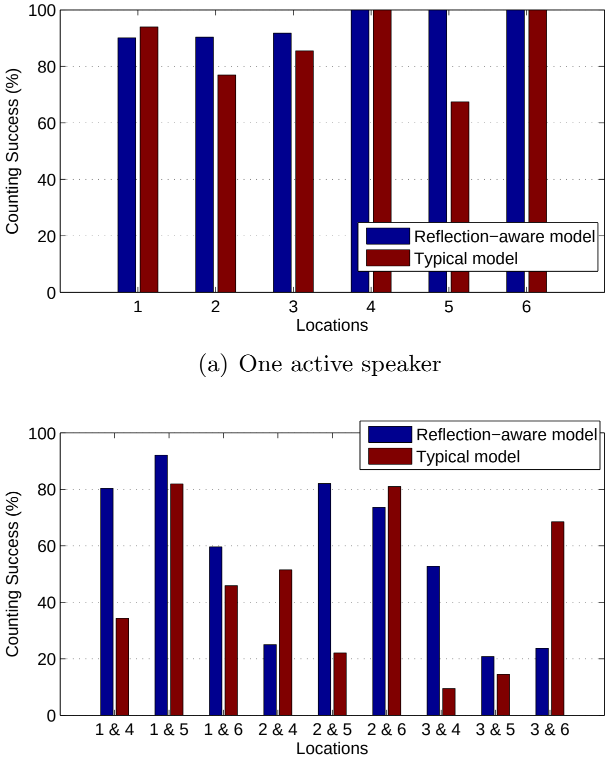
The dataset contains real recordings of speech sources in a typical office room of dimensions $L_x = 6.33$ and $L_y = 4.2$ meters with reverberation time of approximately equal to 400 ms. The recording setup is depicted in Fig. 6.4. The recordings were made at predefined locations (1–6 in Fig. 6.4), by two male speakers. The first speaker (M01) was recorded at locations 1–3, and the second one (M02) at locations 4–6. The speakers were asked to stand in the predefined locations with an orientation towards the center of the room, without further advising them about where to look at or how loud to speak.

We used two uniform circular microphones arrays. The array locations were measured to be $(2.68, 0.086, 1.20)$ meters for the first array (A01) and $(6.248, 2, 1.20)$ meters for the second array (A02). Both arrays were placed very close to walls: A01 is 8.6 cm and A02 is 8.2 cm away from the corresponding walls. Both arrays consisted of 8 Shure SM95 omnidirectional microphones and a radius of 5 cm. They operated individually (i.e., they were connected to different host PCs). Utterances for each speaker and each location were segmented from the original recordings and synchronized by eye-inspection. The signals were recorded at 48 kHz sampling rate.

6.5.3 Results and discussion

We used the recordings (downsampled at 12 kHz) of our dataset to evaluate our method in real-life conditions. For processing, we used an FFT of 512 samples length, with 50% overlap windowed with a square root hanning window and $\omega_c = 4$ kHz. For the per-frequency location estimates we used 1 second history length ($B = 46$ frames). The parameters for outlier rejection were set to $h_X = h_Y = 20$ cm and $q = 0.25$, as these parameters were found to perform best in most of the cases. ISRG was set to $h = 0.9$ and all other parameters were set according to Section 6.4.1 and [93, 97]. Finally, based on the geometry of the space and the positioning of the arrays, each array has a specific range of "allowable" DOAs, i.e., DOAs that can result in locations inside the defined localization area. When a narrowband DOA is not inside this "allowable" range, we reject it and we do not estimate a narrowband location for this frequency. The "allowable" range of DOAs was set to $[0°, 70°] \cup [290°, 360°)$ for A01 and $[0°, 45°] \cup [315°, 360°)$ for A02. To consider scenarios of two sources, we artificially added the microphone array signals at different locations and from different speakers. This resulted in 9 different cases, which represent all combinations of mixing speaker M01 at locations 1–3 with speaker M02 at locations 4–6. The energy of the speakers was not equalized to better model real-life conditions where energies will not be equal.

Fig. 6.5 depicts the counting success rate as the percentage of time-frames where the correct number of sources was found for the single-source and two-source case. The corresponding Root Mean Squared Error (RMSE) with error bars representing one standard deviation is shown in Fig. 6.6. For comparison, we include the results when the sensors use



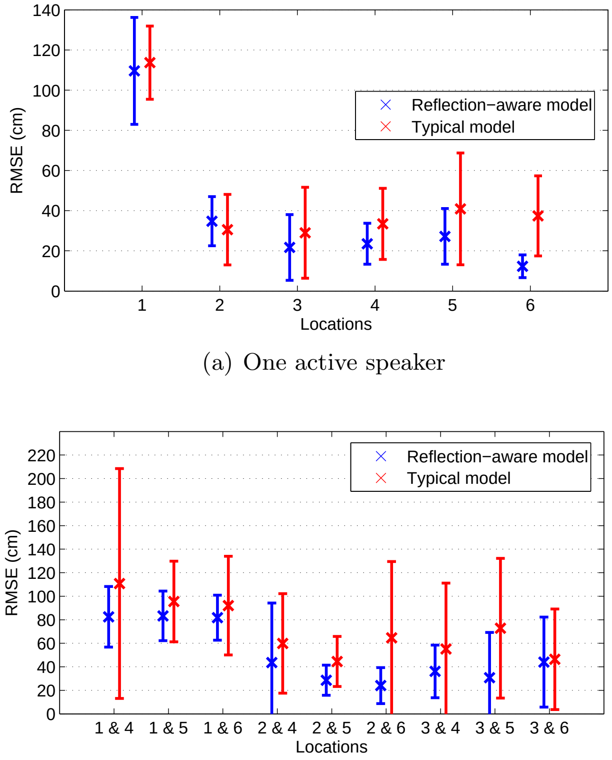
(a) One active speaker

(b) Two active speakers

Figure 6.5: Counting success rates for our dataset, using our "reflection-aware" and the typical model for DOA estimation.

the same DOA estimator but with the typical steering vector (i.e, not accounting for reflections). It is evident that when the sensors utilize our "reflection-aware" DOA estimator, we can achieve better performance: while the counting success rate is not always improved, the location estimation error is significantly reduced for all tested locations, especially in the two sources case.

Generally, our system operates in a functional range of values for these challenging scenarios, both in terms of counting and localization performance. It is also important to note how the performance varies in different locations. A large error is observed at location 1 in the single source case, which can be explained by the fact that this location is much further from the arrays (especially array A02). A performance degradation especially in terms of counting success rate is evident also in some cases of the two sources scenario (e.g, location pairs 2 & 4, 3 & 5), which can be attributed to the small distance between the sources, as well as to the small angular separation of the sources with respect to one (or both) the arrays. Such location pairs could benefit from the deployment of more microphone arrays.



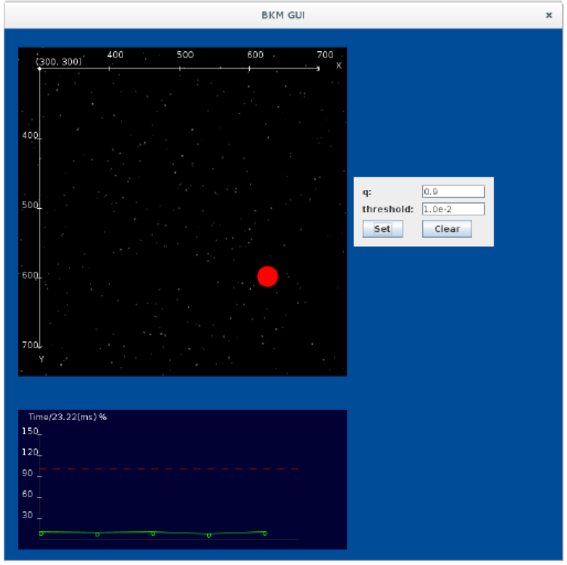
(a) One active speaker

(b) Two active speakers

Figure 6.6: Localization error for our dataset, using our "reflection-aware" and the typical model for DOA estimation.

These results also highlight the importance of evaluation across the entire localization cell, a direction towards which little effort has been made so far.

Fig. 6.8 shows the location estimates using our "reflection-aware" DOA estimation, for 12 out of 15 tested source locations (the three single-source cases with the smallest error have been omitted). The blue dots show the cloud of estimates over the entire duration of the signals for the time-frames where the correct number of speakers was found, revealing again a quite accurate localization. Finally, given that the locations and orientations of the arrays were not finely calibrated and had unintended offsets of a few centimetres and degrees, the conditions were far from ideal, making our results quite encouraging.



q: 0.9
threshold: 1.0e-2

Figure 6.7: The graphical user interface of our real-time localization and counting application.

6.6 Real-time implementation of source localization and counting using digital MEMS microphone arrays

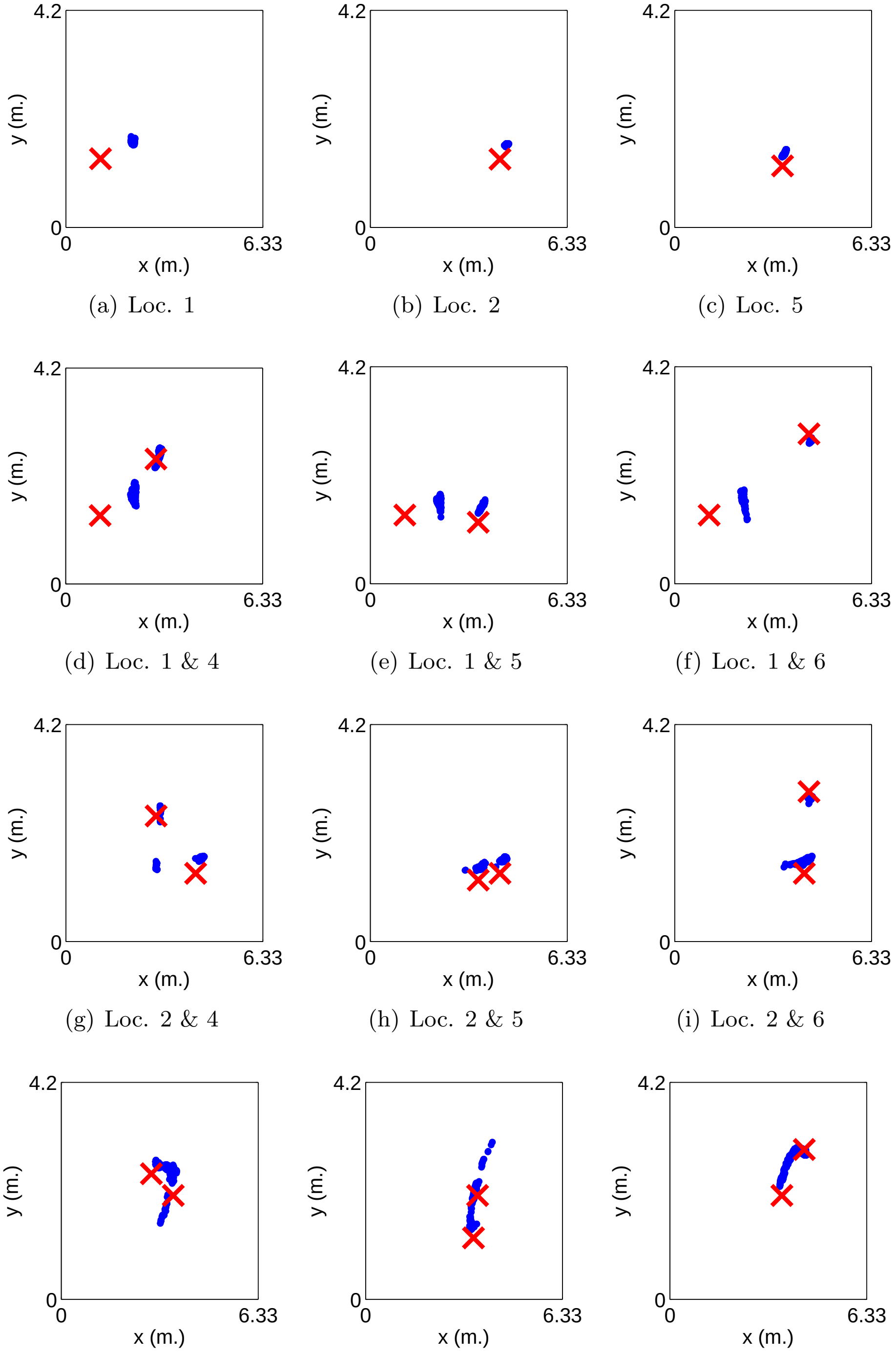
Finally, we implemented the source counting and localization method presented in this chapter on a real WASN where the sensors consist of digital MEMS microphone arrays constructed at FORTH-ICS. As our digital MEMS arrays do not yet have any processing resources, we connected each array to a laptop where the local processing for narrowband DOA estimation is performed according to Section 6.2. The narrowband DOA estimates for each time-frame are then transmitted to another PC which serves as the fusion center where the localization and counting method of Section 6.3 is executed. The estimated number of sources and their corresponding locations are then presented to the user through a Graphical User Interface (GUI) at the fusion center (Fig. 6.7) . The GUI also displays information about the required processing time, which validates that our method can run in real-time. Finally, through the GUI the user can set up the different parameters for the outlier rejection and bayesian K-means.

The local processing at the microphone arrays and the source counting and localization method at the fusion center are implemented in C++, while the graphical interface was implemented as a separate program in Java. We validated that our method can run in real-time and accurately localize a sound source. A video demonstration is available at `https://www.youtube.com/watch?v=bP1aF21BNZ4`.

6.7 Conclusions

In this chapter a method for source counting and location estimation in a wireless acoustic sensor network was presented. The method is based on statistical clustering of per-frequency location estimates in order to estimate the number of active sound sources and their corresponding locations. The per-frequency location estimates are inferred using per-frequency DOA estimates from the sensors. The efficiency of the method was validated using signals in simulated and real environments. To test the potential application of our method in real WASNs we collected a dataset of real recordings with real speakers in a typical office room. To account for the fact that in a practical application the microphone arrays have to be placed near walls, we incorporated our method with a model that explicitly takes into account the first reflection from the wall when performing the DOA estimation. We confirmed that our method with the "reflection-aware" DOA estimation can provide performance within a functional range of values in this challenging dataset of real recordings which is publicly available at `https://github.com/spl-icsforth/WASN-Recordings-OfficeRoom`. Our results on this dataset revealed the importance of evaluating a localization method within the entire localization area, as performance may vary for different source configurations. Finally, a real-time version of our method was developed and tested in a real two-node WASN with two MEMS microphone array sensors.

The next chapter focuses on the per-frequency DOA estimation task. Accurate estimation of the narrowband per-frequency DOAs is an important factor that affects the performance of the entire localization system. Thus, the next chapter will introduce a methodology for inferring more accurate per-frequency DOA estimates.



(a) Loc. 1 (b) Loc. 2 (c) Loc. 5

(d) Loc. 1 & 4 (e) Loc. 1 & 5 (f) Loc. 1 & 6

(g) Loc. 2 & 4 (h) Loc. 2 & 5 (i) Loc. 2 & 6

(j) Loc. 3 & 4 (k) Loc. 3 & 5 (l) Loc. 3 & 6

Figure 6.8: Location estimates (the blue clouds) for the real recordings of one [(a)–(c)] and two [(d)–(l)] speakers (the red X's).

Chapter 7
Improving narrowband DOA estimation

7.1 Introduction

The performance of any DOA-based sound source localization method depends, to a great extent, on the accuracy of the DOA estimates that are transmitted by the sensors. The approaches to the data-association, localization, and counting problems that are presented in the previous chapters utilize narrowband instantaneous DOA estimates for each Time-Frequency (TF) bin of the captured signals. To acquire those instantaneous DOAs, a narrowband DOA estimation method is applied on each TF-bin. Of course, the performance of any localization method depends on how accurately those DOAs are estimated. Irrespective of the narrowband DOA estimation method used, some TF-bins will suffer from erroneous DOA estimates due to noise and/or reverberation. In this chapter, we investigate the narrowband DOA estimation problem and propose a novel approach to more accurately estimate the instantaneous per TF-bin DOA estimates. Note that apart from the localization problem, the extraction of instantaneous DOA estimates is also crucial in other applications, such as source separation [98, 99], parametric spatial audio [100–102] and dereverberation [103].

Currently, the narrowband DOA estimation is applied to the microphone signals at each TF-bin. We propose an approach where instead of using the raw microphone signals in the TF-domain as input to the DOA estimation, an alternative input is generated through statistical modeling of the microphone signals at each TF-bin. More specifically, we utilize the complex Watson distribution to model the microphone signals at a given TF-bin and infer the maximum likelihood estimate of the distribution's mode vector, which is then used as input by the DOA estimation. The choice of the distribution is motivated by directional statistics where it is used to model uncertainties about directions of complex unit-norm vectors.

This distribution has been already used in audio signal processing, although in different ways: in [51, 52, 104] all TF-bins are used together and form a mixture of complex Watson distributions, which is utilized for speech separation [104], while in [51, 52] variational inference on the mixture parameters is employed to determine the number of mixture components which corresponds to the number of active sources. Lastly, the complex Watson distribution is used in [105] as a distance metric for wideband DOA estimation of a single source.

In this chapter we utilize a complex Watson distribution with different parameters for each TF-bin in order to provide an alternative input to a narrowband DOA estimation method. It is important to note that the outcome of our proposed method can be given as input to any narrowband DOA estimation method and thus our approach is independent of the DOA estimation method used and the array geometry. Our results, using simulations and real recordings, indicate that when the proposed method is used to generate the input for the DOA estimation, more accurate instantaneous DOA estimates are acquired. The methodology presented in this chapter was published in [106].

The rest of this chapter is organized as follows: Section 7.2 reviews the complex Watson distribution. Section 7.3 presents the proposed processing of the microphone array signals for more accurate instantaneous DOA estimation and Section 7.4 presents the performance evaluation of the proposed approach.

7.2 The complex Watson distribution

Let $\boldsymbol{y}$ denote a unit-norm d-dimensional complex random variable. The complex Watson distribution maps the observations of $\boldsymbol{y}$ to the d-dimensional complex unit hypersphere. The distribution is governed by the complex mode vector $\boldsymbol{\mu}$ and the real-valued concentration parameter κ, which describes how much the observations are concentrated around the mode vector. When $\kappa = 0$, the observations are uniformly distributed around the complex hypersphere. The probability density function of the complex Watson distribution is defined as [107]:

$$p(\boldsymbol{y}; \boldsymbol{\mu}, \kappa) = \frac{1}{c_{\mathcal{W}}(\kappa)} e^{\kappa |\boldsymbol{\mu}^H \boldsymbol{y}|^2} \tag{7.1}$$

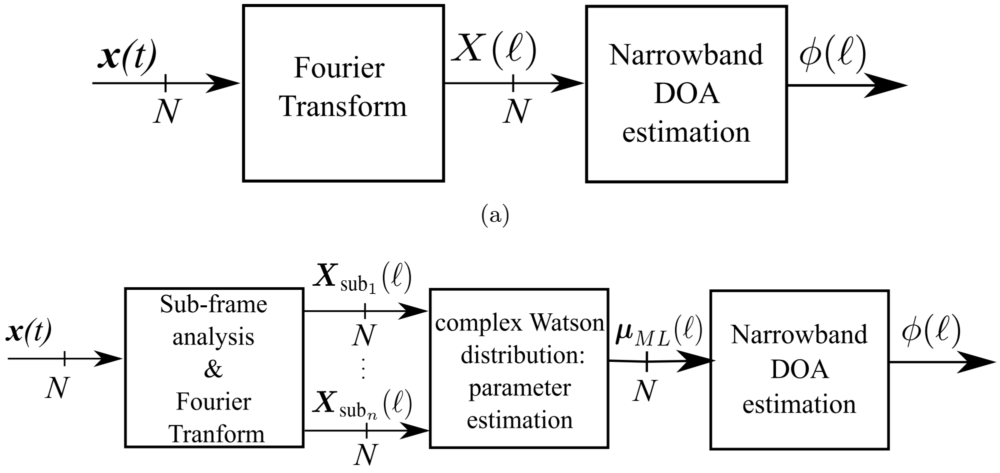
where $(\cdot)^H$ denotes the Hermitian transpose operator and $c_{\mathcal{W}}(\kappa)$ is the normalizing constant which is given by:

$$c_{\mathcal{W}}(\kappa) = \frac{2\pi^d \mathcal{M}(1, d, \kappa)}{(d-1)!} \tag{7.2}$$

with $\mathcal{M}(\cdot)$ being Kummer's confluent hypergeometric function [108].

Given a set $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$ that contains n observation vectors from a complex Watson distribution, maximum likelihood estimates of the distribution's parameters can be found by forming the $d \times d$ matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ as:

$$\boldsymbol{\Phi}_{\boldsymbol{y}} = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^H . \tag{7.3}$$


x(t)
N
Fourier Transform
$X(\ell)$
N
Narrowband DOA estimation
$\phi(\ell)$
x(t)
N
Sub-frame analysis & Fourier Tranform
$X_{\text{sub}_1}(\ell)$
N
$X_{\text{sub}_n}(\ell)$
N
complex Watson distribution: parameter estimation
$\mu_{ML}(\ell)$
N
Narrowband DOA estimation
$\phi(\ell)$

(a)

(b)

Figure 7.1: Block diagram for the estimation of the instantaneous DOA for frequency bin ℓ, using (a) the traditional input from the microphones and (b) our proposed methodology.

Let $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$ be the eigenvalues of $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ the corresponding eigenvectors. Then, the maximum likelihood estimate for the mode vector is given as the eigenvector that corresponds to the largest eigenvalue:

$$\boldsymbol{\mu}_{\text{ML}} = \boldsymbol{u}_1. \tag{7.4}$$

The maximum likelihood estimate of the concentration parameter is found as the solution to the equation:

$$\frac{\frac{\partial}{\partial \kappa_{\text{ML}}} \mathcal{M}(1, d, \kappa_{\text{ML}})}{\mathcal{M}(1, d, \kappa_{\text{ML}})} = \frac{1}{N} \boldsymbol{\mu}_{\text{ML}}^H \boldsymbol{\Phi}_{\boldsymbol{y}} \boldsymbol{\mu}_{\text{ML}} \tag{7.5}$$

which is highly non-linear since it involves ratios of confluent hypergeometric functions. As a result, one has to resort to numerical approximations to estimate κ_{ML} [107].

7.3 Proposed processing

Let $\boldsymbol{X}(\ell) = \left[X_1(\ell), \;\ldots,\; X_M(\ell)\right]^T$ be the $N \times 1$ vector of microphone signals in the frequency domain for frequency bin ℓ, using an L-length Fourier Transform. The traditional approach (Fig. 7.1(a)) would utilize $\boldsymbol{X}(\ell)$ as input to the narrowband DOA estimation method in order to infer the DOA estimate $\phi(\ell)$. In our approach, we model the Fourier coefficients in $\boldsymbol{X}(\ell)$ with a complex Watson distribution. Then, to estimate $\phi(\ell)$ we use the maximum likelihood estimates of the distribution parameters as input to the DOA estimation method

(Fig. 7.1(b)), instead of vector $\boldsymbol{X}(\ell)$ itself.

Our processing starts by dividing an L-length frame into sub-frames of length $L_{\text{sub}} < L$ samples with a time shift of K_{sub} samples windowed with a Hamming window. The number of sub-frames is given by:

$$n = \lfloor L/K_{\text{sub}} \rfloor - 1 \tag{7.6}$$

where $\lfloor \cdot \rfloor$ denotes the floor operator.

The sub-frames are transformed into the frequency domain using an L-length Fourier Transform, resulting for each frequency $\ell = 1, \ldots, L$ in the set of Fourier coefficients

$$\mathcal{Y}(\ell) = \{\boldsymbol{X}_{\text{sub}_1}(\ell), \boldsymbol{X}_{\text{sub}_2}(\ell), \ldots, \boldsymbol{X}_{\text{sub}_n}(\ell)\} \tag{7.7}$$

where $\boldsymbol{X}_{\text{sub}_i}(\ell)$ is a $M \times 1$ vector of Fourier coefficients from the N microphones for the i-th sub-frame.

The samples in $\mathcal{Y}(\ell)$ form the observation vectors which are normalized to unit-norm according to:

$$\bar{\boldsymbol{X}}_{\text{sub}_i} = \boldsymbol{X}_{\text{sub}_i} / ||\boldsymbol{X}_{\text{sub}_i}|| \tag{7.8}$$

where $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^H \boldsymbol{x}}$ denotes the norm of the vector $\boldsymbol{x}$. Unit-norm normalization represents a mapping of the observation vectors to the complex unit hypersphere, albeit preserving the spatial information which is essential for the DOA estimation. The n normalized observation vectors in:

$$\bar{\mathcal{Y}}(\ell) = \{\bar{\boldsymbol{X}}_{\text{sub}_1}(\ell), \bar{\boldsymbol{X}}_{\text{sub}_2}(\ell), \ldots \bar{\boldsymbol{X}}_{\text{sub}_n}(\ell)\} \tag{7.9}$$

can now be assumed to follow a complex Watson distribution with parameters $\boldsymbol{\mu}(\ell)$ and $\kappa(\ell)$. The maximum likelihood estimate $\boldsymbol{\mu}_{\text{ML}}(\ell)$ for the mode vector can then be found using the samples in $\bar{\mathcal{Y}}(\ell)$ as described in Section 7.2. Finally, DOA estimation using an arbitrary narrowband DOA estimation method is applied using the estimated mode vector $\boldsymbol{\mu}_{\text{ML}}(\ell)$ as input. Note that, we only change the input to the narrowband DOA estimation method: instead of the traditional approach that utilizes the frequency domain coefficients from the microphone signals at each TF-bin, we utilize these coefficients to infer an estimate of the mode vector of a complex Watson distribution and give this estimate as input to the DOA estimation. Fig 7.1 presents block diagrams that compare the traditional approach to the proposed one. Also, as we do not alter anything in the DOA estimation procedure, our proposed methodology can be used with any narrowband DOA estimation method and array geometry.

7.4 Results and discussion

To evaluate the performance of our proposed approach, we used simulations and real recorded signals. We considered two narrowband DOA estimation methods: the one described in [4] (denoted as [Karbasi et al.] and the well-known narrowband Multiple Signal Classification (MUSIC) algorithm [109]. We used both DOA estimation methods to estimate the instantaneous DOAs at each TF-bin and compare their performance when using the traditional input to the methods (Fig. 7.1(a)) and the one derived by our proposed methodology (Fig. 7.1(b)). For processing, we used frames of $L = 2048$ samples with 50% overlap. The FFT size was 2048. For our methodology, each frame was divided into sub-frames of $L_{\text{sub}} = 256$ samples with a time shift of $K_{\text{sub}} = 128$ samples which results in $n = 15$ sub-frames, i.e., $n = 15$ observation vectors for each frequency in order to perform the maximum likelihood estimation.

7.4.1 Simulation results

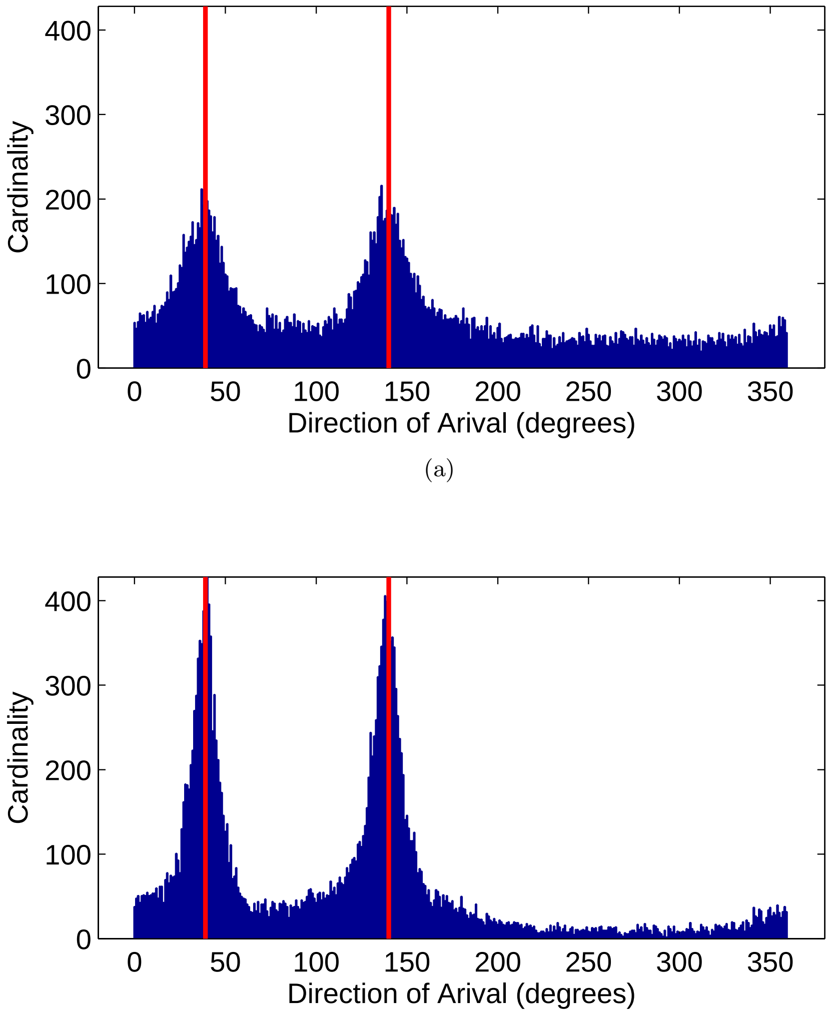
We used the Image-Source method [91] to simulate a room of dimensions of $6 \times 4 \times 3$ meters, characterized by reverberation time of $T_{60} = 400$ ms. We used a uniform circular microphone array with $N = 8$ microphones and a radius of 5 cm. The array was placed at the center of the room at 1 m height. For the given array geometry, the highest frequency of interest in order to avoid spatial aliasing is $\ell_{\text{max}} = 4$ kHz. Thus, in all our results we consider narrowband DOA estimation in all TF-bins below ℓ_{max}.

In each simulation, the sound sources were speech recordings of equal power and duration of 3 seconds sampled at 44.1 kHz. The signal-to-noise ratio (SNR) at each microphone was measured as the power of each source signal to the power of the noise. To simulate different SNR values we added white Gaussian noise at each microphone, uncorrelated with the source signals and the noise signals at the other microphones.

We considered scenarios of two and three simultaneously active sources. To more accurately measure the performance around the array, each scenario was repeated 50 times and the sources were located at random directions around the array with uniform probability and an angular separation between them of at least 30°. The sources were placed at 1 m height and their distance from the array was set to 1.5 m.

First, to qualitatively show the advantage of our proposed methodology, we consider two active sources at 39° and 140° and 20 dB SNR. Fig. 7.2 depicts the histogram of instantaneous DOA estimates using MUSIC with the traditional input (Fig. 7.2(a)) and the input derived by our proposed methodology (Fig. 7.2(b)). It is obvious that when the proposed input is given, MUSIC can more accurately estimate the instantaneous DOAs: the cardinality of DOA estimates very close to the true sources' DOAs is increased substantially, while erroneous estimates which are far away from the sources occur less often.

To compare the DOA estimation accuracy of the two aforementioned DOA estimation methods when using the traditional input from the microphone array and when using as input the one derived by our proposed methodology, we count the percentage of TF-bins in which an accurate instantaneous DOA has been estimated. We consider that a DOA is accurate

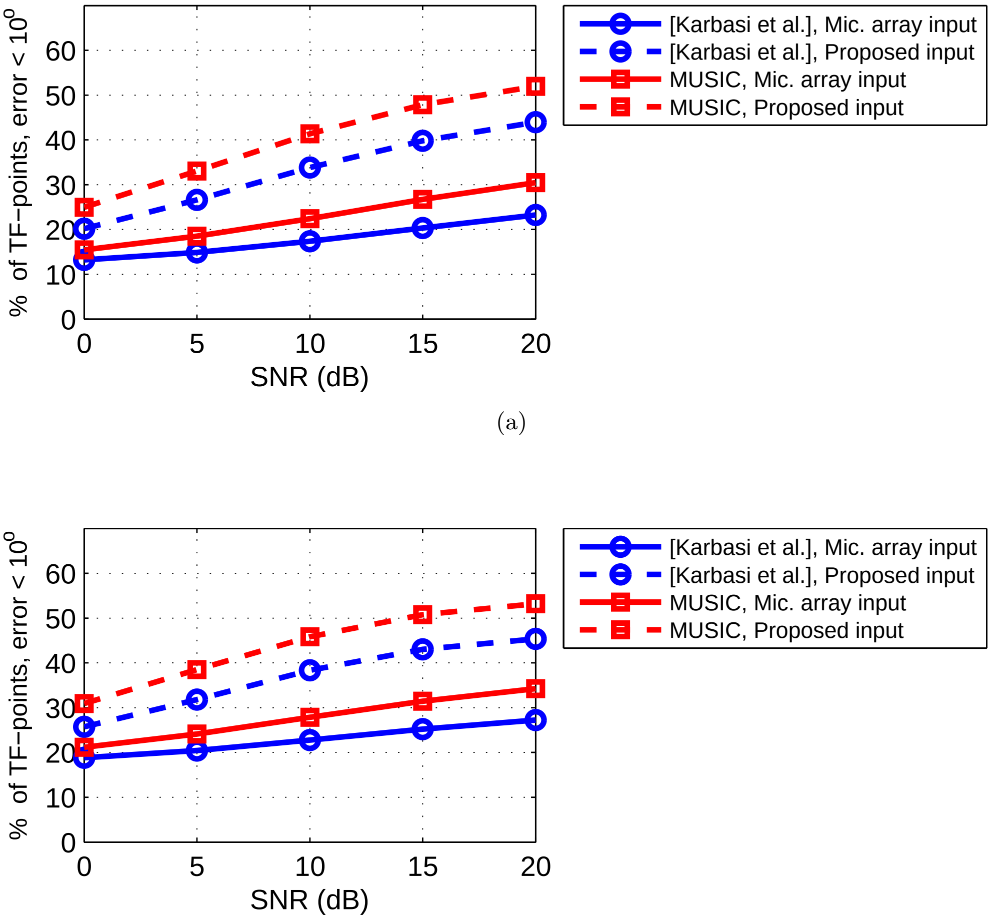


(a)

(b)

Figure 7.2: Histogram of instantaneous DOA estimates obtained using MUSIC with (a) the traditional input and (b) the input derived by the proposed methodology, for a simulated recording of two active sources at 39° and 140° with reverberation time of $T_{60} = 400$ ms and 20 dB SNR. The vertical lines correspond to the true sources' DOAs.

if its absolute error from a source's true DOA is less than 10°. Fig. 7.3 depicts the results for two and three simultaneous sources. It can be observed that our proposed methodology results in more accurate instantaneous DOA estimates compared to using the traditional input, for all SNR cases and number of active sources and for both DOA estimation methods that we consider. Especially at the higher SNR values, the number of TF-bins with accurate

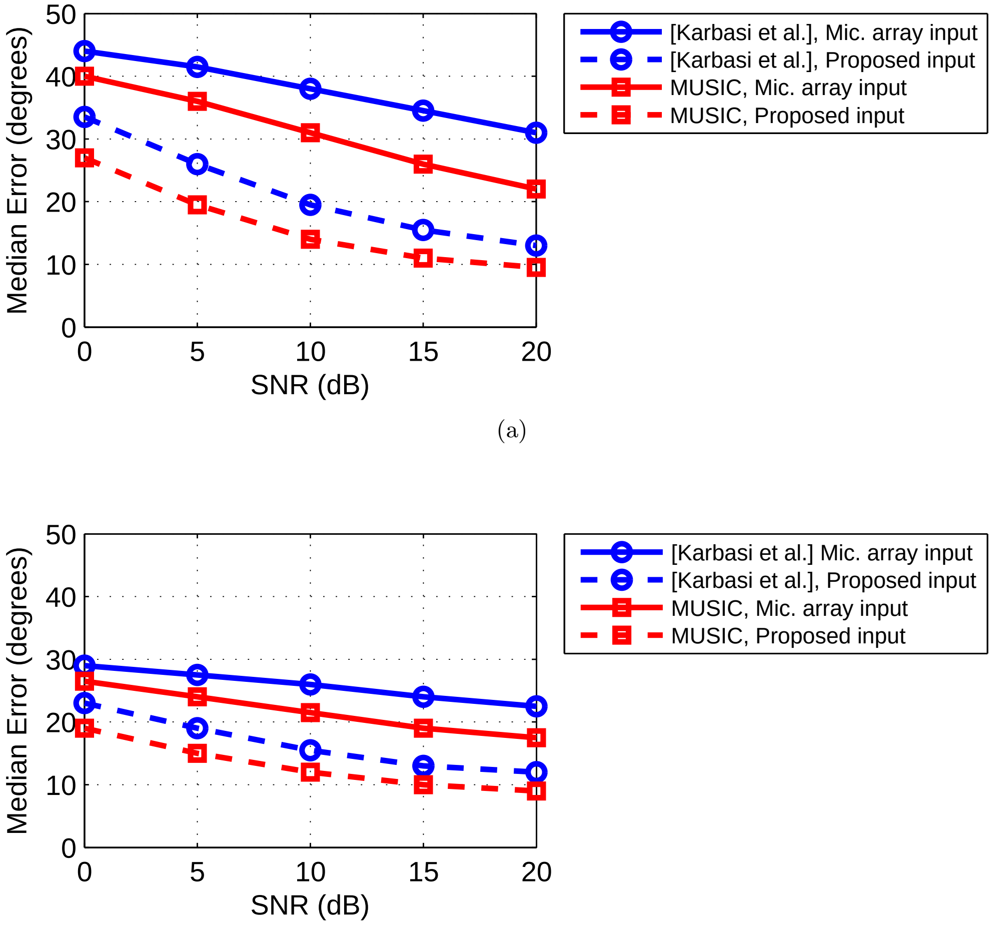


(a)

(b)

Figure 7.3: Percentage of TF-bins that exhibit DOA estimation error less than 10° for various SNR levels for the two DOA estimation methods when using the traditional input from the array and when using the input derived by our methodology for (a) two and (b) three simultaneously active sound sources.

DOA estimates is increased by approximately 20% when using our proposed methodology, showing our method's capability to improve the accuracy of DOA estimation methods and to offer a more accurate parametric spatial modeling of the acoustic environment. Fig. 7.3 also reveals that, irrespective of the DOA estimation method, the traditional approach results in only a fraction of TF-bins with accurate DOA estimates, while the majority of them suffer from noisy estimates. This highlights the need for more accurate approaches to estimate instantaneous DOAs and further motivates this study. Finally, as expected MUSIC performs better than the method of [4] either when the traditional or the proposed input



(a)

(b)

Figure 7.4: Median DOA estimation error for various SNR levels for the two DOA estimation methods when using the traditional input from the array and when using the input derived by our methodology, for (a) two and (b) three simultaneously active sound sources.

to the method is used, which is due to the superior performance of subspace approaches to DOA estimation. Moreover, Fig. 7.4 depicts the corresponding median estimation error, which shows the gain in estimation accuracy that can be obtained when using our proposed methodology to generate the input for the DOA estimation method. It can be observed that our proposed methodology reduces the DOA estimation error by approximately 10° to 15° for all cases. Finally, comparing between the two and three source cases in Figs. 7.3 & 7.4, one can observe a better performance when three active sources are considered. Although, counter-intuitive at a first glance, it can be explained by the fact that we measure the DOA error of each TF-bin from its closest source. Thus, some TF-bins may exhibit reduced error, since it is more probable to find a source close to their estimated DOA when more sources are considered. However, this fact is of no significant importance, since our goal is to compare the two different inputs to the DOA estimation methods for the same scenarios.

7.4.2 Comparison with a steering vector estimation approach

The proposed methodology shares many similarities with steering vector estimation approaches that estimate the steering vector as the principal eigenvector of the spatial correlation matrix of the observation vectors. A difference between this steering vector estimation approach and the proposed methodology is that in the proposed methodology the observation vectors have to be first normalized to unit-norm according to Eq. (7.8). In this section, we compare the performance of this steering vector estimation approach with the proposed methodology. We utilize the method of [4] for DOA estimation and we compare two approaches: in the first approach the input to the DOA estimation method is derived using the proposed methodology (i.e., the maximum likelihood estimate of the mode vector of the complex Watson distribution) and in the second approach the input to the DOA estimation method consists of the steering vector which is estimated as the principal eigenvector of the spatial correlation matrix of the observed signals at the microphones. The basic difference between the two approaches lies on whether a normalization to unit-norm is performed at the set of Fourier coefficients of (7.7).

Fig. 7.5 depicts the percentage of TF-bins in which a DOA with less than 10^o error is estimated and Fig. 7.6 shows the corresponding median DOA estimation error for the same simulations of three active sound sources described in Section 7.4.1. It can be observed that both approaches result in very similar performance, however the proposed methodology performs marginally better.

7.4.3 Results using real recorded signals

We also conducted experiments in a typical office of approximately the same dimensions as in the simulations. We used a circular microphone array of 5 cm radius with eight Shure SM93 microphones and a TASCAM US2000 USB sound card. The reverberation time of the room was measured to be approximately $T_{60} = 400$ ms. We demonstrate the performance gain of our proposed methodology in a real recording of 45 seconds duration with three active speakers at 0°, 160°, and 240°, and 1.5 m away from the array, which was placed on a table at the center of the room.

Fig 7.7 shows the histogram of instantaneous DOA estimates obtained with MUSIC with the traditional microphone array input (Fig. 7.7(a)) and the one obtained using the proposed methodology as input to MUSIC (Fig. 7.7(b)). In accordance with the simulations, it is evident that with our methodology MUSIC can infer much more accurate instantaneous DOA estimates: the cardinality of the DOAs estimated very close to the sources' DOAs

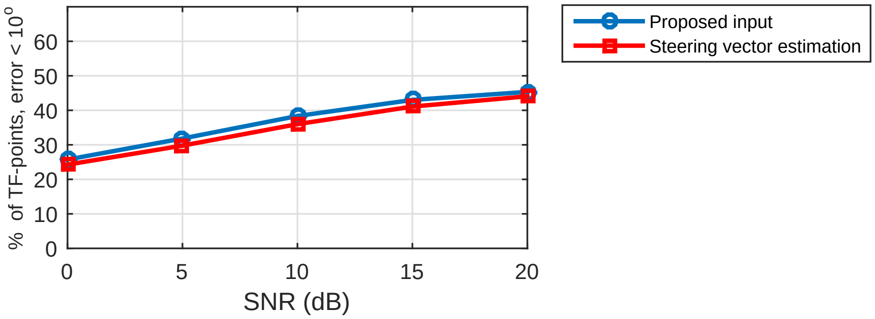


Figure 7.5: Percentage of TF-bins that exhibit DOA estimation error less than 10^o for various SNR levels for three active sound sources for the DOA estimation method of [4] when using as input the one derived by our proposed methodology and when using as input the steering vector estimated as the principal component of the spatial correlation matrix of the observed signals.

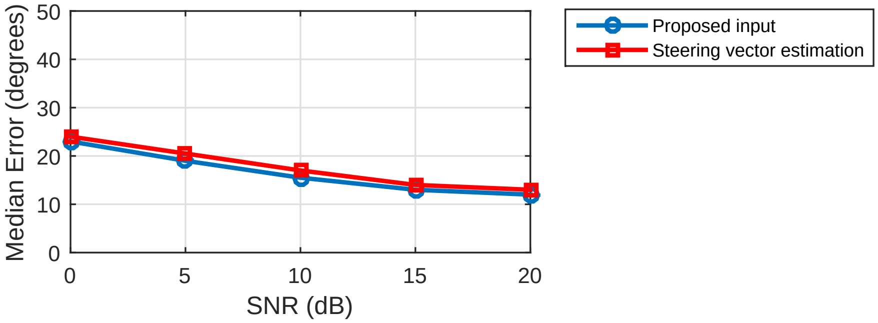


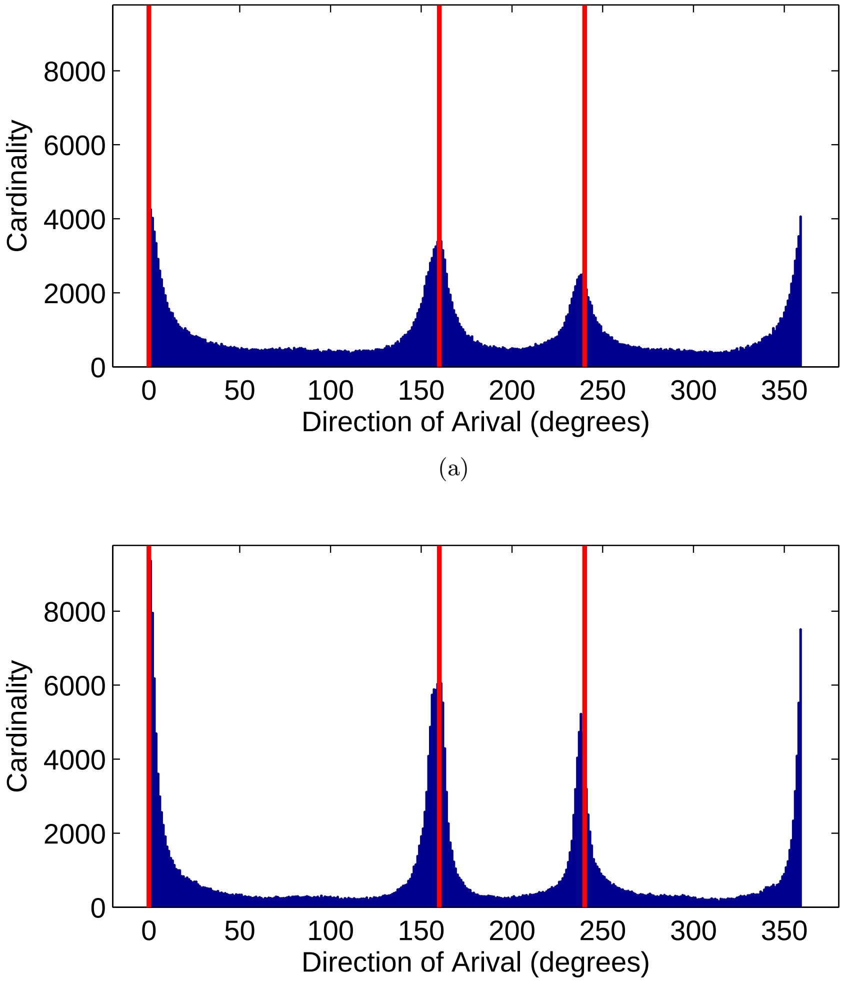
Figure 7.6: Median DOA estimation error for various SNR levels for three active sound sources for the DOA estimation method of [4] when using as input the one derived by our proposed methodology and when using as input the steering vector estimated as the principal component of the spatial correlation matrix of the observed signals.

in much higher in Fig. 7.7(b), while the cardinality of noisy DOA estimates that are away from the sources is reduced. Similar results were obtained in the histograms using the DOA estimation method of [4]. For each TF-bin, we also calculated the absolute DOA estimation error for both DOA estimation methods with the traditional and proposed input. Fig. 7.8 depicts the empirical Cumulative Distribution Function (CDF) of the error. The CDF shows the probability in the y-axis for the error to be less or equal to the corresponding value in the x-axis. The performance gain when using our proposed input to the DOA estimation methods is again evident. According to Fig. 7.8, in 50% of the cases (TF-bins) the estimation error using the traditional input from the microphones is approximately 10 to 20 degrees depending on the method, and reduces to less than 6 degrees for both methods when our proposed input is considered. Finally, using the proposed input an approximately 20% gain is achieved in the number of TF-bins whose DOA error is less than 10°: from 35% when using the traditional input to 56% when using the proposed one for the method of [4], and from 44% to 62% for MUSIC.

7.5 Conclusions

This chapter presented a methodology for more accurate estimation of the narrowband instantaneous directions of arrival. The methodology utilizes the complex Watson distribution in order to model the microphone array signals in each time-frequency bin and infer the maximum likelihood estimate of the distribution's mode vector. The mode vector is then given as input to a DOA estimation method. Thus, instead of using the microphone array signals in order to estimate the DOAs, the estimate of the mode vector of the complex Watson distribution is used as input to the DOA estimation method. The methodology can be combined with any DOA estimation method and microphone array geometry. The evaluation was carried out on simulated and real recordings using two different DOA estimation methods. The results revealed that compared to the traditional approach where the microphone array signals are used as input to the DOA estimation method, our proposed approach can estimate the instantaneous DOAs with higher accuracy.

The next chapter considers a different aspect of source localization and demonstrates some examples on how location information can be used for speech enhancement and separation. It will discuss the potential use of location information for the design of beamformers that separate the sources' signals or enhance the signal of a target source.



(a)

(b)

Figure 7.7: Histogram of instantaneous DOA estimates obtained using MUSIC with (a) the traditional input and (b) the input derived by the proposed methodology, for a a real recording of three active sources at 0°, 160°, and 240° in a room of reverberation time of $T_{60} = 400$ ms. The vertical lines correspond to the true sources' DOAs.

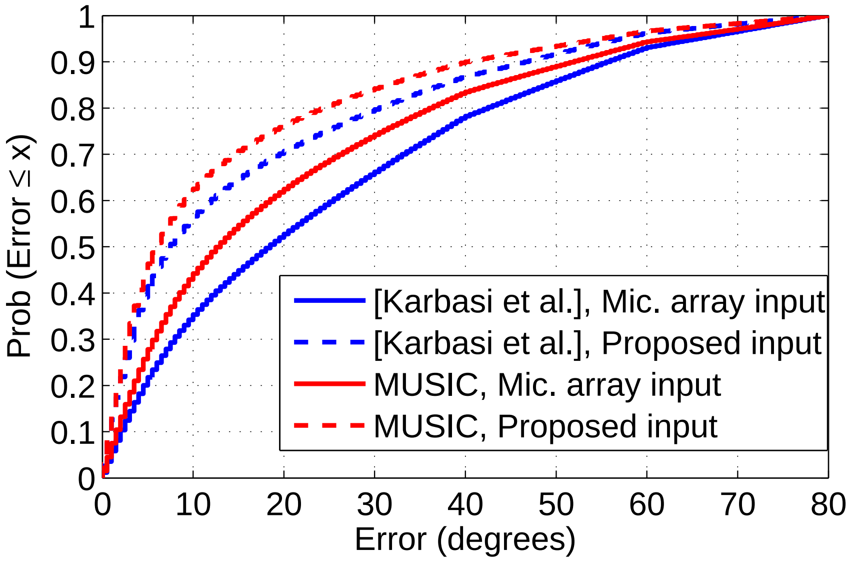


Figure 7.8: Empirical Cumulative Distribution Function of the error of the instantaneous DOA estimates for the two DOA estimation methods when using the traditional input from the array and when using the one derived by our methodology, for a real recording with three active sound sources.

Chapter 8
Location information for speech enhancement and separation

8.1 Introduction

Extracting information about the spatial characteristics of the acoustic environment and the active sound sources enables a parametric spatial modeling of the soundscape. Typically, this parametric spatial modeling is achieved using information related to the number and directions of the active sound sources and the diffuseness of sound, which is an index that indicates whether the sound field at specific frequencies or frequency bands is dominated by directional or diffuse sound (i.e., sound with no prominent direction). This spatial information has been used in numerous applications, such as speech enhancement and source separation [98, 99], robust speech recognition [110, 111], dereverberation [103], and spatial audio [100–102].

In the previous chapters we focused on methods that can estimate the locations of the multiple active sound sources in wireless acoustic sensor networks. In this chapter, we discuss the potential use of location information in applications involving speech enhancement and separation. To this direction, we provide two examples of applications that can benefit from the use of location information about the active sound sources. Our first example considers the problem of spatial audio capturing for immersive audio applications and extends ImmACS—the Immersive Audio Communication System that we have previously proposed—by allowing it to utilize multiple microphone arrays and location information. Currently, ImmACS uses a single microphone array and performs spatial audio capturing by utilizing information related to the DOAs of the active sound sources in the environment. ImmACS utilizes source separation using beamforming and post-filtering in order to separate the source signals and reproduce each source from its corresponding estimated direction. In this example we present an approach that allows multiple arrays to cooperate using location information in order to improve the source separation stage of ImmACS.

In our second example, we consider an MVDR beamformer for speech enhancement in cocktail party situations where multiple sources are simultaneously active. Our beamformer uses location information in order to estimate the steering vector which is needed when computing the beamformer filter coefficients.

8.2 Location information for spatial audio capturing

Extracting the spatial characteristics of the acoustic environment is the most crucial task for spatial audio capturing and reproduction. Usually, the spatial characteristics consist of direction of arrival estimates that provide the angle from which each frequency of the captured signals will be reproduced. Based on our previous work of [112, 113], we have developed ImmACS: an Immersive Audio Communication System. The goal of ImmACS is to capture the soundscape at the recording side using a microphone array and reproduce it using multiple loudspeakers or headphones in real-time. ImmACS encodes the soundscape by separating the signals of the sources that are active in the environment based on their directions. The soundscape is thus described by the separated source signals and their corresponding directions. The capturing and reproducing sides of ImmACS can be located far apart, so that the encoded soundscape is transferred through the Internet. ImmACS also gives the listeners the ability to select the directions they want to hear and attenuate the sources that come from other directions. For these features, source separation is important to provide accurate spatial impression or to reproduce specific sources while attenuating others in the soundscape.

In this section, we first review ImmACS and then present some preliminary results on how our method can be extended to handle multiple microphone arrays. Motivated by situations where a single microphone array cannot provide sufficient spatial coverage—such as when the angular separation of sources is very small or the sources have the same DOA with respect to the array—we extend ImmACS by allowing multiple arrays to cooperate based on location information, in order to improve the source separation stage, thus providing better spatial audio capturing and reproduction. The work presented in this section has been published in [114].

8.2.1 Directional coding of the soundscape with ImmACS

Assume that K active sources are in the far-field of a circular microphone array with N microphones. The microphone array signals are transformed to the Short-Time Fourier Transform (STFT) domain. Then the number of active sound sources and their DOAs are estimated using our previously proposed method of [18], which is capable of estimating the DOAs with high accuracy in reverberant environments. The DOA estimation is applied in each time-frame τ and outputs the number of active sources $\hat{K}_\tau$ and the estimated DOA vector—with 1° resolution—$\boldsymbol{\theta}_\tau = \left[\theta_1 \cdots \theta_{\hat{K}_\tau}\right]$.

Based on the estimated DOA vector, we employ a fixed superdirective beamformer in order to separate the source signals that come from different directions. The beamformer is designed to maximize the array gain while maintaining a minimum constraint on the white noise gain [115]. Thus, the beamformer filter coefficients are given by:

$$\boldsymbol{w}(\ell, \theta_s) = \frac{\left[\epsilon \boldsymbol{I} + \boldsymbol{\Gamma}(\ell)\right]^{-1} \boldsymbol{d}(\ell, \theta_s)}{\boldsymbol{d}(\ell, \theta_s)^H \left[\epsilon \boldsymbol{I} + \boldsymbol{\Gamma}(\ell)\right]^{-1} \boldsymbol{d}(\ell, \theta_s)} \tag{8.1}$$

where $\boldsymbol{w}(\ell, \theta_s)$ is the $N \times 1$ vector of complex filter coefficients for frequency ℓ and steering direction θ_s, $\boldsymbol{d}(\ell, \theta_s)$ is the steering vector of the array, $\boldsymbol{\Gamma}(\ell)$ is the $N \times N$ noise coherence matrix (assumed spherically isotropic diffuse), $(\cdot)^H$ is the Hermitian transpose operation, $\boldsymbol{I}$ is the identity matrix, and ϵ is used to control the white noise gain constraint. The steering vector is estimated from the steering direction under the assumption of plane wave propagation in anechoic conditions. The filter coefficients are fixed and estimated offline.

In the τ-th time frame, based on the number of active sound sources, we employ $\hat{K}_\tau$ concurrent beamformers resulting in the beamformed signals:

$$B_k(\tau, \ell) = \sum_{n=1}^{N} w_n(\ell, \theta_s) X_n(\tau, \ell), \quad k = 1, \cdots, \hat{K}_\tau \tag{8.2}$$

where $X_n(\tau, \ell)$ is the STFT of the signal recorded at the n-th microphone of the array.

The beamformed signals are given as input to a post-filter. The goal of the post-filter is twofold: it produces the final separated source signals and it allows for downmixing all the source signals into one monophonic audio signal. Based on the beamformed signals, the post-filter estimates $\hat{K}_\tau$ binary masks as follows:

$$U_k(\tau, \ell) = \begin{cases} 1, & \text{if } k = \arg\max_p |B_p(\tau, \ell)|^2, p = 1, \cdots, \hat{K}_\tau \\ 0, & \text{otherwise} \end{cases} \tag{8.3}$$

According to Eq. (8.3), for each frequency bin the post-filter keeps only the source with the highest energy (i.e., most dominant) and sets all the other sources at that frequency bin to zero. Thus, the masks are orthogonal to each other, meaning that for each frequency bin only one source is maintained while the other sources are set to zero. Each binary mask is applied to its corresponding beamformed signal to yield the final separated source signals:

$$\hat{S}_k(\tau, \ell) = U_k(\tau, \ell) B_k(\tau, \ell), \quad k = 1, \cdots, \hat{K}_\tau \tag{8.4}$$

Finally, the orthogonality property of the binary masks, allows us to efficiently downmix all the source signals into one full spectrum signal by summing them up. As a result, one audio signal and side-information consisting of the DOA of the source that is dominant in each time-frequency bin, are used to encode the entire soundscape. In [112] we demonstrated that it is possible to encode the audio signal with an MP3 encoder without any loss in spatial impression and we also proposed a coding scheme for the side-information channel.

8.2.2 Incorporating multiple microphone arrays and location information

ImmACS and other related methods usually assume that the microphone array is placed in the middle of the acoustic environment that is encoded. While this is suitable for applications like teleconferencing where people are located around a room, or recording a music performance where the orchestra is placed in the front area of the microphone array, there are other scenarios where a single array cannot provide sufficient spatial coverage. In such scenarios, the sound sources may be located such that their angular separation is too small for the array to isolate them, or the sources may even be located such that they have the same DOA with respect to the array, making the discrimination of the sources impossible.

For these reasons, we investigate the use of multiple microphone arrays combined with location information about the sound sources in order to isolate them and encode the soundscape. Source separation is an important aspect, as in order to provide accurate spatial impression each source signal that will be reproduced from a specific direction must not contain interfering sources. Moreover, it enables the listeners to "focus" the reproduction on a specific sound source by choosing to reproduce that source only and attenuate all the other sources present in the soundscape.

On the recording side, multiple arrays are placed to monitor the area. Assuming that the locations of the sources are known—or can be estimated—each microphone array can calculate the DOAs of the sources with respect to that array by:

$$\theta_{m,k} = \arctan \frac{p_{\mathrm{y},k} - q_{\mathrm{y},m}}{p_{\mathrm{x},k} - q_{\mathrm{x},m}} \tag{8.5}$$

where $\theta_{m,k}$ is the DOA of the k-th source with respect to the m-th microphone array, $\boldsymbol{p}_k = \begin{bmatrix} p_{\mathrm{x},k} & p_{\mathrm{y},k} \end{bmatrix}^T$ and $\boldsymbol{q}_m = \begin{bmatrix} q_{\mathrm{x},m} & q_{\mathrm{y},m} \end{bmatrix}^T$ are the locations of the k-th sound source and the m-th microphone array respectively. We also assume that the microphone arrays are all connected to the central node that carries all the spatial audio capturing operations, thus providing synchronized signals.

We will try to address the following question: what is the best policy for microphone array selection so as to achieve the best source isolation for reproduction?

8.2.3 Beamforming and post-filtering from the closest array for each source

As the locations of the sources are known—or estimated—a natural approach would be to isolate each source using the closest array to the source, as it is expected that this array would have the highest Signal-to-Noise Ratio (SNR) for the source of interest. This approach works in the following way:

1. The microphone array closest to a source is selected, based on the source's location.
2. The DOAs of all the active sources to that array are calculated via (8.5). Beamforming and post-filtering are carried out as described by (8.2)–(8.4) using the signals from the

selected array.

From the $\hat{K}_\tau$ final separated source signals only those of the sources that are closest to that array are maintained, while the separated signals of the other sources are discarded, as they will be estimated from the array that is closest to them. Finally, each microphone array will contribute with the separated signals of the sources that are closest to it.

In this scheme, each microphone array estimates its own post-filter. Thus, the binary masks are no longer orthogonal which does not allow the encoding of the soundscape in one audio signal. Moreover, each array has to beamform to all sources—in order to estimate and apply the post-filter—even though only the closest ones are maintained. As a result, unnecessary beamforming operations are carried out and the computational complexity increases proportionally to the number of microphone arrays. An important problem may arise when the sources are far apart but at a small angular separation with respect to an array. As the post-filter compares energies and energy decreases with distance, the array aiming to separate its closest source will provide poor beamformed signals for the sources that are far away—and act as interferers—degrading the source isolation performance.

8.2.4 Beamforming and cooperative post-filtering

An alternative approach is to allow the microphone arrays to cooperate in order to design a single post-filter that separates all source signals. In this scheme, each microphone array remains responsible for the sources that are closest to it, but it does not individually estimate its own post-filter. This approach works in the following way:

1. Based on the sources' locations, the closest microphone array for each source is selected and the DOA for that source with respect to that array is calculated using (8.5).
2. In contrast to the method in Section 8.2.3, each array beamforms only to the sources that are closest to it using (8.2).
3. The beamformed signals $B_k(\tau, \ell),\ k = 1, \cdots, \hat{K}_\tau$ that now come from different arrays are used to estimate a single post-filter using (8.3).
4. The final separated signals are estimated via (8.4).

This scheme is more computationally efficient as for $\hat{K}_\tau$ number of sources only $\hat{K}_\tau$ beamforming operations are needed. Moreover, as a single post-filter is used, the orthogonality property holds, which allows ImmACS to encode the entire soundscape into one monophonic signal and side-information. Note that, as the locations of the sources are known, the side-information can contain the locations—and not DOAs only—of the sources. Our previously proposed encoding scheme for the side-information channel in [112] can also support the encoding of location information. Finally, this approach is expected to perform better isolation, as the beamformed signals that take part in the post-filtering stage are all beamformed from the closest array (i.e., with the highest SNR) in contrast to the method of Section 8.2.3.

Table 8.1: DOAs for the source locations used in the listening test with respect to each microphone array

	L_1	L_2	L_3
Mic. array 1	48°	42°	18°
Mic. array 2	154°	119°	140°
Mic. array 3	223°	229°	249°
Mic. array 4	294°	328°	313°

8.2.5 Results

In order to evaluate the source isolation performance of the two methods described in Sections 8.2.3 and 8.2.4 we performed a listening test. The test scenario is depicted in Fig. 8.1 and consists of three simultaneously active sources at locations L_1, L_2, and L_3. In a room of dimensions $10 \times 10 \times 3$ meters there are $M = 4$ circular microphone arrays at locations $(1, 1), (9, 1), (9, 9), (1, 9)$ meters. Each microphone array has a radius of 2 cm and consists of $N = 4$ omnidirectional microphones. The DOAs of the sources at the three locations with respect to the 4 microphone arrays are shown in Table 8.1. Note that the sources are located close together in terms of angular separation with respect to all arrays making the source isolation problem quite challenging.

We used the Image-Source Method [91] to produce simulated signals of omnidirectional sources, sampled at 44.1 kHz, in a room with reverberation time $T_{60} = 0.4$ seconds. The signals were processed using frames of 2048 samples with 50% overlap, windowed with a von Hann window. The FFT size was 4096. The approaches of Sections 8.2.3 and 8.2.4 were used in order to isolate the three source signals. The experiment was repeated 6 times with different speakers at locations L_1, L_2, and L_3 (Fig. 8.1), resulting in 18 isolated source signals for each method. We employed a preference test, where listeners used headphones to listen to the reverberant source signal of the target source and the output of the two methods (Section 8.2.3 and Section 8.2.4) and they were asked to indicate which method of the two they preferred in terms of speech quality, intelligibility, and source isolation (always comparing to the original reverberant source). The samples were randomized and the subjects did not know to which method they belonged. Eleven volunteers participated in the listening test.

Fig. 8.2 shows the percentage of listeners that preferred the beamforming with cooperative post-filtering approach of Section 8.2.4 for each location. It is clear that this approach outperforms the method of Section 8.2.3. The cooperative post-filtering approach results in better source isolation and maintains better speech quality and intelligibility, while keeping all the attractive properties for downmixing into a single audio signal and being computationally efficient (the same number of beamforming operations as in the standard ImmACS with one array is required). These results are encouraging for further investigation and analysis of the

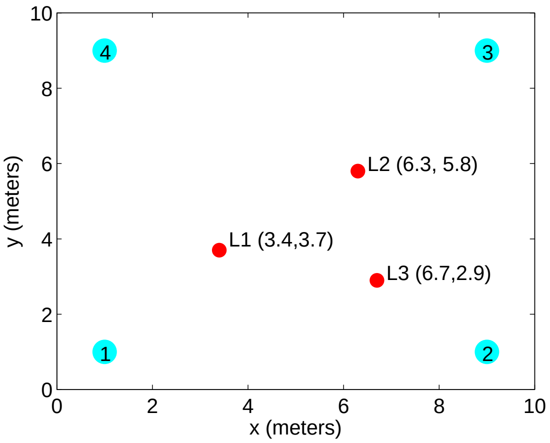


Figure 8.1: Microphone array placement and source locations used for the listening test

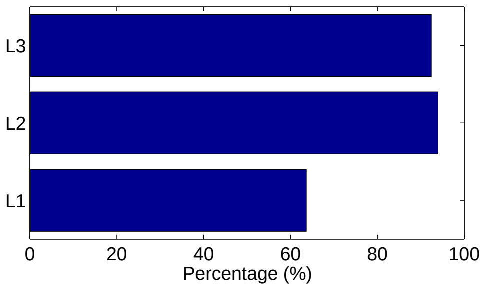


Figure 8.2: Preference test results that indicate the percentage of listeners that preferred the method of Section 8.2.4 over the method of Section 8.2.3 of the three test locations

performance improvement of this method in various scenarios.

8.3 MVDR beamforming with location-based steering vector estimation

In this section, we present a second example of how location estimates can be used to estimate the steering vector which represents the target speaker direction. Accurate estimation of the steering vector is crucial for effective beamforming. In conventional beamformers, the estimation of the steering vector relies on the knowledge of the source's DOA, the array geometry and the plane wave propagation assumption. As an example, the beamformer in the CHiME-3 challenge [110] first estimates the direction of the speaker and then estimates the steering vector based on the speaker's direction and the plane wave propagation assumption.

More powerful schemes for steering vector estimation rely on time-frequency mask estimation that do not require any knowledge of the array geometry, the source direction or the plane wave assumption. These masks can be thought of as the probabilities of source presence in each time-frequency bin. Based on these masks, the target signal statistics can be estimated in order to construct a beamformer to enhance the target source. This approach to beamforming has been shown to produce better results, especially for the task of speech enhancement for robust speech recognition [111, 116]. The most common methods to estimate the time-frequency masks are based on modeling the Fourier coefficients of the microphone array signals using complex Gaussian Mixture models [117, 118] or Watson mixture models [104] or by utilizing Deep Neural Networks [119]. In this section, we propose a methodology to estimate the time-frequency masks based on the per-frequency location estimates obtained by fusing per-frequency DOA estimates from the microphone arrays in a WASN. We combine the location-based mask estimation with an MVDR beamformer and we show that this scheme can reduce the interfering sources and enhance the signal of the target source in a cocktail-party situation.

8.3.1 Problem formulation

Assume that K sources are active in a WASN with M microphone arrays each of which consists of N_m microphones, $m = 1, \ldots, M$. Let N denote the total number of microphones in the network such that

$$N = \sum_{m=1}^{M} N_m. \tag{8.6}$$

The signal at the n-th microphone can be represented in the STFT domain as:

$$Y_n(\tau, \ell) = \sum_{k=1}^{K} H_{k,n}(\ell) S_k(\tau, \ell) + E_n(\tau, \ell) \tag{8.7}$$

where τ,ℓ denote the time-frame and frequency index respectively, $H_{k,n}(\ell)$ denote the impulse response of the acoustic path from the k-th source to the n-th microphone, $S_k(\tau, \ell)$ is the STFT representation of the k-th source's signal and $E_n(\tau, \ell)$ is the STFT representation of the noise at the n-th microphone. Introducing vector notation, (8.7) can be written as:

$$\boldsymbol{Y}(\tau, \ell) = \sum_{k=1}^{K} \boldsymbol{H}_k(\ell) S_k(\tau, \ell) + \boldsymbol{E}(\tau, \ell) \tag{8.8}$$

where

$$\boldsymbol{Y}(\tau, \ell) = \Big[Y_1(\tau, \ell), \quad \ldots \quad Y_N(\tau, \ell) \Big]^T, \tag{8.9}$$

$$\boldsymbol{H}_k(\ell) = \Big[H_{k,1}(\ell), \quad \ldots \quad H_{k,N}(\ell) \Big]^T, \tag{8.10}$$

$$\boldsymbol{E}(\tau, \ell) = \Big[E_1(\tau, \ell), \quad \dots \quad E_N(\tau, \ell)\Big]^T, \tag{8.11}$$

The goal is to extract the source signal $S_k(\tau, \ell)$ from the observed signals $\boldsymbol{Y}(\tau, \ell)$.

8.3.2 MVDR beamforming

The beamformer for the k-th source is a linear filter $\boldsymbol{w}_k(\ell)$ that when applied to the microphone signal vector, it enhances the signal of the k-th source $S_k(\tau, \ell)$:

$$\hat{S}_k(\tau, \ell) = \boldsymbol{w}_k^H(\ell)\boldsymbol{Y}(\tau, \ell) \tag{8.12}$$

where $\boldsymbol{w}_k^H(\ell)$ is the $N \times 1$ vector of beamformer filter coefficients for time-frame τ and frequency bin ℓ. The filter coefficients are calculated as:

$$\boldsymbol{w}_k(\ell) = \frac{\boldsymbol{R}_y^{-1}(\ell)\boldsymbol{H}_k(\ell)}{\boldsymbol{H}_k^H(\ell)\boldsymbol{R}_y^{-1}(\ell)\boldsymbol{H}_k(\ell)} \tag{8.13}$$

where $\boldsymbol{H}_k(\ell)$ is the steering vector for the k-th source and $\boldsymbol{R}_y(\ell)$ is the $N \times N$ spatial correlation matrix of the microphone signals which is calculated as:

$$\boldsymbol{R}_y(\ell) = \frac{1}{T}\sum_{\tau} \boldsymbol{Y}(\tau, \ell)\boldsymbol{Y}^H(\tau, \ell), \tag{8.14}$$

where T denotes the total number of time-frames.

8.3.3 Location-based steering vector estimation

Following the approach of [117, 118] the steering vector is estimated as the principal eigenvector of the spatial correlation matrix of the target speech signal. The spatial correlation matrix can be estimated using time-frequency masks that describe the presence of the target source at each time-frequency bin. Let $\mathcal{M}_k(\tau, \ell)$ denote the time-frequency mask for the k-th source, then the $N \times N$ spatial correlation matrix of the k-th source's signal, $\boldsymbol{R}_k(\ell)$ can be calculated as:

$$\boldsymbol{R}_k(\ell) = \frac{1}{\sum_\tau \mathcal{M}_k(\tau, \ell)} \sum_{\tau} \mathcal{M}_k(\tau, \ell)\boldsymbol{Y}(\tau, \ell)\boldsymbol{Y}^H(\tau, \ell), \tag{8.15}$$

We show that these time-frequency masks can be estimated with the use of per-frequency location estimates, similar to the ones used in Chapter 6 for location estimation and counting. In this scheme, each microphone array estimates and transmits narrowband DOA estimates for each frequency bin and each time-frame. The narrowband DOA estimates are received at the fusion center where the single-source grid-based method is applied in order to infer a location estimate, resulting in the set $L(\tau, \ell)$ of location estimates for each time-frame τ and frequency bin ℓ. In the next step, the K-means algorithm is applied on the set of location estimates $L(\tau, \ell)$ in order to cluster them into K clusters, where K is the number of sources. Based on this clustering, a binary mask $\mathcal{M}_k(\tau, \ell)$ can be computed for each source, based on whether the location estimate at (τ, ℓ) was assigned to source k. Alternatively, a thresholding operation, similar to the outlier rejection scheme of Section 6.3.2 can be applied on the location estimates, prior to K-means clustering, in order to remove noisy location estimates. The location estimates that are removed by this operation are not assigned to any of the sources, i.e., the masks of all sources at the corresponding time-frequency bins are zero.

8.3.4 Results

In order to test the potential of the location-based steering vector estimation for beamforming we considered two example scenarios with three and four simultaneously active sound sources. We employed a WASN with $M = 4$ nodes placed on the corners of a square cell of dimensions of $V = 4$ meters. Each node was an 8-element uniform circular microphone array with 5 cm radius. The sound sources were speech recordings of 3 seconds duration, downsampled at 16 kHz, and had equal power when placed at the center of the cell. The signal-to-noise ratio (SNR) was measured as the ratio of the power of each source signal when located at the center of the cell to the power of the noise signal.

We simulated a room of dimensions of $10 \times 10 \times 3$ meters using the Image-Source method [91] and produced signals of omnidirectional sources at reverberation time of $T_{60} =$ 400 ms. The WASN cell was placed in the middle of the room. Both the nodes and the sources were placed at 1 m. height. More specifically, the nodes were placed at $(3, 3, 1)$, $(7, 3, 1)$, $(7, 7, 1)$, and $(3, 7, 1)$. For processing we used frames of 512 samples with 50% overlap. The FFT size was 512. For narrowband DOA estimation we used the method proposed in [4], while the per-frequency locations are estimated using the single source grid-based method of Chapter 4.

In the first simulation, we considered three simultaneously active sound sources at 40 dB SNR. The locations of the sources are shown in Fig. 8.3. For the mask estimation we used the two approaches described in Section 8.3.3, i.e., processing the entire set of per-frequency location estimates (denoted as *location-based mask estimation*) and applying a thresholding operation similar to the one described in Section 6.3.2 prior to K-means clustering in order to remove noisy location estimates (denoted as *location-based mask estimation with thresholding*). For the outlier rejection scheme we set $q = 0.35$ (see Section 6.3.2).

Fig. 8.4 depicts the estimated per-frequency locations and the assignment of the location estimates to the sources using K-means. The resulted time-frequency masks for each source are shown in Fig. 8.5. Similarly, Fig. 8.6 depicts the location estimates that remained after the thresholding operation and their assignment to the sources using K-means and Fig. 8.7 shows the corresponding time-frequency masks for each source. Comparing the time-frequency

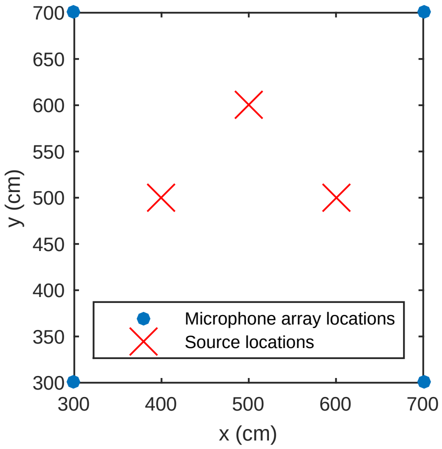
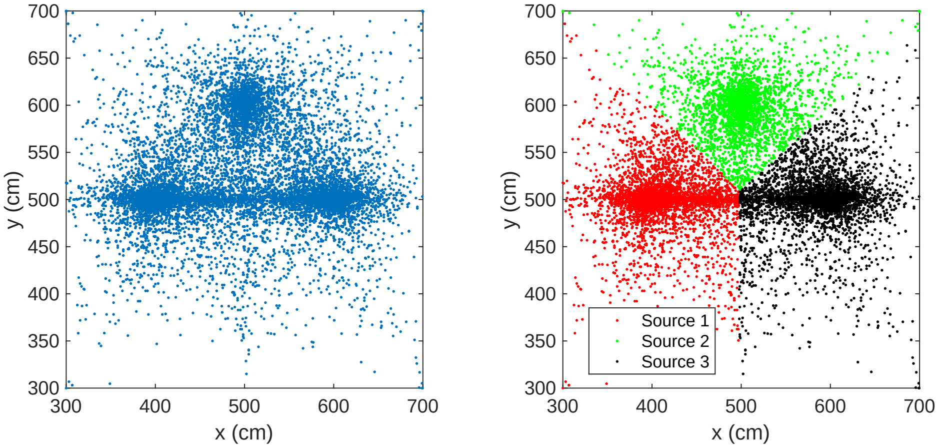


Figure 8.3: Microphone array and source locations for the three sources scenario.



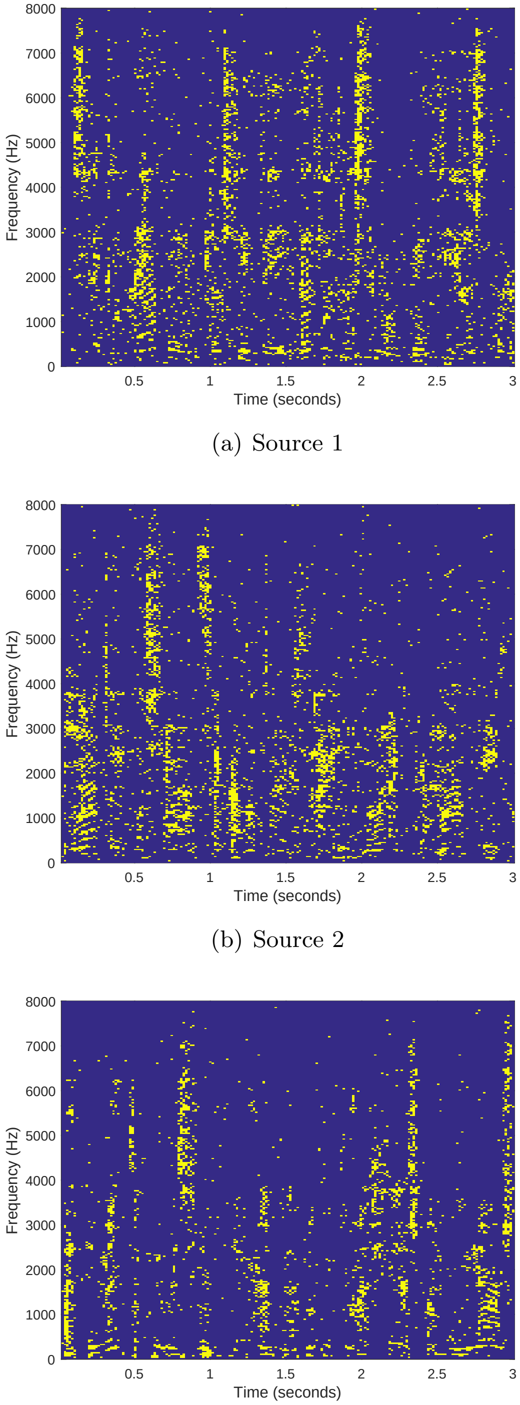
(a)

(b)

Figure 8.4: (a) Per-frequency location estimates and (b) the assignment of location estimates to the sources using K-means.

masks from the two versions of the proposed methodology it is evident that in the location-based mask estimation with thresholding the masks have more zero values which correspond to the rejected frequencies during the thresholding operation that are not assigned to any of the sources.

To evaluate performance, we measured the Signal-to-interference ratio (SIR) before and after the beamforming operation for non-overlapping segments of $T = 32$ ms. The input SIR (before the beamforming operation) at the n-th microphone, for the i-th segment was



(a) Source 1

(b) Source 2

(c) Source 3

Figure 8.5: Time-frequency masks extracted using the location-based mask estimation for the three sources scenario.

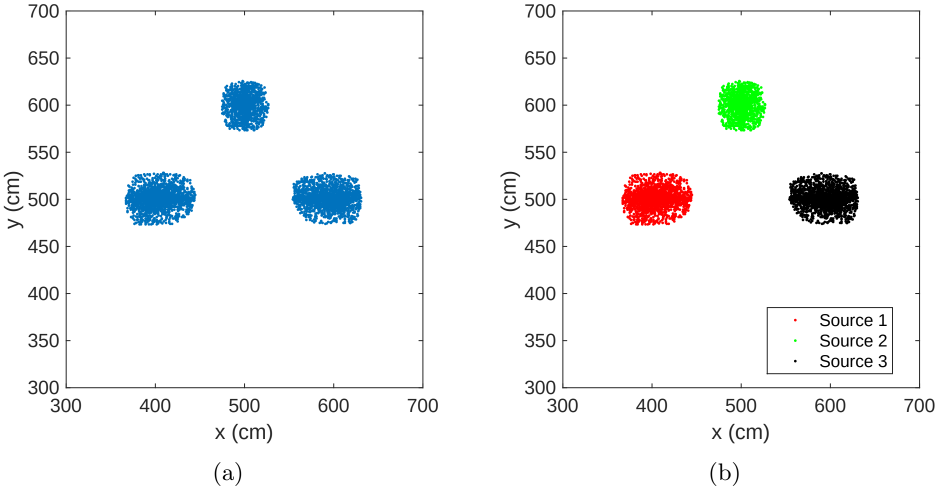


Figure 8.6: (a) Per-frequency location estimates after the thresholding operation with $q = 0.35$ and (b) the assignment of location estimates to the sources using K-means.

computed as:

$$iSIR(i) = 10\log_{10} ||x_n(t)||_2^2/||u_n(t)||_2^2 \tag{8.16}$$

where $x_n(t)$ denotes the target source's signal, $u_n(t)$ denotes the interference sources' signals and $t \in [(i-1)T, iT]$.

The output SIR (after the filtering operation with the beamformer) for the i-th segment was computed as:

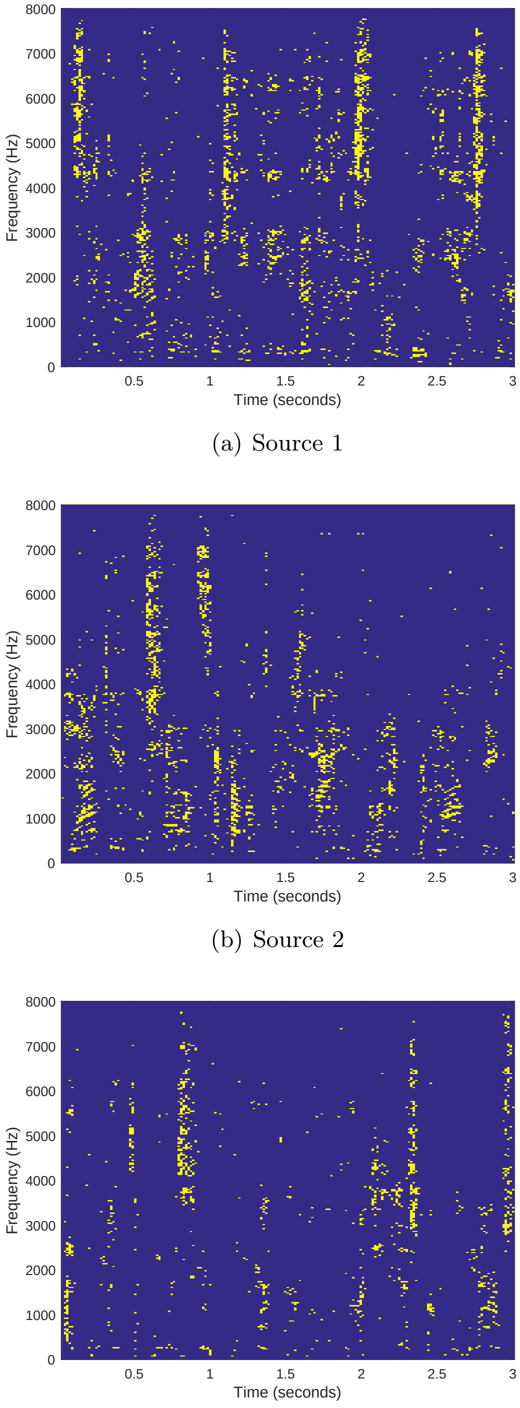
$$oSIR(i) = 10\log_{10} ||\hat{x}(t)||_2^2/||\hat{u}(t)||_2^2 \tag{8.17}$$

where $\hat{x}(t)$ and $\hat{u}(t)$ denotes the filtered version of the target source and interference sources signals. The overall input and output SIR are computed by averaging over all segments i. Finally, the SIR improvement was measured as:

$$\Delta SIR =< oSIR(i) > - < iSIR(i) > \tag{8.18}$$

where $< \cdot >$ denotes averaging over i. As the input SIR may differ significantly among the different microphones in the WASN, we chose the microphone that provides the maximum input SIR, i.e., the microphone that best receives the target source signal.

Fig. 8.8 depicts the SIR improvement for each one of the three simultaneously active sound sources using the two versions of the proposed methodology (location-based mask estimation and location-based mask estimation with thresholding). For comparison, we also included the case where the MVDR beamforming is based on oracle time-frequency masks.



(a) Source 1

(b) Source 2

(c) Source 3

Figure 8.7: Time-frequency masks extracted using the location-based mask estimation with thresholding for the three sources scenario.

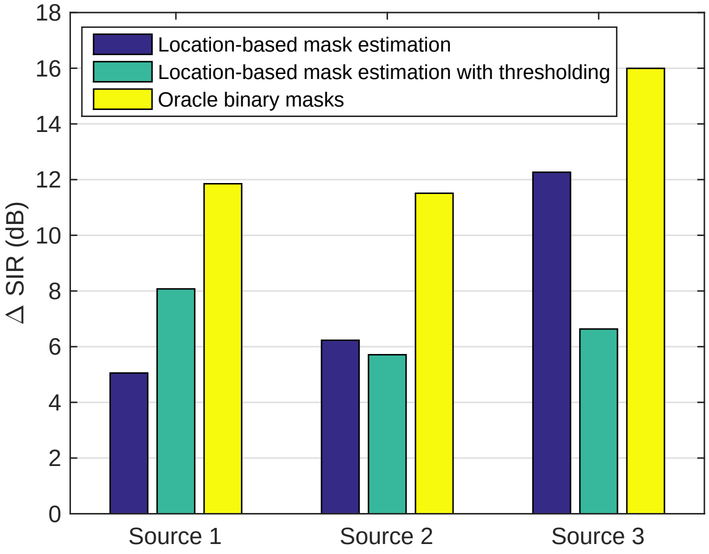


Figure 8.8: SIR improvement for three simultaneously active sound sources at $T_{60} = 400$ ms reverberation time using an MVDR beamformer and different methodologies for time-frequency mask estimation.

We estimate the oracle time-frequency mask for microphone n as:

$$\mathcal{M}_n(\tau,\ell)\begin{cases} 1, & \text{if } |X_n(\tau,\ell)|^2 > |U_n(\tau,\ell)|^2 \\ 0, & \text{otherwise} \end{cases} \tag{8.19}$$

where $X_n(\tau,\ell)$ and $U_n(\tau,\ell)$ is the STFT of the target and interference sources signals at the n-th microphone. In other words, the oracle mask will have the value of one if the energy of the target source is greater than the energy of the interferer sources. The final oracle mask is calculated as the median of the oracle masks over all microphones.

From Fig. 8.8 we can observe that the proposed methodology can achieve satisfying results, indicating the potential of location-based mask estimation for speech enhancement and beamforming. Generally, the methods achieve more than half the performance of the oracle mask estimation. Moreover, there is no version of the proposed methodology that consistently performs the best: the location-based mask estimation with thresholding performs better for the first and second source but significantly worse for the third source.

In the second simulation, we considered four simultaneously active sound sources at 40 dB SNR. The sources' locations are shown in Fig. 8.9 and the corresponding SIR improvement for each one of the four sources is depicted in Fig. 8.10. Again, the proposed location-based mask estimation can achieve more than half the performance of the oracle mask estimation. In

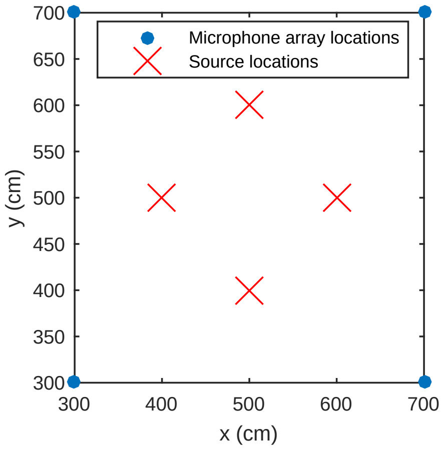


Figure 8.9: Microphone array and source locations for the four sources scenario.

this scenario, the location-based mask estimation without thresholding consistently achieves the best results compared to the mask estimation with thresholding. The discrepancies in performance between the two versions of the proposed methodology indicate that the best location-based mask estimation may lie somewhere in between the two approaches: perhaps in the design of soft-masks. An idea towards this direction is to utilize a gaussian mixture model for clustering and constructing the masks based on the posterior probabilities of each location estimate belonging to each gaussian component. Another important aspect of the proposed methodology is that it does not explicitly take into account the background noise. One approach to take noise into account, is to construct another time-frequency mask that indicates the presence of non-directional noise and to use this mask in order to estimate the noise signal statistics. These two directions of soft-mask estimation and explicit incorporation of noise into the mask estimation procedure are our main focus for future work.

The sounds signals for all the experiments are available for listening at `http://www.csd.uoc.gr/~analexan/loc_based_enhancement.html`.

8.4 Conclusions

This chapter presented some examples of using location information for speech enhancement and separation. The first example dealt with the task of source separation for spatial audio capturing and reproduction. It showed that by utilizing information about the locations of the sources, binary masks can be constructed that serve as post-filters to the output of a superdirective beamformer. These post-filters can provide superior source separation performance compared to an approach where the microphone array which is closer to the source is used. In the same spirit, the second example uses location information in order to construct time-frequency masks for estimating the steering vector and apply an MVDR

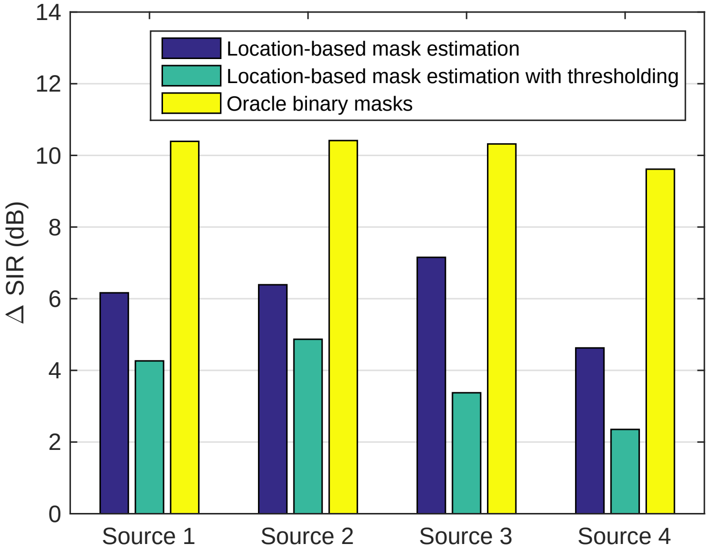


Figure 8.10: SIR improvement for four simultaneously active sound sources at $T_{60} = 400$ ms reverberation time using an MVDR beamformer and different methodologies for time-frequency mask estimation.

beamformer to enhance the signal of a target source. Through simulations on a scenario with three and four simultaneously active sound sources we confirmed that the method can achieve satisfactory performance achieving more than have the performance of ideal binary masks for steering vector estimation. These preliminary results show that location-based methods could be a promising direction for speech enhancement and separation.

Chapter 9
Conclusion

9.1 Synopsis of Contributions

In this thesis, we have studied the problem of location estimation of multiple simultaneously active sound sources using a Wireless Acoustic Sensor Network. Our main focus was to develop localization methods that not only achieve high accuracy, but also attain low communication bandwidth, tolerate unsynchronized input and are computationally efficient, facilitating their implementation in real-life practical WASNs.

We proposed the Grid-based method, a computationally efficient non-linear least squares estimator, for the localization of a single and multiple sources. We identified, defined, and proposed solutions to the core problem that arises in DOA-based approaches when multiple sources are active, which is known as the data-association problem. Our solution is based on additional information transmitted by the sensors, apart from the DOA estimates. We considered the data-association problem in realistic scenarios and showed that our solution is robust to missed detections, reverberation, noise, and moving sources. Keeping in mind the requirements for low communication bandwidth, we incorporated our method with a scheme to attain the bitrate needs of the additional information at low levels. We also proposed a method for the joint problem of source counting and location estimation which is based on statistical modelling of narrowband per-frequency location estimates which are inferred using narrowband per-frequency DOA estimates transmitted by the sensors. Our method was implemented in real-time and tested on a real two-node WASN[1] where the sensors consist of digital MEMS microphone arrays constructed at FORTH-ICS, showing the potential of our approach to be used in real-life applications.

Since the error in the acquired DOA estimates is a determinant factor on the final localization performance, we proposed a methodology to improve the accuracy of the narrowband DOA estimation procedure. Our proposed methodology can be combined with any narrowband DOA estimation method and any array geometry and results in more reliable DOA estimates. Apart from the localization task, the proposed methodology can be advantageous

[1] A video demonstration is available at `https://www.youtube.com/watch?v=bP1aF21BNZ4`

for any task involving the use of narrowband DOA estimates.

Finally, we investigated the use of location information to speech enhancement and separation. We presented preliminary results on how location information can be used to design beamformers and post-filters and we showed that such approaches can achieve satisfactory performance, revealing the potential of location information to improve speech enhancement and separation performance.

9.2 Directions for Future Work and Research

There are several aspects that are worth further work and research:

- **Machine learning-based approaches:** Many of the approaches throughout this dissertation, such as the association algorithm presented in Chapter 5 and the source counting method of Chapter 6 could benefit from the application of machine learning methodologies in order to improve accuracy. For example, better modelling methodologies or model selection algorithms applied on the per-frequency location estimates could improve the source counting accuracy. One direction is to extend the Bayesian K-means algorithm used for source counting and location estimation to other types of mixture models. The Bayesian K-means algorithm assumes that the data have been generated by a mixture of Gaussians, however the Gaussian distribution is known to be sensitive to outliers. Such outliers in the per-frequency location estimates can thus degrade the performance of the algorithm and as a consequence the performance of source counting and location estimation. To make the algorithm more robust to such outliers, other types of distributions can be considered. One possible candidate is the Student-t distribution which is more heavy-tailed and thus more robust to outliers. Mixtures of Student-t distributions have been used in image processing to derive models that are more robust to outlier compared to gaussian mixture models [120].

  Another direction is to utilize deep neural networks (DNNs) for source counting and location estimation. Recently, approaches for single-source DOA estimation have been proposed that are based on deep neural networks. In [121] the Generalized Cross-Correlation vectors are extracted and used as input to train a neural network for DOA estimation, while in [122] the phase information of the captured signals is directly used to train a convolutional neural network (CNN). This approach was later extended to estimate the DOAs of two active sound sources [123].

  Based on our proposed method of source counting and location estimation a deep neural network can be trained on the 2-dimensional histograms of the per-frequency location estimates. At run-time the network will output the number of sources and locations, given as input a two-dimensional histogram of per-frequency location estimates similar to the one presented in the example of Figure 6.1. During training each histogram can be treated as an image, thus a reasonable choice for the network would be to use a convolutional neural network. A possible approach for training the network would be to generate signals in various simulated environments with different conditions (in terms of number of sources, source's locations, reverberation, and noise), construct the two-dimensional histograms of per-frequency location estimates and feed the network with the histograms along with the true number of sources and source's locations for each histogram.

- **Heterogeneity and scalability:** Depending on the specific application, the WASN that needs to be deployed can vary from a small network with a few nodes to a very large network with hundreds of nodes. Also, the network may consist of different types of sensors, such as microphone arrays of different geometries that operate on different sampling frequencies or a combination of microphone and microphone array nodes. The research question that naturally arises is how can we adapt or design localization methods that can easily scale to complex WASNs. One possible direction is the development of distributed approaches that do not rely on the existence of a single fusion center to perform the localization task.

- **Benchmarking:** In terms of evaluating the performance of a localization method there are no consistent benchmarking methodologies and standards. Due to their heterogeneity—in terms of sensors, number of sensors and microphones, network topology and so on—works comparing different localization methods are rare in the literature, while most approaches are evaluated using simulations on specific WASN setups. The definition of formal methodologies, metrics, and datasets that include both simulated and real recorded signals is of paramount importance. While this dissertation has made a step towards this vision by gathering and publishing datasets of real recorded signals a universal and systematic approach is still required in order to set up a common evaluation framework

- **Hardware design:** The practical implementation of WASNs requires low-cost and easily deployable sensors consisting of microphone arrays. To this direction, a digital MEMS microphone arrays was designed at FORTH-ICS and was used as an acoustic sensor in the real-time implementation of our source counting and location estimation method (Section 6.6). However, this sensor is still not fully autonomous, as it requires a host PC that provides the power and processing capabilities. A direction for future work will be to design fully autonomous low-cost microphone arrays that can carry out the capturing, processing and transmission of information.

- **Real-life application:** To facilitate the application of the localization methods in real systems, the integration of methodologies from a diverse range of scientific fields is of paramount importance. Such fields include networks (e.g., to design the communication and synchronization protocols), network administration (e.g., to organize the nodes of the network, the network topology that in practice can be constantly changing and to handle potential node or network failure), signal processing (e.g., for the sound source localization), hardware design (e.g., for the design and maintenance of the acoustic nodes), and so on. While many of these fields have flourished individually, the practical issues that will arise from their integration in practical WASNs still remain unseen and the need for interdisciplinary research is becoming more and more urgent.

- **Location-based speech enhancement:** The preliminary results that we presented show the potential of using location features for speech enhancement and separation. We believe that the use of location features can be a new direction of research for speech enhancement, separation and robust speech recognition that can potentially outperform the current state-of-the-art. However, more extensive research is needed in order to develop and test such approaches in a variety of scenarios (in terms of number of sources, acoustic conditions, etc.) and discover their limits.

  A first step towards this direction would be to compare the location-based speech enhancement with other approaches. Current state-of-the-art methodologies for time-frequency mask estimation include the modelling of the Fourier coefficients of the captured signals with complex Gaussian mixture models [117, 118], mixtures of Watson distributions [104] or utilizing deep neural networks [119]. For the evaluation, metrics for source separation and enhancement, such as signal-to-noise enhancement, signal-to-interference enhancement, and speech distortion, as well as speech recognition metrics, such as the word error rate can be utilize to compare performance between different methodologies.

  Another potential direction in order to further improve the performance of location-based mask estimation for beamforming, is to change the binary nature of the time-frequency masks and construct soft masks that can take all possible values in the range or 0 to 1. A possible way to do that is to use other clustering algorithms, such as clustering the location estimates using gaussian mixture models of fuzzy K-means. Then, each location estimate will be assigned to a source with a certain probability and the time-frequency masks could be constructed based on these probabilities. For beamforming, other schemes apart from the MVDR beamformer could be utilized. Another promising beamformer to use could be the Generalized Eigenvalue (GEV) beamformer that maximizes the output SNR at each time-frequency bin [124].

Bibliography

[1] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[2] J. Tiete, F. Domínguez, B. Silva, L. Segers, K. Steenhaut, and A. Touhafi, "SoundCompass: A Distributed MEMS Microphone Array-Based Sensor for Sound Source Localization," *Sensors*, vol. 14, no. 2, pp. 1918–1949, 2014.

[3] A. N. Bishop and P. N. Pathirana, "A discussion on passive location discovery in emitter networks using angle-only measurements," in *International Conference on Wireless Communications and Mobile Computing.* 2006, ACM.

[4] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," in *European Signal Processing Conference (EUSIPCO)*, 2007, pp. 778–782.

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

[6] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux*, 2011, pp. 1–6.

[7] D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirschel, "Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus," *Journal of Applied Ecology*, vol. 48, no. 3, pp. 758–767, 2011.

[8] D. J. Mennill, M. Battiston, D. R. Wilson, J. R. Foote, and S. M. Doucet, "Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour," *Methods in Ecology and Evolution*, vol. 3, no. 4, pp. 704–712, 2012.

[9] P. M. Stepanian, K. G. Horton, D. C. Hille, C. E. Wainwright, P. B. Chilson, and J. F. Kelly, "Extending bioacoustic monitoring of birds aloft through flight call localization

with a three-dimensional microphone array," *Ecology and Evolution*, vol. 6, no. 19, pp. 7039–7046, 2016.

[10] M. Taseska and E. A. P. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, July 2014.

[11] A. Boukerche, H.A.B. Oliveira, E.F. Nakamura, and A.A.F. Loureiro, "Localization systems for wireless sensor networks," *IEEE Wireless Communications*, vol. 14, no. 6, pp. 6–12, 2007.

[12] G. Mao, B. Fidan, and B. Anderson, "Wireless sensor network localization techniques," *Computer Networks*, vol. 51, no. 10, pp. 2529 – 2553, 2007.

[13] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, (IROS)*, Oct 2007, pp. 2009–2014.

[14] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 954–995, 1989.

[15] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 246–260, Jan 2012.

[16] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Aug. 2012, pp. 2303–2307.

[17] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[18] D. Pavlidi, A. Griffin, M. Puigt, 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, pp. 2193–2206, 2013.

[19] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, August 1976.

[20] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, 2012.

[21] D. Bechler and K. Kroschel, "Three different reliability criteria for time delay estimates," in *European Signal Processing Conference (EUSIPCO)*, Sept 2004, pp. 1987–1990.

[22] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005, vol. 4, pp. iv/985–iv/988 Vol. 4.

[23] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1479–1489, Nov 2008.

[24] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1563–1575, Oct 2015.

[25] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 8, pp. 1223–1225, Aug 1987.

[26] J.O. Smith and J.S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, Dec 1987.

[27] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, Nov 2001.

[28] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1770–1778, May 2008.

[29] L.E. Kinsler, et al. , *Fundamentals of acoustics*, NY: Wiley, 2000.

[30] X. Sheng and Y. Hu, "Energy based acoustic source localization," in *Proceedings of IPSN*, 2003, pp. 285–300.

[31] X. Sheng and Y. Hu, "Sequential acoustic energy based source localization using particle filter in a distributed sensor network," in *IEEE International Conference on*

Acoustics, Speech, and Signal Processing (ICASSP), May 2004, vol. 3, pp. iii–972–5 vol.3.

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

[33] D. Li and Y. Hu, "Energy-based collaborative source localization using acoustic microsensor array.," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 4, pp. 321–337, 2003.

[34] W. Meng and W. Xiao, "Energy-based acoustic source localization methods: a survey," *Sensors*, vol. 17, no. 2, 2017.

[35] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 338–347, Jan. 2003.

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

[37] A. Abad, C. Segura, D. Macho, J. Hernando, and C. Nadeu, "Audio person tracking in a smart-room environment.," in *INTERSPEECH*. 2006, ISCA.

[38] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2008, pp. 69–72.

[39] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 11:1–11:17, Jan. 2010.

[40] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, vol. 4, pp. IV–493–IV–496.

[41] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field.," in *INTERSPEECH*. 2006, ISCA.

[42] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007, vol. 1, pp. I–121–I–124.

[43] H. Do and H.F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation using Coarse-To-Fine Region Contraction(CFRC)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 295–298.

[44] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, Jan 2011.

[45] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," *Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2627–2630, Oct 2013.

[46] M. Taseska, A. H. Khan, and E. A. P. Habets, "Speech enhancement with a low-complexity online source number estimator using distributed arrays," in *22nd European Signal Processing Conference (EUSIPCO)*, Sept 2014, pp. 929–933.

[47] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, Feb 2014.

[48] Y. Dorfan, G. Hazan, and S. Gannot, "Multiple acoustic sources localization using distributed expectation-maximization algorithm," in *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 72–76.

[49] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, Oct 2015.

[50] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 253–256.

[51] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6834–6838.

[52] L. Drude, A. Chinaev, D. H. Tran Vu, and R. Haeb-Umbach, "Towards online source counting in speech mixtures applying a variational EM for complex watson mixture models," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2014, pp. 213–217.

[53] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 33–36.

[54] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric bayesian estimation of an infinite gaussian mixture model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 459–463.

[55] A. Lédeczi, G. Kiss, B. Fehér, P. Völgyesi, and G. Balogh, "Acoustic source localization fusing sparse direction of arrival estimates," in *International Workshop on Intelligent Solutions in Embedded Systems*, June 2006, pp. 1–13.

[56] J.C. Chen, L. Yip, J. Elson, H. Wang, D. Maniezzo, R. E. Hudson, K. Yao, and D. Estrin, "Coherent acoustic array processing and localization on wireless sensor networks," in *Proceedings of the IEEE*, 2003, pp. 1154–1162.

[57] A. M. Ali, S. Asgari, T. Collier, M. Allen, L. Girod, R. Hudson, K. Yao, C. Taylor, and D. Blumstein, "An empirical study of collaborative acoustic source localization," *Journal of Signal Processing Systems*, vol. 57, no. 3, pp. 415–436, 2009.

[58] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, 1993.

[59] L. Girod, M. Lukac, V. Trifa, and D. Estrin, "The design and implementation of a self-calibrating distributed acoustic sensing platform," in *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems*, New York, NY, USA, 2006, SenSys '06, pp. 71–84, ACM.

[60] L. M. Kaplan and Q. Le, "On exploiting propagation delays for passive target localization using bearings-only measurements," *Journal of the Franklin Institute*, vol. 342, no. 2, pp. 193–211, 2005.

[61] A. N. Bishop, B. D. O. Anderson, B. Fidan, P. N. Pathirana, and G. Mao, "Bearing-only localization using geometrically constrained optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 1, pp. 308–320, 2009.

[62] Z. Wang, J. Luo, and X. Zhang, "A novel location-penalized maximum likelihood estimator for bearing-only target localization," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6166–6181, 2012.

[63] R. G. Stansfield, "Statistical theory of D.F. fixing," *Journal of the Institute of Electrical Engineering - Part IIIA: Radiocommunication*, vol. 94, no. 15, pp. 762–770, 1947.

[64] K. Doğançay, "Bearings-only target localization using total least squares," *Signal Processing*, vol. 85, no. 9, pp. 1695–1710, 2005.

[65] K. Doğançay, "On the bias of linear least squares algorithms for passive target localization," *Signal Processing*, vol. 84, no. 3, pp. 475 – 486, 2004.

[66] K. Doğançay, "Bias compensation for the bearings-only pseudolinear target track estimator," *IEEE Transactions on Signal Processing*, vol. 54, no. 1, pp. 59–68, Jan 2006.

[67] M. Gavish and A. J. Weiss, "Performance analysis of bearing-only target location algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 3, pp. 817–828, 1992.

[68] K. Doğançay, "Passive emitter localization using weighted instrumental variables," *Signal Processing*, vol. 84, no. 3, pp. 487 – 497, 2004.

[69] Y. T. Chan and S. W. Rudnicki, "Bearings-only and doppler-bearing tracking using instrumental variables," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 4, pp. 1076–1083, Oct 1992.

[70] K. Doğançay, "Reducing the bias of a bearings-only TLS target location estimator through geometry translations," in *European Signal Processing Conference (EUSIPCO)*, Sept 2004, pp. 1123–1126.

[71] A. Griffin and A. Mouchtaris, "Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013, pp. 1–4.

[72] Steven M. Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[73] A. Farina, "Target tracking with bearings-only measurements," *Signal Processing*, vol. 78, no. 1, pp. 61–78, 1999.

[74] L. M. Kaplan, P. Molnar, and Q. Le, "Bearings-only target localization for an acoustical unattended ground sensor network," in *Proceedigns of SPIE*, 2001, vol. 4393, pp. 40–51.

[75] K.R. Pattipati, S. Deb, Y. Bar-Shalom, and R. B. Washburn, "A new relaxation algorithm and passive sensor data association," *IEEE Transactions on Automatic Control*, vol. 37, no. 2, pp. 198–213, Feb 1992.

[76] S. Deb, M. Yeddanapudi, K. Pattipati, and Y. Bar-Shalom, "A generalized S-D assignment algorithm for multisensor-multitarget state estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 2, pp. 523–538, April 1997.

[77] R.L. Popp, K.R. Pattipati, and Y. Bar-Shalom, “m-Best S-D assignment algorithm with application to multitarget tracking,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 22–39, Jan 2001.

[78] A.N. Bishop and P.N. Pathirana, “Localization of emitters via the intersection of bearing lines: A ghost elimination approach,” *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 3106–3110, Sep 2007.

[79] J.D. Reed, C.R.C.M. da Silva, and R.M. Buehrer, “Multiple-source localization using line-of-bearing measurements: Approaches to the data association problem,” in *IEEE Military Communications Conference (MILCOM)*, Nov 2008, pp. 1–7.

[80] M. Swartling, M. Nilsson, and N. Grbic, “Distinguishing true and false source locations when locating multiple concurrent speech sources,” in *5th IEEE Sensor Array and Multichannel Signal Processing Workshop, (SAM)*, Jul 2008, pp. 361–364.

[81] M. Swartling, N. Grbić, and I. Claesson, “Source localization for multiple speech sources using low complexity non-parametric source separation and clustering,” *Signal Processing*, vol. 91, no. 8, pp. 1781–1788, 2011.

[82] M. Swartling, N. Grbic, and I. Claesson, “Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, May 2006, vol. 4, pp. IV–IV.

[83] S. Schulz and T. Herfet, “On the window-disjoint-orthogonality of speech source in reverberant humanoid scenarios,” in *Proc. of DAFx–08*, 2008, pp. 241–248.

[84] S. T. Roweis, “Factorial models and refiltering for speech separation and denoising,” in *Proceedings of the European Conference on Speech Communication (EUROSPEECH)*, 2003.

[85] A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, and A. Mouchtaris, “Localizing multiple audio sources in a wireless acoustic sensor network,” *Signal Processing, Special Issue on wireless acoustic sensor networks and ad hoc microphone arrays*, vol. 107, pp. 54 – 67, 2014.

[86] A. Griffin, A. Alexandridis, D. Pavlidi, and A. Mouchtaris, “Real-time localization of multiple audio sources in a wireless acoustic sensor network,” in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Sep 2014, pp. 306–310.

[87] Malcolm J. Crocker, *Handbook of Acoustics*, Wiley, 1998.

[88] A. Alexandridis, G. Borboudakis, and A. Mouchtaris, "Addressing the data-association problem for multiple sound source localization using DOA estimates," in *European Signal Processing (EUSIPCO)*, 2015, pp. 1576–1580.

[89] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation in wireless acoustic sensor networks using doa estimates: The data-association problem," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 342–356, Feb 2018.

[90] G. Carpaneto and P. Toth, "Algorithm for the solution of the bottleneck assignment problem," *Computing*, vol. 27, no. 2, pp. 179–187, 1981.

[91] 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*, vol. 18, no. 6, Aug. 2010.

[92] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, Academic Press, 4th edition, 2008.

[93] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

[94] K. Kurihara and M. Welling, "Bayesian k-means as a "maximization-expectation" algorithm," *Neural Computation*, vol. 21, no. 4, pp. 1145–1172, Apr 2009.

[95] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002, vol. 1, pp. I–529–I–532.

[96] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "SMEM Algorithm for Mixture Models," *Neural Computation*, vol. 12, no. 9, pp. 2109–2128, Sep 2000.

[97] N. Stefanakis and A. Mouchtaris, "Direction of arrival estimation in front of a reflective plane using a circular microphone array," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[98] O. Thiergart, M. Taseska, and E.A.P. Habets, "An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates," in *European Signal Processing Conference (EUSIPCO)*, Sept 2013, pp. 1–5.

[99] O. Thiergart, M. Taseska, and E.A.P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2182–2196, Dec 2014.

[100] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, 2007.

[101] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *HSCMA, 2008.*, May 2008, pp. 37–40.

[102] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 2:1–2:13, 2010.

[103] M. Kallinger, G. Del Galdo, F. Kuech, and O. Thiergart, "Dereverberation in the spatial audio coding domain," in *Audio Engineering Society Convention 130*, May 2011.

[104] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2010, pp. 241–244.

[105] L. Drude, F. Jacob, and R. Haeb-Umbach, "DOA-estimation based on a complex watson kernel method," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 255–259.

[106] A. Alexandridis and A. Mouchtaris, "Improving narrowband DOA estimation of sound sources using the complex watson distribution," in *European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 1468–1472.

[107] K. V. Mardia and I. L. Dryden, "The complex watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[108] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST Handbook of Mathematical Functions*, Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[109] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, Mar 1986.

[110] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 504–511.

[111] T. Menne, J. Heymann, A. Alexandridis, Irie K., Zeyer A., M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schluter, H. Ney, R. Haeb-Umbach, and A. Mouchtaris, "The

RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *The 4th International Workshop on Speech Processing in Everyday Environments*, 2016, pp. 39–44.

[112] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Directional coding of audio using a circular microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 296–300.

[113] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Capturing and Reproducing Spatial Audio Based on a Circular Microphone Array," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[114] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Breaking down the coctail-party: capturing and isolating sources in a soundscape," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2014.

[115] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[116] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 436–443.

[117] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5210–5214.

[118] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, April 2017.

[119] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[120] G. Sfikas, C. Nikou, and N. Galatsanos, "Robust image segmentation with mixtures of student's t-distributions," in *IEEE International Conference on Image Processing*, Sept 2007, vol. 1, pp. I – 273–I – 276.

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

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

[123] S. Chakrabarty and E.A.P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," in *Proc. of the Machine Learning for Audio Signal Processing Workshop at the Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[124] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.