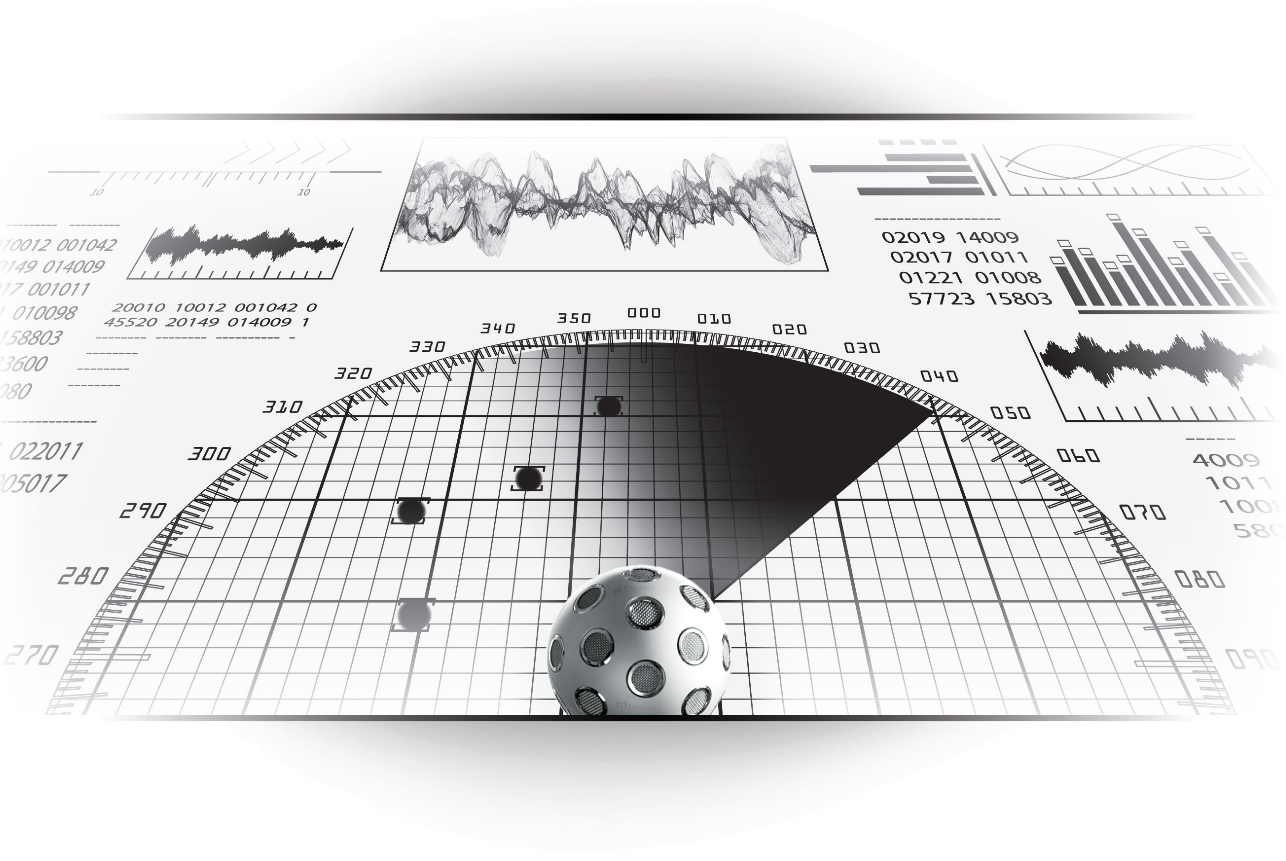
Multiple Source Localization using Spherical Microphone Arrays

by
Sina Hafezi



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

Communications & Signal Processing Group
Department of Electrical & Electronic Engineering
Imperial College London
2018

Copyright declaration

Statement of Originality

I declare that this thesis and the research to which it refers are the product of my own work under the guidance and supervision of Dr. Patrick A. Naylor and Dr. Alastair H. Moore. Any ideas or quotations from the work of others, published or otherwise, are fully acknowledged in accordance with standard referencing practice. The material of this thesis has not been accepted for any degree, and has not been concurrently submitted for the award of any other degree.

Abstract

Direction-of-Arrival (DOA) estimation is a fundamental task in acoustic signal processing and is used in source separation, localization, tracking, environment mapping, speech enhancement and dereverberation. In applications such as hearing aids, robot audition, teleconferencing and meeting diarization, the presence of multiple simultaneously active sources often occurs. Therefore DOA estimation which is robust to Multi-Source (MS) scenarios is of particular importance.

In the past decade, interest in Spherical Microphone Arrays (SMAs) has been rapidly grown due to its ability to analyse the sound field with equal resolution in all directions. Such symmetry makes SMAs suitable for applications in robot audition where potential variety of heights and positions of the talkers are expected. Acoustic signal processing for SMAs is often formulated in the Spherical Harmonic Domain (SHD) which describes the sound field in a form that is independent of the geometry of the SMA. DOA estimation methods for the real-world scenarios address one or more performance degrading factors such as noise, reverberation, multi-source activity or tackled problems such as source counting or reducing computational complexity.

This thesis addresses various problems in MS DOA estimation for speech sources each of which focuses on one or more performance degrading factor(s). Firstly a narrowband DOA estimator is proposed utilizing high order spatial information in two computationally efficient ways. Secondly, an autonomous source counting technique is proposed which uses density-based clustering in an evolutionary framework. Thirdly, a confidence metric for validity of Single Source (SS) assumption in a Time-Frequency (TF) bin is proposed. It is based on MS assumption in a short time interval where the number and the TF bin of active sources are adaptively estimated. Finally two analytical narrowband MS

DOA estimators are proposed based on MS assumption in a TF bin.

The proposed methods are evaluated using simulations and real recordings. Each proposed technique outperforms comparative baseline methods and performs at least as accurately as the state-of-the-art.

Acknowledgment

First and foremost, I would like to thank my supervisor, Dr. Patrick A. Naylor for his support, encouragement, optimism and guidance through my entire PhD. I would like to express my appreciation to Dr. Alastair H. Moore for his secondary supervision, support and guidance.

I am thankful to Mr. Mike Brookes and Dr. Christine Evers for their constructive advice and feedback during the research. I also highly appreciate the PhD scholarship provided by the Electrical and Electronic Engineering Department of Imperial College London. In addition, I would like to thank my friends and colleagues in the Communication and Signal Processing group including Costas, Hamza and Constantinos for all the good times.

Finally, my sincere appreciation goes to my family including my lovely parents Mohammad Ali and Nassrin, my sweetheart sister Zoha, and my dearest brother Adib for their support during the passed years of research.

Sina Hafezi

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

AIC	Akaike Information Criterion
AIR	Acoustic Impulse Response
AIV	Augmented Intensity Vectors
ANOVA	Analysis Of Variance
BIC	Baysian Information Criterion
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DIV	Dual Intensity Vector
DOA	Direction Of Arrival
DPD	Direct Path Dominance
DPDc-MUSIC	Direct Path Dominance with Coherent MUSIC
DPDi-MUSIC	Direct Path Dominance with Incoherent MUSIC
EC	Estimation Consistency
EVD	Eigenvalue Decomposition
FS	Frequency Smoothing
GD	Gradient Descent
GS	Grid Search
IC	Information Criterion
ML	Maximum Likelihood
MS	Multi-Source
MS PIV	Multi-Source PIV
MS MUSIC	Multi-Source MUSIC
MSEC	Multi-Source Estimation Consistency

MUSIC	Multiple Signal Classification
NoDS	Number of Detected Sources
PCA	Principal Component Analysis
PIV	Pseudo-intensity Vector
PWD	Plane Wave Decomposition
RSS	Residual Sum of Squares
RT	Reverberation Time
SH	Spherical Harmonic
SHD	Spherical Harmonic Domain
SHT	Spherical Harmonic Transform
SIR	Signal to Interference Ratio
SMA	Spherical Microphone Array
SMIRgen	Spherical Microphone array Impulse Response generator
SNR	Signal to Noise Ratio
SR	Sources Ratio
SRP	Steered Response Power
SS	Single Source
SS PIV	Subspace PIV
SS DIV	Subspace DIV
SSZ	Single Source Zone
STFT	Short-Time Fourier Transform
SVD	Singular Value Decomposition
SVR	Singular Value Ratio
TF	Time-Frequency
WDO	W-Disjoint Orthogonality

Operators

$(.)^*$	Complex conjugate
$(.)'$	Derivative
$(.)^T$	Vector/matrix transpose
$(.)^H$	Conjugate or Hermitian transpose
$\|.\|$	Absolute value of a number or cardinality of a set
$\Re(.)$	Real part of a complex number
$\Im(.)$	Imaginary part of a complex number
$\|\|.\|\|$	ℓ_2-norm
$E[.]$	Expectation operator
$\angle(.)$	Phase of a complex number or angle between two vectors
$\sum$	Summation operator
$\nabla(.)$	Gradient operator
RSS(.)	Residual sum of squares
sgn(.)	Sign operator
$(.)^{-1}$	Matrix inverse
diag(.)	Diagonal operator
dist(.)	Distance function between two points
erank(.)	Effective Rank, rank of matrix under a particular condition

Symbols and variables

a_{lm}	Eigenbeam (compensated spherical harmonic coefficient) of order l and degree m
b_l	Microphone array mode-strength of order l
$\mathbf{c}$	Centroid
C	Cluster
D	Direct path from source n
h_l	Spherical Hankel function of order l
i	Complex number, $i^2 = -1$
$\mathbf{I}$	Intensity vector
$\mathbf{I}_I$	Identity matrix
j_l	Spherical Bessel function of order l
J_τ	Size of STFT window across time domain
J_k	Size of STFT window across frequency domain
k	Frequency index
K	Number of clusters
l	Spherical harmonic order
L	Maximum spherical harmonic order
m	Spherical harmonic degree
n	Source index
N	Assumed number of active sources
N_s	Overall number of sources
N_ε	Set of neighbouring DOAs within ε neighbourhood
p	Soundfield pressure or probability value for ANOVA test
$\hat{\mathbf{p}}$	DOA estimate unit vector
p_{lm}	Spherical harmonic coefficient of order l and degree m
P_{lm}	Legendre function of order l and degree m
q	Microphone index
$\hat{\mathbf{q}}$	DOA estimate unit vector
Q	Number of microphones

r	Range
$\mathbf{R}$	Covariance matrix
S	Source plane wave
t	Time frame variable
T	Temporal window size
$\mathbf{u}$	Unit vector
U	Set of DOA estimates
$\mathbf{U}$	Matrix of eigenvectors
Y_{lm}	Spherical harmonic basis function of order l and degree m
ε	Angular distance
ϵ	Threshold
η	Singular value ratio (ratio of the largest to the second largest eigenvalue)
θ	Inclination $\in [0, \pi]$
κ	Wavenumber
σ	Standard deviation
$\boldsymbol{\Sigma}$	Matrix of eigenvalues
τ	Time frame index
$\mathit{\Upsilon}$	Set of time-frequency bins
φ	Azimuth $\in [0, 2\pi)$
Ω	Direction

Chapter 1

Introduction

MULTI -Source (MS) Direction-of-Arrival (DOA) estimation is the process of estimating the direction of multiple acoustic sources using their received acoustic signals. It is a fundamental acoustic signal processing task and has been used in areas including source separation/localization/tracking, spatial filtering, environment mapping, dereverberation and speech enhancement. It addresses the often-occurring case in real-world scenarios where multiple sources are simultaneously active. As such it can be used in applications for hearing aids, robot audition, meeting diarization and teleconferencing. The performance of MS DOA estimation can be challenged by several degrading factors including sensor or environmental noise, coherent reflections, referred to as reverberation, low angular separation of sources, simultaneous activity and/or unknown number of sources. The movement of the sources is not considered in DOA estimation since it is mainly treated as a problem in source tracking. Hence it is mostly assumed that the sources are stationary in DOA estimation and also in this thesis.

The hardware tool used for DOA estimation is referred to as a microphone array, which employs multiple microphones configured in a particular geometry to exploit the spatial information encoded in the differences of signals captured by each microphone. There are various types of microphone array such as linear, planar, circular and spherical, each of which benefits from their unique geometry in different applications. The array geometry and the number of microphones respectively indicate the directionality and the spatial resolution of the array's capability. For example, a planar array provides a finer spatial resolution for the sound field to which the array faces compared to the side sound field. Circular, and linear, arrays suffer from limitation to estimation of azimuth only and not the inclination. Such limitation makes the mentioned types of array unsuitable for scenarios in which the 3D full-sphere sound field is of interest, e.g. robot audition where varying azimuths and inclinations, due to various height and position of talkers, are expected. Spherical Microphone Arrays (SMAs) [1, 2] can be used for such applications. This work focuses on Spherical Microphone Arrays, which have recently had growing interest in various fields such as array design [2, 3, 4, 5, 6, 7, 8], spatial filtering [9, 10, 11, 12, 13] and localization [1, 14, 15, 16, 17, 18].

1.1 Spherical Microphone Arrays

The SMAs consist of multiple microphones configured in a spherical geometry. For the purpose of DOA estimation only, the microphones are all distributed with the same distance from the centre of the array (on a spherical shell). Such geometry enables the

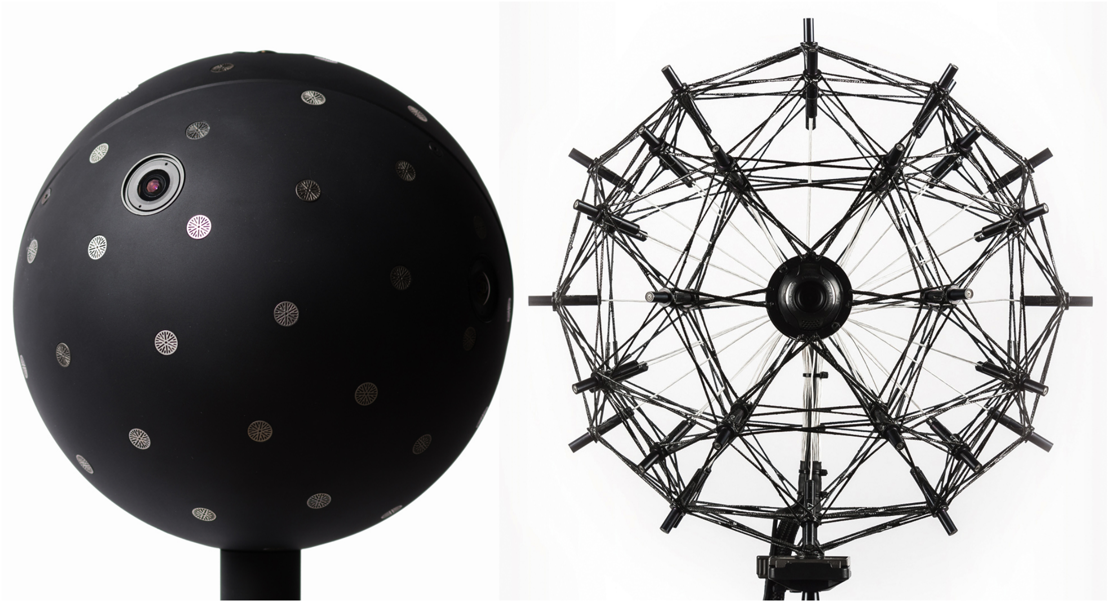


Figure 1.1: The example of (right) an open SMA, Sphere48-35 AC Pro, and (left) a rigid SMA, VisiSonics 5/64. Photo credits: (Right) Acoustic Camera (www.acoustic-camera.com), (Left) VisiSonics (www.visisonics.com)

analysis and processing of the sound field in 3D with equal resolution in all directions. They come with either an open or rigid body as seen in Figure 1.1. The rigid SMAs are often more preferred over the open spheres since the scattering effect of the rigid baffle [19] reinforces the inter-channel time and level differences as shown in [1]. The first works on DOA estimation using SMAs can be found in [20] and the first works on signal processing for SMAs are presented in [21, 4, 22].

Acoustic signal processing for SMAs is mainly done in the Spherical Harmonic Domain (SHD), which provides an elegant mathematical framework to represent the sound field on the surface of the sphere independent of the array properties such as radius, body type and the number of microphones (sampling scheme). In addition, it decouples the range-and-direction dependency of the signals into two separate range- and direction-dependant components making it a preferred domain for the formulation of the problem and solution in the context of DOA estimation. In the SHD, an arbitrary sound field is decomposed into a weighted sum of predefined orthogonal and spatially harmonic sound fields ordered by their spatial resolution. The weights and the pre-defined sound fields are respectively called SH coefficients and basis functions.

DOA estimation for SMAs, as for other types of arrays, can be discussed in two frequency scales: narrowband and wideband. The narrowband DOA estimation assumes a narrowband signal for which the frequency can be assumed to be fixed. On the other hand, wideband DOA estimators aim to estimate the DOA of a wideband signal consisting of

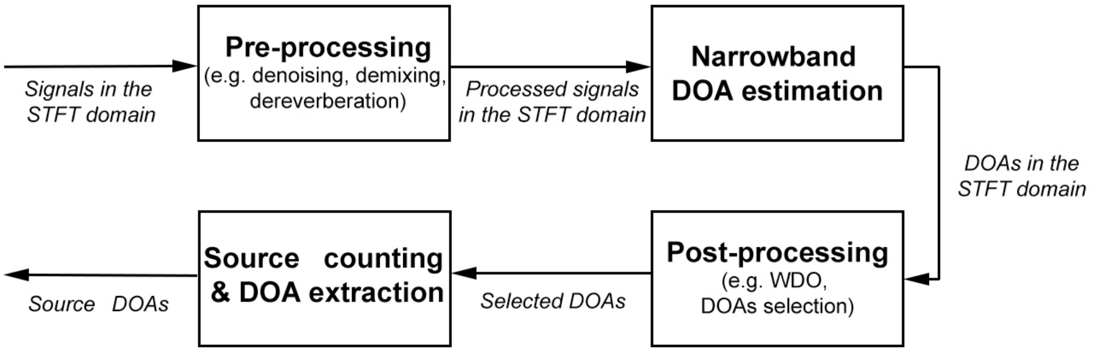

Pre-processing
(e.g. denoising, demixing, dereverberation)
Signals in the STFT domain
Processed signals in the STFT domain
Narrowband DOA estimation
DOAs in the STFT domain
Post-processing
(e.g. WDO, DOAs selection)
Selected DOAs
Source counting & DOA extraction
Source DOAs

Figure 1.2: The system block diagram for conventional wideband MS DOA estimation.

various frequencies. The conventional DOA estimation methods for the wideband signals, such as speech, usually perform narrowband DOA estimation in the Short-Time Frequency Domain (STFT). The DOA outcomes from each Time-Frequency (TF) bin are gathered and post-processed to obtain the source(s)' DOA(s).

Figure 1.2 illustrates the conventional system for wideband DOA estimation. The chain consists of four main blocks as follow: 1) A pre-processing stage where signals in the STFT domain are denoised, demixed or dereverberated in order to extract the signals representing isolated sources. Some techniques, as will be discussed later, may be used in this stage to select some TF bins rather than all for the next stage of processing. 2) A narrowband DOA estimation, mostly based on single source assumption, is performed per TF bin or maybe only for some pre-selected TF bins. 3) The DOAs in the STFT domain (one per TF bin) may be post-processed to keep the reliable DOAs and remove the outlier DOAs. 4) The remaining DOAs are processed for source counting (if number of sources is unknown) and extract the final DOAs belonging to sources. Such a procedure can be seen as a chain of operating blocks each of which deals with one or more particular challenge(s) in DOA estimation.

1.2 Challenges in DOA estimation

Several degrading factors including noise, simultaneously active multi-source, reverberation, low separation and unknown number of sources can challenge the problem-solving for DOA estimation. An efficiently realistic computational complexity must also be considered for real-time applications in real-world scenarios.

1.2.1 Noise

The core problem formulation of the conventional narrowband DOA estimation methods in the simplest case assumes a noise-free scenario. The noise-robustness is mainly achieved by employing de-noising, which is either performed on the input noisy signals (pre-processing block in Figure 1.2) or on the erroneous outcome DOAs (post-processing block in Figure 1.2). A well-known and commonly-used technique for de-noising the input signal is subspace decomposition [23, 24] in which the covariance of the observed noisy correlated signals is decomposed into linearly uncorrelated and weighted principal components using Principal Component Analysis (PCA). The de-noising is performed by splitting the principal components into two sets of signal and noise subspaces distinguished by a thresholding-approach based on the principal components' weight. The well-known MUltiple Signal Classification (MUSIC) DOA estimator [14, 25] uses direction-dependent minimization on the noise subspace. Maximum-Likelihood (ML) approaches [26, 27, 28, 29] are also often used in which an optimization is performed on an objective function to find the model with the best fit to noisy observations. They require accurate models for the noise and mostly assume the presence of isotropic noise (meaning equal noise power from all directions).

Another technique used by some noise-robust DOA estimators [30, 31] is to perform de-noising on the set of DOA estimates obtained from the noisy data. This is based on the assumption that there are often-occurring TF bins with relatively high Signal-to-Noise Ratio (SNR) in which the resulting narrowband DOA estimate is relatively reliable. The non-deterministic behaviour of the noise results in erroneous DOAs with stochastic direction and inaccuracy. Such characteristics are employed to distinguish the reliable and noisy DOA estimates using their spatial density/spread [32, 33, 34, 35]. These techniques are used in the post-processing block in Figure 1.2.

1.2.2 Multi-source (correlated and uncorrelated)

A conventional approach for uncorrelated MS-robustness is the use of W-Disjoint Orthogonality (WDO) [36], assuming sparseness of each speech in the TF domain. In a MS scenario with multiple simultaneously active talkers there are often TF bins (or regions) where only one source is significantly active. This happens because of the differences in the timing, pauses, voices timbre and the utterances of different talkers. PCA can also be used to decompose the correlated signals of a mixture of uncorrelated sources into multiple uncorrelated components each belonging to a source [23]. Such techniques are employed in the pre-processing block in Figure 1.2.

For scenarios with correlated sources, early reflections, the mixture signal consists of the direct path signal (from the true source) and the reflections (from the image sources). The linear dependency of the reflections on the direct path reduces the rank of the covariance matrix potentially resulting into an erroneous division between signal and noise subspaces. In order to overcome this problem, Frequency Smoothing (FS) [37] can be used to decorrelate coherent reflections by combining information across multiple frequency bands.

1.2.3 Unknown number of sources

Source enumerating (counting) is a fundamental problem in scenarios without *a priori* knowledge of the number of sources. The source counting techniques can be categorised into two groups. The first group estimates the optimum number of the sources using an Information Criterion (IC) metric [38, 37] which trades-off between the model complexity and distortion. The second approaches [39, 40, 30] assume a constraint, e.g. minimum cardinality or density of accurate DOAs, in the distribution of DOAs and consequently estimate the number of resulting clusters ignoring the DOAs which do not meet the constraint. These techniques are used in the final block of the chain in Figure 1.2.

1.2.4 Adjacent sources

Spatial resolution of DOA estimation highly depends on the number of microphones used in the array. A higher number of microphones provides a more accurate model of the sound field and therefore higher spatial resolution, as will be shown in Chapter 2 and Figure 2.2. The use of non-spatial features of the talkers, such as voice timbre or harmonic relations, is an alternative approach to distinguish the adjacent sources. Such challenge is not addressed in this thesis.

1.2.5 Computational cost

In terms of the solution's mathematical form and computational complexity, DOA estimators can be categorised in two groups: (1) **Analytical** methods [16, 41, 42] formulate the problem as a determined system of equations and provide a closed-form solution. Although they benefit from their low computational complexity, which makes them suitable for fast DOA estimation, they are prone to noise as their problem formulation is based on either a noise-free scenario or low-order spatial information. (2) **Steering** methods [43, 14], on the other hand, formulate the problem as an underdetermined system of equations assuming the noise as variables in the equations. Given an assumed statistical model of the noise such as isotropic, an objective function as a function of direction is formed. Since the solution is in the wrapped spatial domain, an optimization is performed on the objective function in the form of either maximizing a closeness-of-fit function or minimizing a cost/loss function where the maxima or minima represent the solutions respectively. The higher spatial resolution in the look direction domain results in a higher accuracy as well as a higher computational cost. If *a priori* knowledge of the minimum angular separation of the sources is assumed, an efficient spatial resolution of look directions can be obtained to minimize the complexity with maximum accuracy. Interpolation can be used to improve the accuracy of discrete outcomes. Although steering-based techniques are noise-robust, they require an accurate statistical model of the noise and suffer from high computational complexity if they are applied as a narrowband estimator per TF bin with fine spatial resolution of look directions.

1.3 Motivation and Aims

As introduced and discussed in the previous section, there are various solutions proposed for the challenges in DOA estimation. However, they are mostly based on some assumptions which may be violated in a real-world scenario. For example, one important limitation of previously discussed methods in a multi-source scenario is the violation of WDO assumption. This is likely to occur as the number of sources increases or when one source is constantly masked due to lower loudness, shorter activity or further distance compared to other sources. This motivating problem is addressed in more depth in Chapters 5 and 6 where multiple solutions are proposed.

Another example of limited assumption is in source counting methods. The previously named solutions such as information criterion are based on a wide set of statistical assumptions and are prone to radical conditions and spatial distribution of DOAs as shown later in this thesis. The presence of a reliable source counting with more relaxed constraints and less user engagement is yet an existing challenge. Chapter 4 addresses this problem and proposes an autonomous source counting which is more successful in source detection than conventional methods in challenging conditions.

As discussed previously, the reduction of computationality in analytical algorithms is achieved at the cost of loss in accuracy and spatial resolution due to utilizing less spatial information compared to steering-based solutions. To overcome this limitation, Chapter 3 proposes two alternative approaches which provide the accuracy and robustness of steering-based methods with much less computational complexity.

As a summary, this thesis is motivated by the limitations and violation of assumptions in the existing conventional solutions stated above. Each chapter addresses one operating units of a wideband MS DOA estimation system shown in Figure 1.2, analyses the limitations of the baseline and the state-of-the-art methods for that particular unit and proposes various novel methods for each operating unit to overcome the challenges stated previously.

1.4 Thesis contributions

1.4.1 Research Statement

The aim of this thesis is to propose multiple methods for wideband MS DOA estimation using SMAs each addressing one or multiple degrading factors including sensor noise, reverberation, unknown number of simultaneously active sources and computational complexity.

1.4.2 Publications and Software

The following publications and software were produced during the course of this work:

Journal publications

1. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Augmented Intensity Vectors for Direction of Arrival Estimation in the Spherical Harmonic Domain"* in IEEE/ACM Transactions on Audio, Speech and Language Processing. Vol. 25, Issue. 10, pp. 1956-1968, Oct 2017. [44]

2. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Spatial Consistency for Multiple Source Direction-of-Arrival Estimation and Source Counting"* in Journal of the Acoustic Society of America. [Submitted: Under Review]

3. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Narrowband Multi-Source DOA Estimation in the Spherical Harmonic Domain"* in IEEE/ACM Transactions on Audio, Speech and Language Processing. [Submitted: Under Review]

Conference publications

1. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Robust Source Counting for DOA estimation using Density-based Clustering"* in IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), Sheffield, UK, July 2018. [45]

2. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Multiple DOA estimation based on Estimation Consistency and Spherical Harmonic MUSIC"* in Proceeding European Signal Processing Conference (EUSIPCO), Kos Island, Greece, Sep 2017. [46]

3. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *"Multi-Source Estimation Consistency for Improved Multiple Direction-Of-Arrival Estimation"* in Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), San Francisco, USA, Mar 2017. [47]

4. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *“Multiple Source Localization using Estimation Consistency in the Time-Frequency Domain”* in Proceeding IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), New Orleans, USA, Mar 2017. [48]

5. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *“Multiple Source Localization in the Spherical Harmonic Domain using Augmented Intensity Vectors based on Grid Search”* in Proceeding European Signal Processing Conference (EUSIPCO), Budapest, Hungary, Sep 2016. [49]

6. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *“3D Acoustic Source Localization in the Spherical Harmonic Domain based on Optimized Grid Search”* in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar 2016. [50]

7. **S. Hafezi**, A. H. Moore, and P. A. Naylor, *“Modelling Source Directivity in Room Impulse Response Simulation for Spherical Microphone Arrays”* in Proceeding European Signal Processing Conference (EUSIPCO), Nice, France, Sep 2015. [51]

Software

1. **S. Hafezi**, *”Room Impulse Response generator for Directional source (RIRD)”* http://www.ee.ic.ac.uk/sap/rirdgen/ [52]

2. **S. Hafezi**, *”Spherical Microphone array Impulse Response generator for Directional source (SMIRD)”* http://www.ee.ic.ac.uk/sap/smirdgen/ [53]

1.4.3 Original Contributions

The following aspects of thesis are, to the best of the author’s knowledge, original contributions:

1. Development of a narrowband DOA estimation method using high order harmonics in two optimized ways. *(Chapter 3, published in [50, 49, 44])*

- Proposing an efficient optimization to utilize high order harmonics in two versions. *(Chapter 3.2.3 and 3.2.4, published in [50, 49, 44])*
- Theoretical analysis of DOA estimation error caused by isotropic noise. *(Chapter 3.2.2, published in [44])*
- Comparison of two variations of the proposed method in Single Source (SS) scenario for varying sensor noise level and Reverberation Time (RT). *(Chapter 3.3.1, published in [50, 44])*
- Quantification of the improvement in SS scenario for varying sensor noise level and RT. *(Chapter 3.3.1, published in [50, 44])*
- Comparison of the best performing proposed method to the baseline and the state-of-the-art methods in SS scenario for varying sensor noise level and RT. *(Chapter 3.3.1, published in [50, 44])*
- Comparison of the proposed method to the baseline and the state-of-the-art methods in MS scenario for varying RT, number and angular separation of the sources. *(Chapter 3.3.2, published in [49, 44])*
- Performance evaluation and illustrative validation of the proposed method compared with others with real-world data recordings. *(Chapter 3.4, published in [44])*
- Analysis and comparison of the computational complexity of the proposed methods and the comparative ones. *(Chapter 3.5, published in [49, 44])*

2. Development of an autonomous robust source counting method using density-based clustering in an evolutionary framework. *(Chapter 4, published in [45])*

    - The use of density-based clustering in the context of DOA estimation. *(Chapter 4.1, published in [45])*
    - Proposing an evolutive approach for source counting and sources direction extraction. *(Chapter 4.2, published in [45])*
    - Proposing a density-based strategy to trade-off between accurate and inaccurate DOA estimates. *(Chapter 4.3, published in [45])*

- Performance evaluation of the proposed technique compared with baseline techniques using generated and estimated DOA estimates for varying separation and number of sources. *(Chapter 4.4, published in [45])*

3. Development of a SS-validity confidence metric using Estimation Consistency (EC) based on MS assumption and a variation of MUSIC using the proposed metric. *(Chapter 5, published in [48, 47, 46] and to be published as a comprehensive journal paper)*

    - Proposing a SS-validity confidence metric based on SS assumption in a time frame. *(Chapter 5.2, published in [48])*
    - Evaluation of the effect of the choice of setting parameters for the proposed and the comparative methods in MS scenario with varying setting parameters values, number and angular separation of sources. *(Chapter 5.3, published in [48])*
    - The use of adaptive distance-based and density-based clustering for estimation of the number of active sources per frame. *(Chapter 5.2.1, published in [47, 46])*
    - Proposing a SS-validity confidence metric based on MS assumption in a time frame. *(Chapter 5.2.2, published in [47, 46])*
    - Evaluation of the average accuracy of DOAs selected by the variations of the proposed and the comparative metrics in MS scenario with varying number and angular separation of sources. *(Chapter 5.3.1, to be published.)*
    - Evaluation of the correlation between the accuracy of DOAs and their weight for both variations of the proposed and the comparative metrics in MS scenario with varying number and angular separation of sources. *(Chapter 5.3.2, to be published.)*
    - Evaluation of the effect of each metric in the performance of source counting and wideband DOA estimation with varying number and angular separation of sources. *(Chapter 5.3.3, to be published.)*

- Performance evaluation and illustrative validation of the proposed metric using real-world data recording. *(Chapter 5.4, to be published.)*
- Proposing a variation of MUSIC using the previously proposed metric. *(Chapter 5.6, published in [46])*
- Evaluation of the effect of the choice of setting parameter for the proposed technique. *(Chapter 5.7.1, published in [46])*
- The performance evaluation of the proposed DOA estimation compared to state-of-the-art method with varying number and angular source separation. *(Chapter 5.7.2, published in [46])*

4. Development of two analytical narrowband DOA estimators based on multi-source and double-source assumption. *(Chapter 6, published in [48, 47, 46] and to be published as a comprehensive journal paper)*

   - Proposing an extension to an analytical subspace DOA estimator based on MS assumption. *(Chapter 6.1, to be published.)*
   - Proposing an analytical DOA estimator based on double-source assumption. *(Chapter 6.2, to be published.)*
   - Proposing a subspace variation of the proposed method based on double-source assumption to improve noise-robustness. *(Chapter 6.4, to be published.)*
   - Illustrative validation in narrowband and wideband scenario for both proposed methods and the comparative ones with varying sensor noise level and sources mixing ratio. *(Chapter 6.3 and 6.5, to be published.)*

1.4.4 Thesis outline

The content of this thesis is structured as follows:

- **Chapter 2**: Reviews of the SHD, signal model, subspace decomposition, various baseline and the state-of-the-art methods for DOA estimation are provided.

- **Chapter 3**: An Augmented Intensity Vector (AIV) is proposed which improves the accuracy of Pseudo-intensity Vectors (PIVs) by exploiting higher order spherical harmonics in two optimized versions. A comparison is performed using our proposed AIVs against PIVs, Steered Response Power (SRP) and subspace methods where the number of sources, their angular separation, RT and the sensor noise level are varied. The results show that the proposed approach outperforms the baseline methods and performs at least as accurately as the state-of-the-art method with strong robustness to reverberation, sensor noise and number of sources with significantly less computational cost. In the single and multiple source scenarios tested, which include realistic levels of reverberation and noise, the proposed method had average error of 1.5° and 2°, respectively.

- **Chapter 4**: A method of source counting for DOA estimation using density-based clustering is proposed. Multiple Density-based Spatial Clustering of Applications with Noise (DBSCAN) with varying noise sensitivity is applied in an evolutionary procedure to obtain weighted centroids. An autonomous DBSCAN is finally run on the weighted centroids to extract the sources' DOAs. The results using generated and estimated DOAs show that the proposed technique significantly outperforms the conventional histogram peak picking as well as the original DBSCAN and variations of Kmeans with $\leq 4°$ DOA estimation accuracy and improves the source counting.

- **Chapter 5**: A novel SS-validity confidence metric is proposed that exploits a dynamic MS assumption over relatively large TF regions. The proposed metric first clusters the initial DOA estimates (one per TF bin) and then uses the members' spatial consistency as well as its cluster's spread to weight each TF bin. Distance-based and density-based clustering are employed as two alternative approaches for clustering DOAs. A noise-robust density-based clustering is also used for source counting and source direction estimation. The evaluation results show that the proposed weighting significantly improves the accuracy of source counting and MS DOA estimation compared to the state-of-the-art. As a result, a version of MUSIC is also proposed in which all the SS bins for each talker across the TF domain are globally used to improve the quality of covariance matrix for MUSIC. The simulation shows that the proposed technique significantly outperforms the state-of-the-art with $< 6.5^\circ$ mean estimation error and strong robustness to widely varying source separation for up to 5 sources in the presence of realistic reverberation and sensor noise.

- **Chapter 6**: Two novel analytical approaches are proposed for narrowband DOA estimation based on MS assumption in a bin for low reverberant environment. In the first approach Eigenvalue Decomposition (EVD) is used to decompose a MS scenario into multiple SS scenarios on each of which a SS-based analytical DOA estimation is performed. The second approach analytically estimates up to two DOAs per bin assuming the presence of two active sources per bin. EVD is used to extend the second approach to scenarios where more than two sources are active. The evaluation validates an improvement of double accuracy and robustness to sensor noise compared to the baseline methods.
- **Chapter 7**: The thesis is concluded and potential future works are discussed.

Chapter 2

Background

THIS section reviews the signal representation and problem formulation in the SHD, as well as the baseline and the state-of-the-art DOA estimation methods used in our evaluations and comparisons. It finally introduces and justifies various evaluation metrics as well as evaluation datasets used in this thesis. Each Chapter addresses a specific challenging scenario for each of which there are different comparative methods available in the literature. Hence the literature reviews and discussion of the methods are provided at the beginning of each Chapter where only the methods suitable for that specific scenario are discussed.

2.1 Coordinate system

In this thesis, spatial information is presented in the spherical coordinate system unless otherwise stated. The spherical coordinates are $(r, \Omega) = (r, \theta, \varphi)$ with range r, inclination $\theta \in [0, \pi]$, and azimuth $\varphi \in [0, 2\pi)$ as illustrated in Fig. 2.1. Note that the term elevation in this work refers to the complementary angle of the inclination and is defined within $[\frac{\pi}{2}, -\frac{\pi}{2}]$.

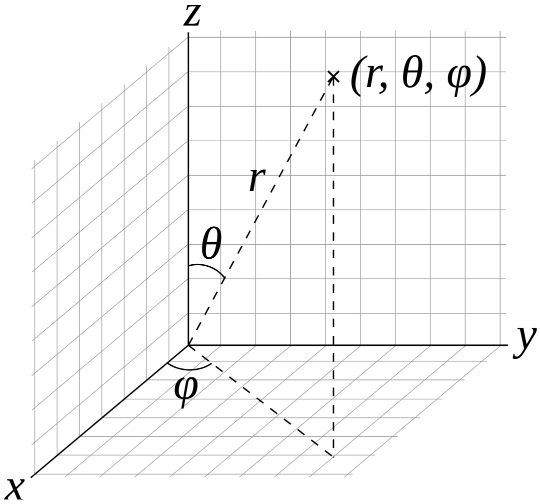


Figure 2.1: The spherical coordinate system.

2.2 Spherical Harmonic Domain (SHD)

Consider the sound pressure field $p(\kappa, r, \Omega)$ as a function of wavenumber κ and the point location (r, Ω). The Spherical Harmonic Transform (SHT) of this field is given by [54]

$$p_{lm}(\kappa, r) = \int_{\Omega \in S^2} p(\kappa, r, \Omega) Y^*_{lm}(\Omega)\, d\Omega, \tag{2.1}$$

where $\int_{\Omega \in S^2} d\Omega = \int_0^{2\pi} \int_0^{\pi} \sin(\theta)\, d\theta d\varphi$, and $(.)^*$ denotes the complex conjugate.

The complex-valued SH basis function $Y_{lm}(\Omega)$ of order l and degree m (satisfying $|m| \leq l$) is given by [54]:

$$Y_{lm}(\Omega) = \sqrt{\frac{(2l+1)}{4\pi} \frac{(l-m)!}{(l+m)!}} P_{lm}(\cos(\theta))\, e^{im\varphi}, \tag{2.2}$$

where P_{lm} is the associated Legendre function and $i^2 = -1$.

The real form SHs are defined as

$$Y_l^m = \begin{cases} \sqrt{2}\,(-1)^m\, \Im\left(Y_{l|m|}\right), & \text{if } m < 0 \\ Y_{lm}, & \text{if } m = 0 \\ \sqrt{2}\,(-1)^m\, \Re\left(Y_{lm}\right), & \text{if } m > 0 \end{cases} \tag{2.3}$$

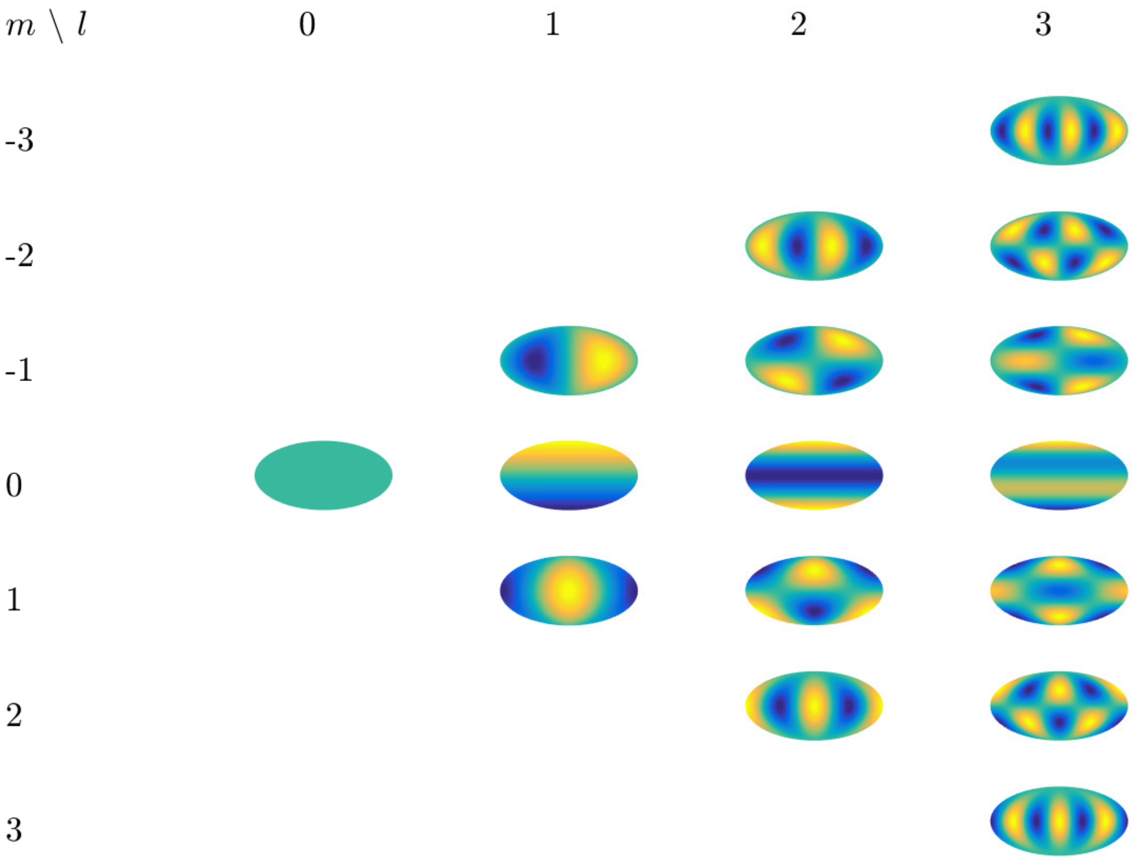


Figure 2.2: Mollweide projection of the real form SH basis functions, Y_l^m, up to the 3rd order.

which are illustrated in Figure 2.2. Note that (Ω) is omitted for notational simplicity.

Using the inverse SHT, the sound pressure field can be reconstructed as [2]

$$p(\kappa, r, \Omega) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} p_{lm}(\kappa, r) Y_{lm}(\Omega), \tag{2.4}$$

where the coefficients p_{lm} are spherical harmonic coefficients.

Considering a SMA with radius r_a and Q microphones each with angle Ω_q, the integral in (2.1) is approximated as

$$p_{lm}(\kappa, r_a) = \sum_{q=1}^{Q} p(\kappa, r_a, \Omega_q) w_{q,lm}, \tag{2.5}$$

where in the case of a uniform sensor distribution, the weights $w_{q,lm}$ are

$$w_{q,lm} = \frac{4\pi}{Q} Y_{lm}^{*}(\Omega_q). \tag{2.6}$$

For up to harmonic order L, there are $(L+1)^2$ independent harmonics (including all the degrees m for each order l satisfying $|m| \leq l$). In order to avoid spatial aliasing, the number of microphones Q must satisfy [6]

$$Q \geq (L+1)^2. \tag{2.7}$$

Further information regarding the analysis of spatial aliasing errors and the selection of an appropriate spatial sampling scheme can be found in [1].

Since the interested points to analyse the sound pressure field are on the surface of the SMA which has a known constant radius $r = r_a$, the coefficients $p_{lm}(\kappa, r)$ can be simplified to be range-independent. This allows the formulation of the problem and solution to be independent of the array geometry, configuration and type of its body, e.g. rigid or open. Such dependency is compensated by the mode strength $b_l(\kappa r_a)$ and is used to form the eigenbeams $a_{lm}(\kappa)$ such that

$$a_{lm}(\kappa) = \frac{p_{lm}(\kappa, r_a)}{b_l(\kappa r_a)}. \tag{2.8}$$

For a rigid SMA, as used in our experimental study, with radius r_a the mode strength $b_l(\kappa r_a)$ is given by [54]

$$b_l(\kappa r_a) = 4\pi i^l \left[j_l(\kappa r_a) - \frac{j_l^{'}(\kappa r_a)}{h_l^{(2)'}(\kappa r_a)} h_l^{(2)}(\kappa r_a) \right], \tag{2.9}$$

where j_l is the spherical Bessel function of order l, $h_l^{(2)}$ is the spherical Hankel function of the second kind and of order l, and $(.)^{'}$ denotes the first derivative. For an open sphere the mode strength is $b_l(\kappa r_a) = 4\pi i^l j_l(\kappa r_a)$.

2.3 Problem formulation in the SHD

Consider multiple N_s sources in the far-field with arriving plane waves $\{S_n\}_{n=1}^{N_s}$ and DOAs $\{\Omega_n = (\theta_n, \varphi_n)\}_{n=1}^{N_s}$, the plane-wave representation of the eigenbeams is [1, 2]

$$a_{lm}(\tau, k) = \left(\sum_{n=1}^{N_s} S_n(\tau, k) Y_{lm}^*(\Omega_n) \right) + n_{lm}(\tau, k), \tag{2.10}$$

where n_{lm} is the sensor noise in the SHD, τ and k are time frame and frequency respectively. The matrix representation of (2.10) gives

$$\mathbf{a_{lm}}(\tau, k) = \mathbf{Y}^H(\mathbf{\Omega})\mathbf{S}(\tau, k) + \mathbf{n_{lm}}(\tau, k), \tag{2.11}$$

where

$$\mathbf{Y}(\mathbf{\Omega}) = \begin{bmatrix} \mathbf{Y_{lm}}^T(\Omega_1) \\ \vdots \\ \mathbf{Y_{lm}}^T(\Omega_{N_s}) \end{bmatrix}, \tag{2.12}$$

and $\mathbf{a_{lm}} = \left[a_{00}, a_{1(-1)}, a_{1(0)}, a_{1(1)}, \ldots, a_{LL}\right]^T$ is the $(L+1)^2$-length column vector of eigenbeams, $\mathbf{Y_{lm}} = \left[Y_{00}\, Y_{1(-1)}, Y_{1(0)}, Y_{1(1)}, \ldots, Y_{LL}\right]^T$ is the same-length column vector of SH basis functions, $\mathbf{S} = [S_1 \ldots S_{N_s}]^T$ and $\mathbf{\Omega} = [\Omega_1 \ldots \Omega_{N_s}]^T$ are the N_s-length column vectors of plane-wave signals and their associated DOA respectively and $(.)^H$ denotes the conjugate transpose. Note that (τ, k) and (Ω) are omitted for notational simplicity.

2.4 Steered Response Power (SRP)

One of the baseline DOA estimation methods used for comparison is the Steered Response Power approach [43] implemented in the SHD using a Plane Wave Decompositions (PWD) beamformer [55]. The output of the beamformer steered to look direction Ω is given as [11]

$$y(\tau, k, \Omega) = \sum_{l=0}^{L_b} \sum_{m=-l}^{l} a_{lm}(\tau, k) Y_{lm}(\Omega), \tag{2.13}$$

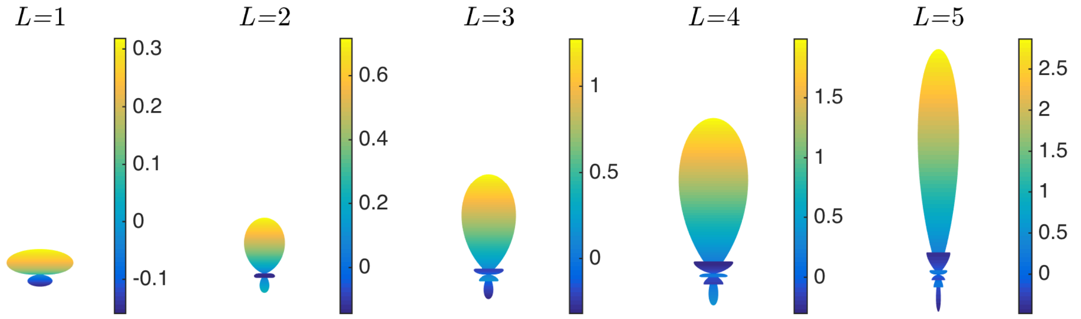


Figure 2.3: PWD beam pattern for up to 5th order.

where L_b is the maximum harmonic order used in the beamforming. Figure 2.3 illustrates the beam directivity pattern of PWD for look direction towards the z-axis with varying L_b. As L_b increases, the beam becomes narrower and more directional. In addition to the main lobe (look direction), the presence of the side and back lobes results in leakage from other directions other than the look direction.

The narrowband SRP spectrum is the power of the beamformer's output

$$P_{SRP}(\tau, k, \Omega) = \mid y(\tau, k, \Omega) \mid^2 . \tag{2.14}$$

Assuming the presence of a single source, the narrowband DOA estimate is the global peak in the SRP spectrum while the wideband DOA estimate is the peak in the SRP spectrum summed over all TF bins

$$\Omega_{SRP} = \arg\max_{\Omega} \sum_{(\tau,k)} P_{SRP}(\tau, k, \Omega). \tag{2.15}$$

2.5 Pseudo-intensity Vectors (PIVs)

In acoustics, sound intensity is a measure of the flow of sound energy through a surface per unit area, in a direction perpendicular to this surface. The intensity vector I, which defines the magnitude and the direction of the energy flow can be determined by calculating the flow of sound energy through the three unit surfaces perpendicular to the Cartesian axes as [56]:

$$\mathbf{I}(k) = \frac{1}{2}\Re\left\{p^{*}.\mathbf{v}\right\}, \tag{2.16}$$

where p is the sound pressure, $\mathbf{v} = [v_x, v_y, v_z]^T$ is the particle velocity in Cartesian coordinates, and $\Re\{.\}$ denotes the real part of a complex number.

Since in practice it is difficult to measure the particle velocity, an alternative is to use pseudo-intensity vectors. PIVs are calculated using the zeroth- and the first-order eigenbeams as [16]

$$\mathbf{I_{piv}}(\tau, k) = \frac{1}{2}\Re\left\{ a_{00}(\tau, k)^{*} \begin{bmatrix} D_{-x}(\tau, k, \mathbf{a_{lm}}) \\ D_{-y}(\tau, k, \mathbf{a_{lm}}) \\ D_{-z}(\tau, k, \mathbf{a_{lm}}) \end{bmatrix} \right\}, \tag{2.17}$$

where

$$D_{\nu}(\tau, k, \mathbf{a_{lm}}) = \sum_{m=-1}^{1} Y_{1m}(\phi_{\nu}) a_{1m}(\tau, k),\ \nu \in \{-x, -y, -z\} \tag{2.18}$$

are dipoles steered in the negative direction of Cartesian axes, given by $\phi_{-x} = (\pi/2, \pi)$, $\phi_{-y} = (\pi/2, -\pi/2)$ and $\phi_{-z} = (\pi, 0)$.

The DOA unit vector $\mathbf{u}(k)$ is given by

$$\mathbf{u_{piv}}(\tau, k) = -\frac{\mathbf{I_{piv}}(\tau, k)}{\|\mathbf{I_{piv}}(\tau, k)\|}, \tag{2.19}$$

where $\|.\|$ indicates ℓ_2-norm.

2.6 Subspace Decomposition

In order to improve the noise-robustness, subspace decomposition can be used to split the eigenbeams' covariance matrix into signal and noise subspaces.

Assuming no correlation between the noise and the source signal, the covariance matrix of the noisy signals in the SHD is decomposed as

$$\begin{aligned}\mathbf{R}(\tau,k) =& E\left[\mathbf{a_{lm}}(\tau,k)\mathbf{a_{lm}}^H(\tau,k)\right]\\ =&\mathbf{Y}^H(\mathbf{\Omega_n})\mathbf{R_s}(\tau,k)\mathbf{Y}(\mathbf{\Omega_n})+\mathbf{R_v}(\tau,k),\end{aligned} \tag{2.20}$$

where $\mathbf{R_s} = E\left[\mathbf{S}\mathbf{S}^H\right]$ and $\mathbf{R_v}$ are respectively the source and noise signals' covariance matrices and $E\left[.\right]$ denotes the expectation. In the case of spatially white sensor noise, the noise covariance matrix is proportional to the identity matrix where the proportion is the noise variance.

In order to obtain the covariance matrix $\mathbf{R}$, it is approximated as the average covariance matrix over a local TF region [14, 17]

$$\begin{aligned}\mathbf{R_a}(\tau,k) =& \frac{1}{J_\tau J_k}\sum_{j_\tau=0}^{J_\tau-1}\sum_{j_k=0}^{J_k-1}\mathbf{a}(\tau-j_\tau,k+j_k)\\ &\times \mathbf{a}^H(\tau-j_\tau,k+j_k),\end{aligned} \tag{2.21}$$

where J_τ and J_k are the widths (number of bins) of the averaging windows over time and frequency respectively.

Using EVD, the covariance matrix of the eigenbeams is decomposed as

$$\mathbf{R_a} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H, \tag{2.22}$$

where $\mathbf{\Sigma}$ is the diagonal matrix containing the decreasingly sorted eigenvalues and $\mathbf{U}$ is the square matrix containing the eigenvectors as its columns. Note that (τ,k) are omitted here for notational simplicity.

Assuming the presence of N ($N \leq (L+1)^2$) active sources, let $\mathbf{U_s}^{(N)}$, representing the signal subspace, be the first N columns of $\mathbf{U}$ and $\mathbf{U_v}^{(N)}$, representing the noise subspace, be the rest

$$\mathbf{U_s}^{(N)} = [\mathbf{U}_1 \ldots \mathbf{U}_N] \tag{2.23}$$

$$\mathbf{U_v}^{(N)} = \left[\mathbf{U}_{N+1} \ldots \mathbf{U}_{(L+1)^2}\right]. \tag{2.24}$$

2.7 Subspace-PIV

To improve the noise-robustness of PIV, in [17] the authors proposed Subspace-PIV (SSPIV), which assumes $N = 1$. In SSPIV, the raw eigenbeams are replaced with the signal subspace $\mathbf{U_s}^{(1)} = \mathbf{U}_1$, which is the first column of $\mathbf{U}$ in (2.22) giving

$$\mathbf{I_{sspiv}}(\tau, k) = \frac{1}{2}\Re\left\{ a_{00}(\tau, k)^* \begin{bmatrix} D_{-x}(\tau, k, \mathbf{U}_1) \\ D_{-y}(\tau, k, \mathbf{U}_1) \\ D_{-z}(\tau, k, \mathbf{U}_1) \end{bmatrix} \right\}. \tag{2.25}$$

Note that in the original SSPIV paper [17], Singular Value Decomposition (SVD) is used instead of EVD, which is equivalent since $\mathbf{R_a}$ is symmetric.

2.8 Multiple Signals Classification (MUSIC)

One of the most well-known subspace-based DOA estimators is MUSIC, in which EVD is used to decompose the observed noisy signals of a mixture of sources into signal and noise subspaces. MUSIC uses the spatial steering vector and the noise subspace to obtain the MUSIC spectrum given by [23]

$$P_{MUSIC}(\Omega) = \frac{1}{\sum \|(\mathbf{U_v}^{(N)})^H \mathbf{Y_{lm}}^*(\Omega)\|^2}. \tag{2.26}$$

Note that (τ, k) is omitted for notational simplicity. The estimated DOAs of N sources are the first N highest peaks in the MUSIC spectrum.

2.9 Direct-Path-Dominance (DPD) MUSIC

In [14] authors proposed a reverberation-robust variation of MUSIC based on a narrowband single source assumption. A Direct-Path-Dominance (DPD) test is designed to identify the TF regions with significant contribution from the direct-path signal of a single source where the significance of dominance is defined by a user-controlled threshold. By selecting only those TF bins passing the test, it aims to improve the robustness to reverberation and noise. The TF bins that pass the DPD test are

$$\Upsilon_{\mathrm{DPD}} = \left\{(\tau, k) : \ \mathrm{erank}\left(\mathbf{R}_a(\tau, k)\right) = 1\right\}, \tag{2.27}$$

where

$$\mathrm{erank}\left(\mathbf{R}_a(\tau, k)\right) = 1 \text{ if } \eta_{DPD}(\tau, k) > \epsilon \tag{2.28}$$

is the effective rank, η_{DPD} is Singular Value Ratio (SVR), the ratio of the largest and the second largest singular values, of $\mathbf{R}_a$ and ϵ is a user-defined threshold. The two alternative approaches [14] to apply MUSIC on the outcome of the DPD test are discussed next.

Incoherent DPD-MUSIC

In the first approach the MUSIC spectra in (2.26) are summed over the selected TF bins $(\tau, k) \in \Upsilon_{DPD}$ so that

$$P_{incoh-MUSIC}(\Omega) = \sum_{(\tau,k)\in\Upsilon_{DPD}} P_{MUSIC}(\tau, k, \Omega), \tag{2.29}$$

where the set of N highest peaks in the final spectrum indicates the overall estimated DOAs.

Coherent DPD-MUSIC

The second approach performs coherent fusion of the directional information from the selected TF bins. The set of one dimensional signal spaces from the selected TF bins, $\left\{\mathbf{U}_{\mathbf{s}}{}^{(1)}\left(\tau, k\right)\right\}_{(\tau,k)\in\Upsilon_{\mathrm{DPD}}}$, are clustered using one-run K-means clustering [14] with random initialization into N clusters with centroids $\{\mathbf{U}_{\mathbf{s}}{}^{n}\}_{n=1}^{N}$ where each centroid signal space is associated with one source. The DOA of an individual source n is selected as the global peak in the coherent MUSIC spectrum of source n which is given as

$$
\begin{aligned}
P^{n}_{coh-MUSIC}(\Omega) =& \frac{1}{\| \left(\mathbf{U_v}^{n}\right)^{H} \mathbf{Y_{lm}}^{*}(\Omega) \|^{2}} \\
=& \frac{1}{\mathbf{Y_{lm}}^{T}(\Omega)(\mathbf{I}_{I} - \mathbf{U_s}^{n} \left(\mathbf{U_s}^{n}\right)^{H})\mathbf{Y_{lm}}^{*}(\Omega)},
\end{aligned}
\tag{2.30}
$$

where $\mathbf{I}_I$ is the identity matrix.

2.10 Evaluation metrics

In order to evaluate and compare the performance of methods in different context, various metrics are used for evaluation in this thesis depending on the scenario and the aim of evaluation.

2.10.1 Estimation error

DOA estimation error ε between a true DOA unit vector $\mathbf{u}_o$ and an estimated DOA unit vector $\mathbf{u}_e$ is computed in degrees as

$$
\varepsilon = \cos^{-1}\left(\mathbf{u}_o^T \mathbf{u}_e\right). \tag{2.31}
$$

For multiple sources with an equal number of estimated and true DOAs, estimation error is the average of ε between all pairs of estimated-true DOAs. Each estimated DOA is paired with a true DOA using best case data association in order to avoid any ambiguity due to data association uncertainty. This metric is used for evaluations in Chapters 3-6.

2.10.2 Mean error

Mean error is simply the average of estimation errors among the trials with equal number of estimated and true DOAs. Each evaluation usually consists of various sets of settings, e.g. noise level, number of sources, separation or reverberation time. Each set of setting includes various trials in each of which only the distribution of true DOAs varies while all other configurations and settings are fixed. This metric is used in Chapters 3-6.

2.10.3 Mean Number of Detected Sources (NoDS)

In the context of DOA estimation for real-world application accuracy is defined within a limited range of error. For example, 120° error is considered as bad and erroneous as 140° error and one does not necessarily provide a better performance. In order to avoid the effect of highly erroneous estimated DOAs on the mean error, NoDS is also used as a metric next to mean error in the performance evaluations. In the best case assignment of estimated and true DOAs, the number of pairs with estimation error $\leq 15°$ (half of the minimum source separation used in our evaluations) is considered as NoDS. The mean NoDS is the average NoDS across all trials. In evaluations where NoDS is used, the mean error is the average estimation error among trials with equal number of NoDS and true DOAs. This metric is used in Chapter 3.

2.10.4 Successful Localization Rate (SLR)

SLR is the percentage of the trials in which NoDS is equal to the number of sources. This metric is used for evaluation of source counting and estimation accuracy in Chapters 4-5.

2.10.5 Number of real multiplications

The number of multiplications between two real numbers are used as a metric for theoretical computational complexity. Note that the multiplication of two complex numbers is counted as four real multiplications while $| \, . \, |^2$ is counted as two real multiplications. This metric is used for evaluation of computational complexity in Chapters 3 and 6.

2.10.6 Peak Strength

As shown later in experimental verifications using recordings in a real room, because of well separation of sources, all methods may successfully estimate DOAs corresponding to all sources. In addition, due to lack of exact measurement of the source position in a real-room with volumetric loudspeaker as sound sources, it is not possible to have an exact numerical true DOA and therefore none of the estimation metric defined so far can be used. In these scenarios the distinctness, sharpness and the height of the peaks in the DOA histogram may still indicate the relative performance of the methods. In order to provide a numerical evaluation under such scenarios, for each peak, a measure of 'peak strength' is proposed which is the ratio of the peak height over the peak smoothness where the peak smoothness is defined as the average height in the normalized peak distribution within its range of $r_p = 30°$ neighbourhood (half of the minimum source separation used for real-room recordings). This metric is used in Chapters 3 and 5.

2.10.7 Accuracy

The normalized accuracy, as used in Chapter 5, is

$$1 - \varepsilon/180. \tag{2.32}$$

2.10.8 Inferential Statistical Analysis

In order to evaluate the significance of difference in the performance of the methods, one-way Analysis of Variance (ANOVA) is used. It tests the hypothesis that the samples are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. The output of the test is presented as probability value p where $p < 0.001$ indicates that there is at least two methods with significantly different mean error. In such cases, a Tukey post hoc analysis is also performed to determine between which pairs of groups the difference is significant. In cases of comparison between two methods only, an independent-samples t-test is conducted to determine if the methods perform significantly different. One-way ANOVA is used in Chapters 3 to 6 and t-test is used in Chapter 5.

2.11 Evaluation Testbeds

In this thesis, various methods are proposed for different conditions and sometimes different assumptions. In order to fairly evaluate the performance of the methods under re-

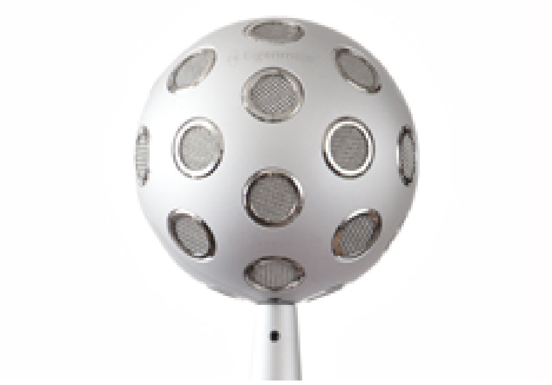


Figure 2.4: 32-channel rigid SMA, em32 Eigenmike®, used in the simulations and real recordings. Photo credit: mh acoustics (www.mhacoustics.com).

lated conditions and configurations, variety of testbed materials are needed. The testbed datasets, including both simulated and real recordings, used in this thesis for the performance evaluation are categorized into five groups.

2.11.1 Single source simulated recordings

The Acoustic Impulse Responses (AIRs) of a 32-element rigid SMA with radius of 4.2 cm (corresponding to the em32 Eigenmike®, as shown in Figure 2.4) in a $5 \times 6 \times 4$ m shoebox room were simulated using Spherical Microphone arrays Impulse Response Generator (SMIRgen) [57] based on Allen & Berkley's image method [58]. The reflection coefficient of the walls is calculated using reverberation time [59]. The array was placed at (2.52, 3.11, 1.97) m and the source signal consists of an anechoic speech signal, using the same utterance for all trials [60] with duration 5 s, convolved with the simulated AIRs to each microphone and white Gaussian sensor noise added. Forty different source positions were considered at the distance of 1 m from the centre of array with a DOA randomly selected from a uniform distribution around the sphere. For each source position, the test was repeated over a range of RT, $T_{60} = \{0.2, 0.3, 0.4, 0.5, 0.6\}$ s, and signal-to-noise ratio, SNR=$\{10, 20, 30\}$ dB.

2.11.2 Multi-source simulated recordings

This dataset is used to evaluate the effect of reverberation, number of sources and angular separation of the sources in scenarios with multiple sources. The room dimension, type of microphone array and the simulation method are the same as in Section 2.11.1.

In order to systematically evaluate the effect of source separation with varying number of sources, multiple sources were distributed with equal angular separation between them. For simplicity of understanding of the result and maximizing the clarity of systematic evaluation of the effect of source separation, the sources were distributed on the same horizontal plane as the microphone array. The datasets in Sections 2.11.1 and 2.11.3 demonstrate the effectiveness of the method in varying azimuth-inclination condition. In total, 100 trials were used where in each trial the azimuth of the first source is chosen randomly from a uniform distribution around the circle and the subsequent sources are placed at regularly spaced azimuth intervals $\Delta\phi_s$. The number of sources N_s and the angular separation $\Delta\phi_s$ vary in each experiment, as described below.

The constraint of fixed inclination reduces the range of variations in distance-to-the-closest-wall and so the strongest reflection in compare to the single source scenario in which both the azimuth and inclination vary per trial. In order to compensate for the reduction in range of variation of the strongest reflection, the microphone array is slightly displaced away from the centre of the room at (2.52, 4.48, 1.45) m in multi-source scenario while the distance of the sources to the centre of SMA stays 1 m.

The source signals consist of different anechoic speech signals randomly selected for each trial from the APLAWD database [61]. The active level of each speech source according to ITU-T P.56 [62], as measured at p_{00}, is set to be equal across all trials. Spatio-temporally white Gaussian noise is added to the microphone signals to produce a signal to incoherent noise ratio (SNR) of 25 dB at p_{00} for each source. The dataset is divided into two subsets: 1) Varying $T_{60} = \{0.2, 0.4, 0.6\}$ s with $N_s = 2$ and $\Delta\phi_s = 45°$; 2) Varying $N_s = \{2, 3, 4, 5\}$ and $\Delta\phi_s$ from 5° to maximum possible separation with $T_{60} = 0.4$ s.

2.11.3 Multi-source real recordings

This dataset is used to demonstrate the performance of each method in a real-world multi-source scenario. A real recording of 4 s speech in a room with approximate dimensions of $10 \times 9 \times 2.5$ m and reverberation time of 0.4 s was used using an Eigenmike 32-channel rigid spherical microphone array, shown in Figure 2.4, with radius of 4.2 cm placed close to the centre of the room. Four talkers were simultaneously active and were located 1.5 m away from the centre of the array at approximately 60° intervals while their inclinations alternated to be above or below the horizontal plane of the array.

2.11.4 Multi-source simulated DOAs

This dataset is used to expand the systematic evaluation of source counting methods to scenarios with varying controlled spatial distribution of DOAs in addition to scenarios with varying controlled environmental parameters such as T_{60}.

The simulator for DOA spatial distribution uses two variables defined as Noise Proportion (N_p), which is the ratio of the number of noise DOAs to the total DOAs, and the Sources Proportion (S_p), which is the ratio of the DOAs from a single distribution associated with a single source to all DOAs (excluding the noise DOAs). For example, with 500 total DOAs and $\{N_s, N_p, S_p\} = \{2, 0.8, 0.3\}$, there are 400 noise DOAs plus 30 and 70 DOAs generated for source 1 and 2 respectively. The Von Mises-Fisher distribution with true DOA as mean direction and $\kappa = 30$ as concentration parameter were used to generate the DOAs. The noise DOAs were randomly selected from a uniform distribution around the sphere. For N_s sources, 100 trials per experiment were used. The first true DOA was randomly selected from uniform distribution around the sphere and the subsequent true DOAs placed with Sep separation on the randomly-orientated great circle passing through the first true DOA. A total of 2000 DOAs and $\{N_s, N_p, S_p, Sep\} = \{2, 0.5, 0.5, 60^\circ\}$ was used per trial unless otherwise stated.

2.11.5 Multi-source simulated recordings with masking

This dataset demonstrates the challenging scenario of multi-source where one source being strongly masked by others. It uses recorded anechoic speech signals convolved with direct-path impulse response with varying additive white Gaussian sensor noise for a 32-element rigid SMA with radius of 4.2 cm (corresponding to the em32 Eigenmike®). Four anechoic speech sources were employed in total (3 males and 1 female) each with duration of 2 seconds selected from the APLAWD database [61]. Some segments (with overall duration of 1 second) of source 2 were cut off to make inconsistency of activity as illustrated in later in Chapter 6. The sources were mixed with mixing coefficient of 1 for sources 1, 3 and 4 and 0.2 for source 2 so that source 2 is strongly masked by others. The sources were all 1.5 m away from the array. In each trial, four random source positions were selected from a uniform distribution around the sphere where the angular separation of each two sources was guaranteed to be $\geq 70°$ and no two sources had equal azimuths or inclinations. 100 trials were used for each SNR=$\{10, 20, 30, \text{Inf}\}$ dB.

Chapter 3

Augmented Intensity Vectors (AIVs)

THE PIVs, as shown in the previous chapter, only use zeroth- and first-order SHs ignoring the higher order SHs. Since the spatial frequency of Y_{lm} increases with the SH order as illustrated in Figure 2.2, employing higher order information increases the spatial resolution. In this chapter high order ($l > 1$) information is used to improve the PIV accuracy while avoiding exhaustive search.

3.1 Signal Model

Consider a plane wave $S(\tau, k)$ with DOA $\Omega_u = (\theta_u, \varphi_u)$ arriving from a single source in the far-field in an anechoic (reverberation-free) environment. Using (2.10) with $N_s = 1$ the eigenbeams of the sound field are

$$a_{lm}(\tau, k) = S(\tau, k) Y^*_{lm}(\Omega_u) + n_{lm}(\tau, k), \tag{3.1}$$

where $n_{lm}(\tau, k)$ represents the noise.

In case of a noise-free scenario, $n_{lm} = 0$, considering (3.1) for $l \in [0, L]$ and $-l \leq m \leq l$, there would be $(L+1)^2$ complex equations with two unknowns S and Ω_u. In this case even for $L = 1$ the system of equations is overdetermined and the solution can be accurately obtained. Increasing L still results in the same solution that is the accurate true DOA.

Considering the noisy case with non-zero n_{lm}, (3.1) is an underdetermined system of equations. In such scenario, the aim is to find the Ω which best satisfies all the $(L+1)^2$ equations of (3.1) for up to SH order L.

3.2 AIV

3.2.1 Cost function

The zeroth SH order has the noise-reducing characteristic since the noise signals at the individual sensors are averaged and reduced as in (2.5). For spatially-white noise this reduction is approximately 10 dB based on (2.6). Using this noise-reducing property of $l = 0$, it is assumed that $n_{00}(\tau, k) = 0$ (noise-free omnidirectional eigenbeam only), which for moderate sensor noise level is a suitable approximation, to approximate $S(\tau, k)$ by substituting (2.2) into (3.1) for $l = 0$ and $m = 0$ giving

$$S(\tau, k) = \sqrt{4\pi} a_{00}(\tau, k). \tag{3.2}$$

Substituting (3.2) into rearranged (3.1), for an arbitrary look direction Ω, a direction-dependent error $E_{lm}(\tau, k, \Omega)$ is defined as

$$E_{lm}(\tau, k, \Omega) = a_{lm}(\tau, k) - \sqrt{4\pi} a_{00}(\tau, k) Y^*_{lm}(\Omega), \tag{3.3}$$

which leads to the cost function

$$\Psi(\tau, k, \Omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \mid E_{lm}(\tau, k, \Omega) \mid^2 . \tag{3.4}$$

The corresponding optimized DOA $\Omega_{aiv}(\tau, k)$ is

$$\Omega_{aiv}(\tau, k) = \arg\min_{\Omega} \Psi(\tau, k, \Omega). \tag{3.5}$$

To form the AIV, the optimized direction $\Omega_{aiv}(\tau, k)$ is combined with the norm of the original PIV in (2.17)

$$\mathbf{I_{aiv}}(\tau, k) = -\mathbf{u_{aiv}}(\tau, k)\|\mathbf{I_{piv}}(\tau, k)\|, \tag{3.6}$$

where $\mathbf{u}_{aiv}(\tau, k)$ is the Cartesian unit vector of $\Omega_{aiv}(\tau, k)$.

Figure 3.1 shows an example which demonstrates that $\Psi(\tau, k, \Omega)$ is non-convex. However, calculation of the cost function (3.4) over all possible directions at each TF bin is computationally expensive. At first a theoretical error analysis in the presence of noise is provided and then two grid search and gradient descent approaches, both of which use the PIV solution to form an initial estimate for optimization, are presented.

3.2.2 DOA Error Analysis

This Section presents an analysis of theoretical DOA error for AIV. For the formulation in this Section, (τ, k) are omitted for notational simplicity. As proven in Appendix A, for a close-to-zero inclination error, the azimuth error can be written as a function of SNR, eigenbeams and $\Lambda_{lm}(S, \Omega_u, n_{lm}) = \angle S + \angle Y^*_{lm}(\Omega_u) - \angle n_{lm}$ that is the combined phase from the direct path and the noise eigenbeams.

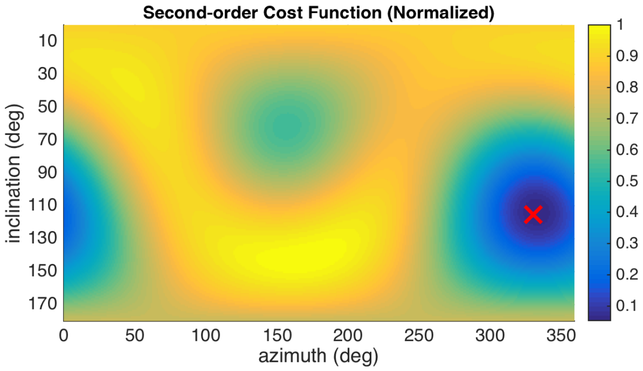


Figure 3.1: Normalized second-order cost function for the entire space at a particular TF-bin for a single source with true DOA marked by a red cross, $T_{60} = 0.5$ s and sensor noise level with SNR=20 dB.

For up to the first SH order ($L = 1$), using (A.10), the azimuth error is

$$\triangle\varphi_s = \arctan\left(\frac{B_1}{A_1}\right), \tag{3.7}$$

where A and B are defiend in (A.11) and (A.12) respectively.

Figure 3.2 presents the azimuth error $\triangle\varphi_s$ in degrees for all possible combinations of Λ_{lm} for $L = 1$ with 1° angle resolution for varying SNR with a random true DOA $\Omega_u = (\phi_u, \theta_u) = (20, 45)^\circ$ and $\mid S \mid= 1$. Since the overall behaviour of the function is smooth, the choice of such coarse angle resolution does not affect the outcome. It clearly shows the decrease in maximum error as the SNR increases.

For up to the second order ($L = 2$), (A.10) can be simplified into a quartic equation with one variable $\sin(\triangle\varphi_s - \arctan(B_1/A_1))$ and parameters as a function of Aj and B_j. Among the real roots of the quartic equation, the one with the minimum $\triangle\varphi_s$ is considered. Figure 3.3 presents the median of azimuth errors $\triangle\varphi_s$ in degrees across all possible combinations of Λ_{lm} with 1° angle resolution for varying SNR and $L = \{1, 2\}$ with the same true DOA as in Fig. 3.2. It can be clearly seen that the increase in the maximum SH order of AIV cost function at least doubles the expected accuracy.

3.2.3 Grid search optimization

In the discrete spatial domain sampled with 1° resolution across azimuth and inclination, the search domain is defined as the set of look directions $\{\Omega_M\}$ covered within a spherical cap with a chosen radius centred at the initial DOA estimated by PIV. Note that larger

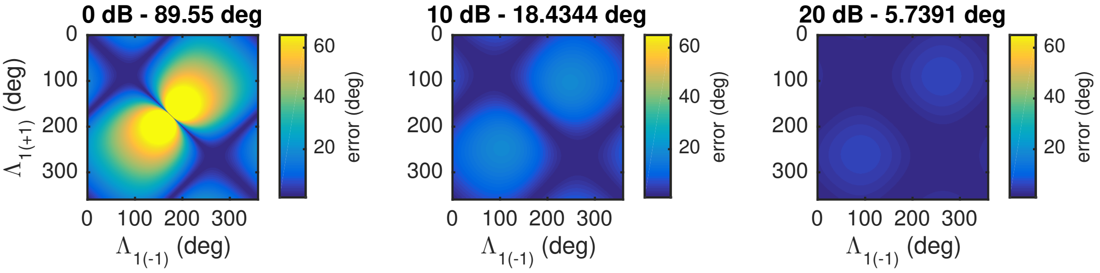


Figure 3.2: Azimuth error for all possible combinations of Λ_{lm} for $L = 1$. The title contains the SNR=$\{0, 10, 20\}$ dB and the worst error.

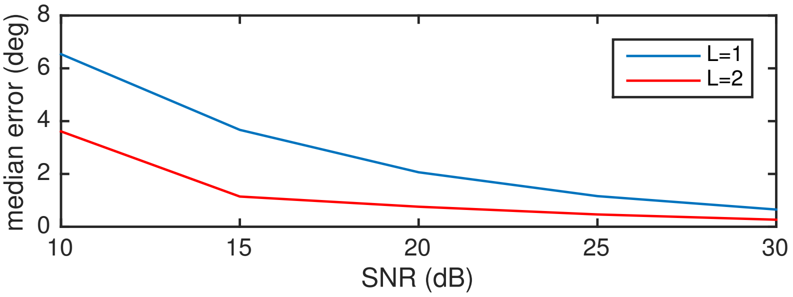


Figure 3.3: Median error among all possible combinations of combined phases for varying SNR and maximum SH order L.

search window size and higher grid resolution both increase the accuracy of estimation as well as the computational cost. The optimized DOA $\Omega_s(k)$ is obtained using (3.5) for $\Omega \in \{\Omega_M\}$. The performance of this approach is investigated in Section 3.3.1.

3.2.4 Gradient descent optimization

The grid search approach in 3.2.3 suffers from limited spatial domain or high computational cost for small or large window sizes respectively. A gradient descent approach can be used to overcome the problem of spatial limitation with low computational cost. Starting from the initial angle Ω_0, using the objective cost function in (3.4) an iterative gradient descent can be formulated as

$$\Omega_{n+1}(k) = \Omega_n(k) - \gamma_n(k)\nabla\Psi(k,\Omega_n), \quad n \geq 0 \tag{3.8}$$

where γ_n is the step at the n^{th} iteration and $\nabla(.) = \frac{\partial}{\partial\theta}(.)\hat{\theta} + \frac{\partial}{\partial\varphi}(.)\hat{\varphi}$ denotes the gradient operator.

For a convex cost function, convergence to a global minimum can be guaranteed. However when there are multiple active talkers or an active direct acoustic path plus one or more active reflections at the same TF bin, $\Psi(k,\Omega)$ is nonconvex. In this case, convergence to a global minimum is only achieved if the initial point is close enough to the global minimum.

The gradient at angle $\Omega = (\theta, \varphi)$ can be found by substituting (3.3) into (3.4) giving

$$\begin{aligned}
\nabla\Psi(k,\Omega) &= \nabla\{\sum_{lm}^{L} |E_{lm}(k,\Omega)|^2\} \\
&= \sum_{lm}^{L} \nabla\{|a_{lm}(k) - \sqrt{4\pi}a_{00}(k)Y^*_{lm}(\Omega)|^2\} \\
&= \sum_{lm}^{L} \nabla\{|a_{lm}(k)|^2 + |\sqrt{4\pi}a_{00}(k)|^2|Y^*_{lm}(\Omega)|^2 \\
&\quad - 2|a_{lm}(k)||\sqrt{4\pi}a_{00}(k)||Y^*_{lm}(\Omega)|\cos\left(\lambda_{lm}(k) - \angle Y^*_{lm}(\Omega)\right)\} \\
&= 4\pi|a_{00}(k)|^2\sum_{lm}^{L}\left\{\nabla\{|Y^*_{lm}(\Omega)|^2\}\right\} - 2\sqrt{4\pi}|a_{00}(k)| \\
&\quad \times \sum_{lm}^{L}\{|a_{lm}(k)|\nabla\{|Y^*_{lm}(\Omega)|\cos\left(\lambda_{lm}(k) - \angle Y^*_{lm}(\Omega)\right)\}\}, \quad (3.9)
\end{aligned}$$

where $\sum_{lm}^{L} = \sum_{l=0}^{L}\sum_{m=-l}^{l}$ is the summation over all the harmonic orders and degrees up to the maximum order L, $\lambda_{lm} = \angle a_{lm} - \angle a_{00}$. The gradient of the components in the final expression in (3.9) can be calculated individually for each harmonic order and degree using (2.2) as shown in Tables B.1 and B.2 in Appendix B.

3.2.5 DOA extraction from Intensity Vectors

Considering either PIVs or AIVs, the intensity vectors are calculated for all TF bins. A 2D histogram (inclination vs azimuth) using the quantized directions of all intensity vectors is formed. Note that only the directions of the intensity vectors are used and not their vector length. As shown in [17] in case of multiple arriving plane-waves in a TF bin, it is possible to have an erroneous resulting intensity vector with direction in between or away from the sources and a norm higher than the intensity vector norm in the presence of a single source depending on the relative amplitude and phase of the impinging plane waves. In order to avoid the accuracy-loss effect of the erroneous intensity vector with high norm, the norms are ignored and only the cardinality of the quantized directions are considered in the histogram. An advantage of the histogram is to eliminate the weakening impact of the erroneous directions with low cardinality on the position of the peaks in the histogram.

Hence it is preferred over the averaging technique in [16] in which all intensity vectors are summed to estimate the final DOA where erroneous directions reduce the accuracy if they are not spatially diffused.

Due to noisy observations and the presence of multiple irregular peaks, the constructed DOA histogram is smoothed using a Gaussian kernel. The Gaussian kernel, centred on the look direction Ω, for an angle $\Omega_{\theta_i,\varphi_i}$ with inclination θ_i and azimuth φ_i is expressed as

$$\xi(\Omega, \Omega_{\theta_i,\varphi_i}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\angle(\Omega, \Omega_{\theta_i,\varphi_i})^2}{2\sigma^2}\right), \tag{3.10}$$

where σ denotes the standard deviation, which is chosen empirically as described in Section 3.3. The kernel is truncated by removing the entries with $\xi < 0.001$. For N_s sources, the positions of the largest N_s peaks in the smoothed histogram are taken as the estimated DOAs. Figure 3.4 shows an example of unsmoothed (raw) and smoothed histograms for two sources with 45° separation with simulation configuration as in Section 3.3.2. The choice for the σ, which represents the smoothness of the histogram, is studied in [48], which concludes that a suitable σ requires the knowledge of the minimum angular separation of the sources and the choice of $3 \leq \sigma \leq 6$ was shown to provide robust and well distinguished peaks for $\geq 30°$ separation.

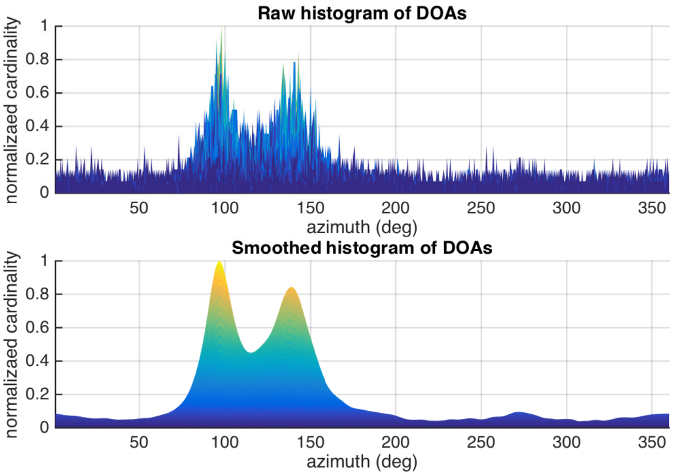


Figure 3.4: An example side view of the raw and the smoothed 2D histograms of the estimated narrow-band DOAs with $\sigma = 4°$.

3.3 Evaluation

The proposed DOA estimation algorithms are evaluated in terms of their accuracy and robustness to RT [63], sensor noise level, number of sources, and angular separation of sources using simulated data for one talker and for multiple simultaneous talkers. For all methods, a sampling frequency of 8 kHz was used with a STFT window size of 8 ms and 50% overlap of time frame. The processing band was set to 500 Hz to 3850 Hz to avoid spatial aliasing and ensuring $kr < L$ for $L = 3$ as in [64, 14] and to avoid excessive noise amplification due to mode strength compensation at lower frequencies.

The proposed methods, denoted AIV Grid Search (AIV-GS) and AIV Gradient Descent (AIV-GD), are compared to the previously presented PIV, SSPIV, PWD-SRP as the baseline methods and both variations of DPD-MUSIC as the state-of-the-art. For AIVs, the effect of the choice of maximum SH order $L = 2, 3$ is evaluated. Accordingly, the evaluated algorithms are denoted AIV-GS2, AIV-GS3, AIV-GD2 and AIV-GD3. For the rest of the evaluation $L = 3$ is considered for the methods using high order SHs.

In order to compare the narrowband PWD-SRP with our AIV-GS under the same spatial limitation, Spatially Constrained (SC)-SRP is employed which uses the same search window as AIV-GS and PIV as its centre of window. The employed SC-SRP is the generalized variation of the proposed method in [65] in which SC-SRP is applied only on the TF bins with an active single source unlike our SC-SRP which is applied on all TF bins. The radius of spherical cap search window in AIV-GS and SC-SRP was set to 10° as, in our experiments, more than 95% of PIVs were within 10° of the true DOA. For AIV-GD, the optimization function 'fminunc' based on 'trust-region' algorithm from MATLAB Optimization Toolbox$^{\text{TM}}$ was used and was set to be terminated if the new angle is less than 0.5° (maximum error with 1° spatial resolution) away from the current angle or if the number of calls to the cost function exceeds 100. These termination conditions were determined empirically. The approximate covariance matrix $\mathbf{R_a}$ in (2.21) used by SSPIV and DPD-MUSIC had an averaging window with $J_\tau = 6$ and $J_k = 4$ over time and frequency respectively giving 500Hz and 32ms of window size in the TF domain based on our frequency and time resolution. The threshold ϵ in (2.28) for DPD-MUSIC was empirically set to 6, which is also the choice in its original paper [14]. The DOA histogram and MUSIC spectrum were constructed with 1 ° resolution along inclination and azimuth (181 × 360 points respectively). In DOA histogram smoothing, the kernel had the standard deviation of $\sigma = 4^\circ$ which was chosen empirically from a range of 2° to 6° giving the lowest mean error.

3.3.1 Single Source

A dataset using simulated AIRs convolved with anechoic speech for a single source was used in this section. The details of configuration can be found in Section 2.11.1. The results are presented in two parts. In the first a comparison of the second- and third-order of two variations of our methods, grid search and gradient descent AIVs with PIV method are presented. In the second, the most accurate AIV approach is compared with the baseline and the state-of-the-art methods.

Quantification of the improvements due to higher order spherical harmonics

Figure 3.5 shows the mean DOA estimation error for each method as a function of T_{60} for all SNRs. As expected due to utilization of higher spatial resolution from higher SHs, the AIV approaches significantly outperform PIV for all T_{60} and SNRs. AIV approaches also show noticeably more robustness to reverberation and noise. Comparing AIV-GS and AIV-GD, the advantage of gradient descent becomes noticeable as the noise increases. This is due to spatial freedom of search for gradient descent since the AIV-GS's search window centred on PIVs, which are prone to noise, are more likely to not include the global minimum of cost function. Moreover, the results demonstrate that the increase in SH order (2 vs 3) has a larger impact on improvement of accuracy and robustness than the change in optimization method (GS vs GD) highlighting the higher importance of the cost function quality over the optimization method used for it.

Comparison with baselines and state-of-the-art

In this section our AIV-GD and AIV-GS are compared with SSPIV, SC-SRP and incoherent DPD-MUSIC.

As shown in Figure 3.6, AIV-GD shows the second best accuracy of $\leq 2^{\circ}$ after DPD-MUSIC. The worst performance is by PIV as it only uses the low order eigenbeams. Although SSPIV uses the same formulation as PIV, it performs significantly better than PIV as it employs high order SH in SVD to estimate de-noised low order eigenbeams. SC-SRP and AIV-GS outperform the previous two methods due to utilization of high order eigenbeams but perform very similar to each other as they use the same eigenbeams, search window and TF bins although their formulations differ. AIV-GD outperforms all previously stated methods due to the use of high order eigenbeams, compared to PIV and SSPIV, and lack of spatial limitation compared to SC-SRP and AIV-GS. DPD-MUSIC leads in the performance with 0.5° accuracy due to utilization of denoised high order eigenbeams using SVD, spatial freedom due to full grid search and DPD test which estimates reliable TF bins in which accuracy of DOA estimation would be high.

In terms of robustness to noise and reverberation, subspace techniques such as SSPIV and DPD-MUSIC lead as they take advantage of decomposition of noisy eigenbeams

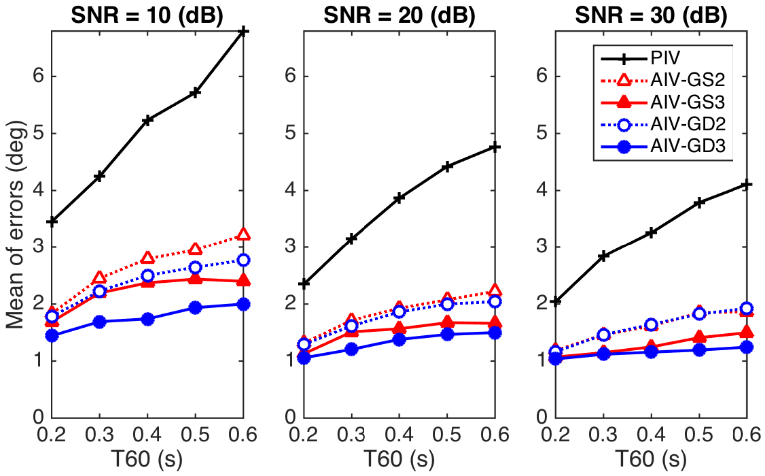


Figure 3.5: Effect of T_{60} and SNR on mean DOA estimation error for PIV and AIV methods in the simulated single source scenario.

into signal and noise subspaces. Although AIV-GD uses noisy eigenbeams, it shows almost as strong robustness as subspace techniques due to its spatially-unconstrained optimization which minimizes the effect of noise on the optimized DOA estimate.

3.3.2 Multiple Sources

This Section uses the testbed dataset introduced in Section 2.11.2 to evaluate the effect of reverberation, number of sources and angular separation of the sources in the multiple source scenario.

Experiment 1

The effect of T_{60} is evaluated here for the illustrative case with $N_s = 2$ and $\Delta\phi_s = 45°$. Figure 3.7 shows the distribution of DOA estimation errors of all methods for $T_{60} = \{0.2,\, 0.4\,, 0.6\}$ s. The black dot in each box shows the median while the boxes show the upper and lower quartiles, and the whiskers extend to 1.5 times the interquartile range. As studied in [17], the accuracy of PIV is severely degraded in strong reverberation when multiple sources are simultaneously active. This can be seen in Figure 3.7 where PIV median error increases from 2.2° at 0.2 s to up to 8.7° (out of y-axis limit) at 0.6 s as

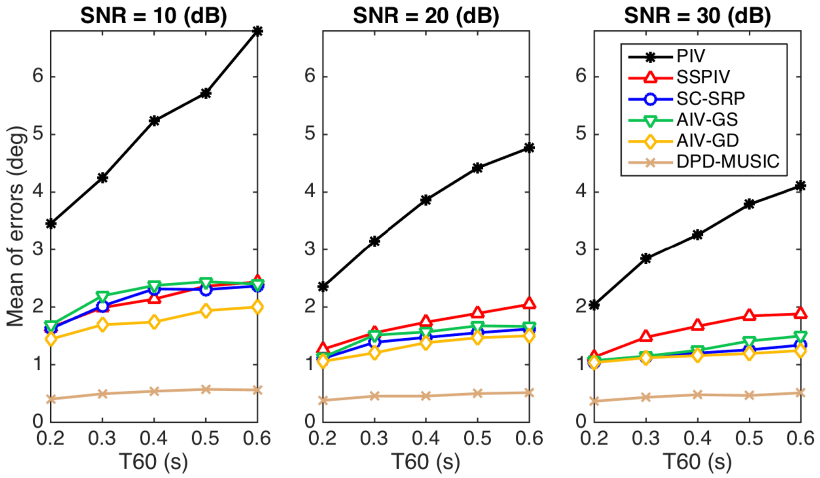


Figure 3.6: Effect of T_{60} and SNR on mean DOA estimation error for the proposed, baseline and state-of-the-art methods in the single source scenario.

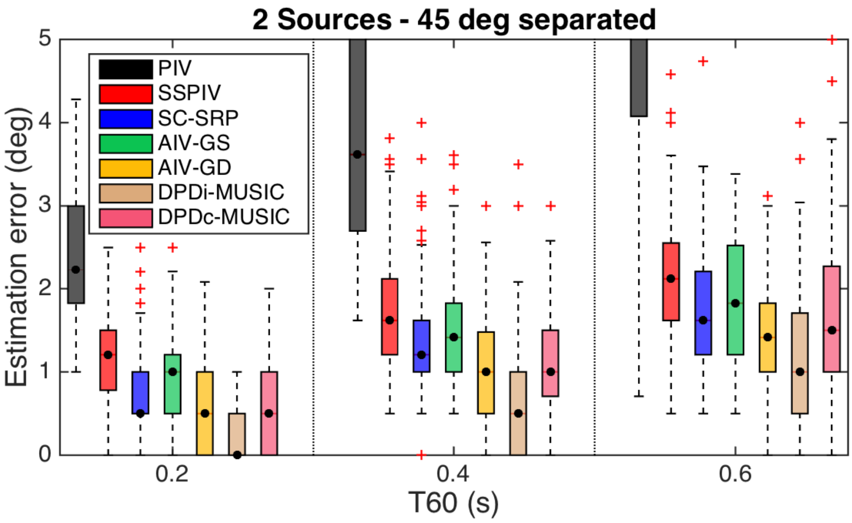


Figure 3.7: Distribution of DOA estimation errors for two sources 45° apart with varying T_{60}.

T_{60} increases. AIV-GD, after incoherent DPD-MUSIC, shows the second best accuracy with median errors of $\{0.5,\, 0.9,\, 1.4\}^\circ$ and robustness to reverberation of 1° similar to DPDc-MUSIC for all T_{60}s. Although DPDi-MUSIC leads in accuracy it shows the same robustness to reverberation as others as its median error increases by around 1° from lowest (0.2 s) to highest (0.6 s) T_{60}. The DPDc-MUSIC provides less accuracy than DPDi-MUSIC since DPDc-MUSIC is more prone to TF bins with erroneous signal space due to high sensitivity of clustering to outliers.

The results for $T_{60} = 0.4$ s are used for inferential statistical analysis. There was a statistically significant difference between groups as determined by one-way ANOVA ($F(6, 742) = 63.42, p < 0.001$). A Tukey post hoc test revealed that the estimation error was statistically significantly lower for AIV-GD ($0.95^\circ \pm 0.19^\circ, p < 0.001$) compared to PIV method ($5.39^\circ \pm 0.19^\circ$).

Experiment 2

The effect of the number of sources and the angular separation of the sources is evaluated for incremental N_s from 2 to 5 sources with $\Delta\phi_s = \{30,\, 45\,, 90\}^\circ$ for up to 4 sources and $\Delta\phi_s = \{30,\, 45\}^\circ$ for 5 sources with $T_{60} = 0.4\,$s. The performance of each method is evaluated using mean NoDS and mean error as defined in 2.10.

Figure 3.8 shows the mean errors and mean NoDS for all methods with incremental N_s and varying $\Delta\phi_s$. In terms of accuracy, AIV-GD shows the second best performance with the worst mean error of 1.8° after DPDi-MUSIC with the worst mean error of 1.0°. AIV-GS performs as accurately as SC-SRP with the worst mean error of 2.1° as they both utilize the same eigenbeams, initial DOA estimates and search window although differ in cost function. DPDc-MUSIC shows noticeable accuracy loss of 2.0° for adjacent sources with $\Delta\phi_s = 30°$ for all the values of N_s. In contrast to the results in the original work [14] on DPD-MUSIC where the sources are widely separated by 60° and 70°, in scenarios with lower separation, as in our evaluation, DPDc-MUSIC, compared to DPDi-MUSIC, does not show a better accuracy. This is caused as the clustering in DPDc-MUSIC becomes highly prone to adjacent sources and results in the merge of two adjacent clusters of signal

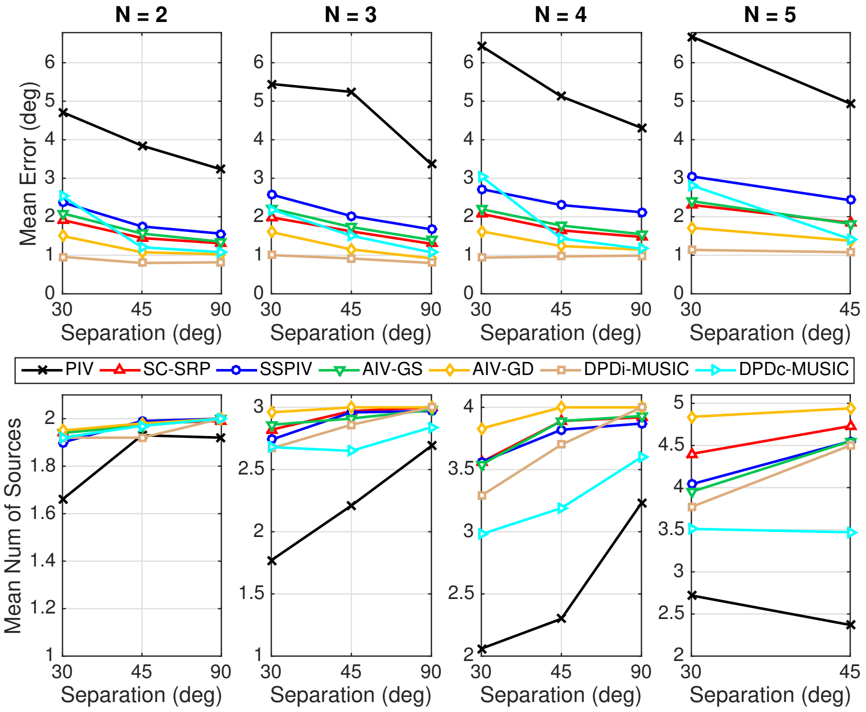


Figure 3.8: Mean error and Mean NoDS for incremental N_s and varying $\Delta\phi_s$ with $T_{60} = 0.4\,s$.

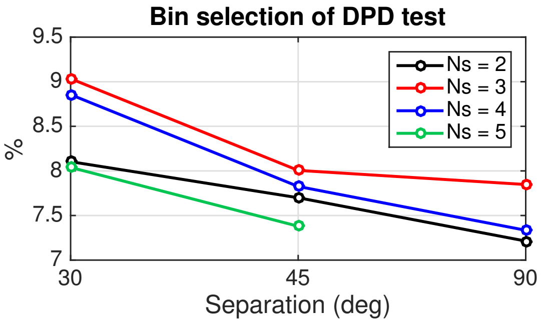


Figure 3.9: Percentage of the bins passed in DPD test for varying N_s and $\Delta\phi_s$.

subspaces giving erroneous centroid signal subspace especially in the presence of outliers signal subspaces.

In terms of mean NoDS, AIV-GD has the highest robustness to angular separation and number of sources. Apart from PIV, which generally performs poorly in the multi-source scenario in all cases, SSPIV, SC-SRP, AIV-GS and AIV-GD show more robustness to the number of sources than DPD-based methods. Both DPD-MUSIC variations significantly lose mean NoDS as the number of sources increases due to the static threshold for SVR in the DPD test which causes the reduction of number of TF bins that passed the test with the increase of number of sources. Figure 3.9 presents the percentage of the passed TF bins in the DPD test for an incremental number of sources. As expected the percentage reduces with the increase of N_s as the likelihood of a dominant single source in a bin drops. It can also be observed that the percentage increases as the angular separation of the sources decreases since the likelihood of strong unity effective rank in (2.28) increases as the two adjacent sources tend to be considered as a single erroneous intermediate dominant source. Although an increase in source separation decreases the percentage of the passed bins, the accuracy and robustness increase as shown in Fig. 3.8 due to having fewer erroneous passed TF bins.

3.4 Experimental Verification

To demonstrate the performance of each method in a real-world scenario, the methods are evaluated using real recordings in a real reverberant room. The details of configuration can be found in Section 2.11.3.

Figure 3.10 shows the normalized smoothed histogram for PIV, SSPIV, SC-SRP, AIV-GS and AIV-GD as well as normalized MUSIC spectrum for incoherent DPD-MUSIC. Due to approximate knowledge of the position of sources and array, accurate numerical estimation error cannot be obtained. The approximate mean estimation error for all methods is 3° except PIV which is 4.5°. Although, because of well angular separation, all methods successfully estimate peaks corresponding to all four sources, the measure of 'peak strength', as defined in 2.10, clearly present the relative performance of the methods. Table 3.1 presents the peak strength of each peak for all methods.

AIV-GD and SSPIV lead as they both estimate the most prominent peaks. AIG-GS and SC-SRP performs similarly as explained in previous sections. SSPIV, due to noise suppression in eigenbeams by sub-space decomposition, manages to successfully estimate accurate DOAs in the majority of TF bins where PIV estimates an erroneous DOA. The performance improvement of AIV-GD compared to AIV-GS, due to utilization of spatially unconstrained optimization, is clearly observable in Fig. 3.10 and in Table 3.1 as erroneous DOAs in AIV-GS, which are mainly distributed between and around the peaks, are more concentrated around the peaks in AIV-GD resulting in sharper peaks. Note that the number of DOA estimates in all methods except DPD-MUSIC are equal. Comparing the sharpness and the sparsity of the histograms of PIV and AIV-GD, there is a significant accuracy improvement by AIV-GD since majority of the erroneous DOAs in PIV histogram have had their accuracy improved in AIV-GD due to employment of high order eigenbeams. DPD-MUSIC shows a poor peak strength in Table 3.1 due to having a very low-height, although sharp, peak (peak 4) as seen in Fig. 3.10. This is caused since an increase in the number of sources results in reduction of the number of passed bins in the DPD test, as previously shown in Fig. 3.9, which can result in having low-height peaks in the MUSIC spectrum and therefore potentially missing a source.

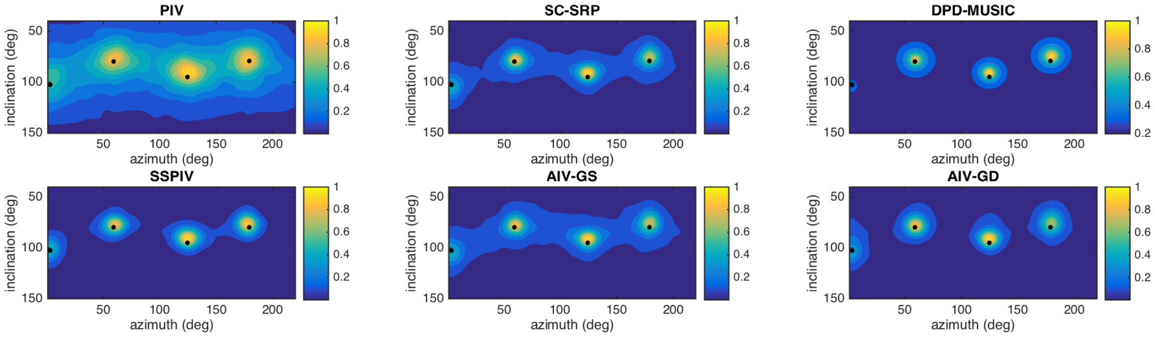


Figure 3.10: Normalized smoothed histogram of PIV, SSPIV, SC-SRP, AIV-GS, AIV-GD and normalized incoherent DPD-MUSIC spectrum using real recording. The black dot represents the approximate true DOA.

3.5 Computational Complexity

This Section discusses the number of computations required in each method for a single TF bin in terms of the number of real multiplications. Note that the multiplication of two complex numbers is counted as four real multiplications while $| \, . \, |^2$ is counted as two real multiplications. The number of multiplications in (2.2) is not included as Y_{lm} can be pre-calculated and stored for all directions in a discrete spatial domain.

For subspace methods, there are $4(L+1)^4 J_\tau J_k$ multiplications to compute the estimated covariance matrix in (2.21) as well as $3(L+1)^6$ for SVD in (2.22).

For PIV, there are 48 $(3\times3\times4+3\times4)$ operations where the numbers in parentheses respectively represent the number of axes, harmonic modes and the real multiplications in (2.18) and the number of axes and the real multiplications in (2.17). For AIV and SRP cost functions using (3.4) and (2.14), there are $48 + (L+1)^2 \times (2+4+2)$ operations for a single look direction where the numbers respectively correspond to the PIV, number of

Peak	PIV	SC-SRP	DPD-M	SSPIV	AIV-GS	AIV-GD
1	2.08	5.66	3.31	6.23	5.24	6.88
2	1.96	4.45	3.04	6.12	4.18	5.37
3	1.82	3.54	2.39	5.32	3.50	3.35
4	0.99	1.50	0.49	2.31	1.46	1.17
Mean	1.71	3.79	2.31	5.00	3.59	4.19

Table 3.1: Peak Strength of each peak for all methods where DPD-M = DPDi-MUSIC.

eigenbeams up to the order L, a real-complex followed by a complex-complex multiplications, and squared magnitude. For DPD-MUSIC, as well as subspace computation there are $4((L+1)^2-1)(L+1)^2$ multiplications for MUSIC spectrum in (2.26) for a single look direction. Coherent DPD-MUSIC is excluded from this consideration due to unknown and non-deterministic complexity in clustering.

An average of 5 iterations for gradient descent in AIV-GD was achieved. Considering numerical gradient using the four neighbouring look directions, the AIV cost function is called 5 times per iteration which results in an average of 25 look directions for AIV-GD. With a spherical cap window of radius 10° for AIV-GS and SC-SRP, there is an average of 100 look directions. MUSIC uses a full grid search of 181×360 look directions.

Using the settings in this Chapter, the overall approximate number of real multiplications of each method per TF bin is as follows: 48 for PIV, 37×10^3 for SSPIV, 13×10^3 for SC-SRP and AIV-GS, 3×10^3 for AIV-GD, and 25×10^7 for DPD-MUSIC assuming an average of 10% of the bins passing the DPD test. Our proposed AIV-GD leads in computation after PIV while the state-of-the-art DPD-MUSIC shows an expensive computational cost due to covariance matrix calculation, SVD and full grid search although it is performed on a small percentage of the total TF bins.

3.6 Conclusions

This Chapter proposed a novel narrowband DOA estimation method for spherical microphone arrays. It uses high order spherical harmonics while avoiding exhaustive search to enhance the accuracy and robustness of DOA estimates in PIV. Two alternative implementations of the method were evaluated, one based on grid search and the other on gradient descent optimization. It is shown that the gradient descent approach shows a better performance in accuracy and robustness compared to spatially limited grid search approach. Simulations and real recording results have been presented for single and multiple sources with different sensor noise levels, reverberation times, number of sources, and angular separation of sources. The results also show that using up to the third order spherical harmonics has significant advantages over second order harmonics for AIV and the increase of order has more impact on accuracy than the choice of optimization technique. For the third-order gradient descent AIV in the presence of realistic reverberation and sensor noise level, the worst average error was 1.5° for single source and 2° for up to 5 sources with down to 30° angular separation. It also outperforms the baseline PIV, SSPIV and SC-SRP. It is shown that AIV-GD leads in terms of robustness to the number of sources and separation. In addition, an analysis of computational complexity indicates that the proposed AIV-GD technique outperforms the state-of-the-art method in terms of computational complexity with a few thousand real multiplications per bin with only less than 1.5° accuracy loss compared to DPD-MUSIC.

Chapter 4

Evolutive density-based source counting

THE conventional procedure of DOA estimation for multiple wideband sources typically consists of two stages: (1) Temporal narrowband DOA estimation in the TF domain and (2) final DOAs extraction from local DOA estimates. The latter stage has a direct impact on the outcome accuracy and robustness, and therefore it is important to be robust to outlier (highly erroneous) DOA estimates. Working in scenarios without knowledge of the number of sources and/or outlier DOAs remains a further on-going challenge in multi-source DOA estimation.

The two main categories of final DOA extraction techniques are (1) conventional peak picking and (2) clustering-based techniques. In conventional peak picking, the peaks with highest cardinality in the smoothed histogram of the DOA estimates are directly [48] or iteratively [31] extracted where the number of sources is known *a priori*. As shown in [48], the required settings for the smoothing function depend on the angular separation of the sources, the noise level and the irregularity of the peaks. With fixed smoothness, low reliability may also be expected in scenarios with varying peak irregularity among different distributions of DOAs. In the second category, a clustering technique such as Kmeans [32] or mixture models using Gaussian [33], Laplacian [34] or Von Mises [35] distributions are applied. These approaches typically require *a priori* knowledge of the source number and/or outlier removal in order to perform reliably. In [37], the authors propose the use of Akaike Information Criterion (AIC) [38] to estimate the number of sources.

Density-based clustering has received much less attention than distance-based or model-based clustering techniques for acoustic DOA estimation. This chapter investigates the use of Density-based Spatial Clustering of Applications with Noise (DBSCAN) [39] to propose Evolutive DBSCAN, an evolutionary framework for DOA clustering and source counting. The term 'Evolutive' refers to the evolutionary process of the birth, growth, death and merge of the clusters over iterative DBSCAN as will be discussed more in depth later in this Chapter. DBSCAN, as opposed to other clustering techniques, does not require *a priori* knowledge of the source number and is robust to outliers, although it is not fully autonomous. It is assumed to have no knowledge of the number of sources and outlier DOAs. The only assumption is that the distribution of accurate DOAs (non-outliers) is not strongly skewed and the outlier DOAs are spatially white.

4.1 DBSCAN Clustering

Unlike distance-based clustering techniques, density-based DBSCAN clustering does not consider the number of clusters to be known *a priori* but instead is based on a user-defined minimum density for a cluster. Therefore DBSCAN considers an assumption on the density of clusters rather than the number of clusters, which makes it robust to noise and suitable for autonomous cluster counting.

The terms used in DBSCAN clustering are defined as follows [39].

Neighbourhood DOAs

The set of neighbourhood DOAs for a DOA estimate $\hat{\mathbf{p}}$ is defined as

$$N_\varepsilon(\hat{\mathbf{p}}) = \{\hat{\mathbf{q}} \in U | \angle(\hat{\mathbf{p}}, \hat{\mathbf{q}}) \leq \varepsilon\}, \tag{4.1}$$

where $\angle(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ is the angular separation (in degrees) between two DOA estimates $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ and ε is chosen to define the angular extent of the neighbourhood in degrees.

Density

The density at a DOA estimate $\hat{\mathbf{p}}$ is defined as the number of DOA estimates (including $\hat{\mathbf{p}}$ itself) within its neighbourhood $|N_\varepsilon(\hat{\mathbf{p}})|$, where $|.|$ indicates cardinality.

Threshold density

The threshold density denoted as MinPts is the minimum density for a potential cluster.

Directly density-reachable

A DOA estimate $\hat{\mathbf{p}}$ is directly density-reachable from another DOA estimate $\hat{\mathbf{q}}$ if

- $\hat{\mathbf{p}} \in N_\varepsilon(\hat{\mathbf{q}})$ and
- $|N_\varepsilon(\hat{\mathbf{q}})| \geq \text{MinPts}$ (core point condition).

Density-reachable

A DOA estimate $\hat{\mathbf{p}}$ is density-reachable from another DOA estimate $\hat{\mathbf{q}}$ if there is a chain of DOA estimates $\{\hat{\mathbf{p}}_i\}_{i=1}^{L}$, where $\hat{\mathbf{p}}_1 = \hat{\mathbf{q}}$ and $\hat{\mathbf{p}}_L = \hat{\mathbf{p}}$, such that $\hat{\mathbf{p}}_{i+1}$ is directly density-reachable from $\hat{\mathbf{p}}_i$.

Density-connected

A DOA estimate $\hat{\mathbf{p}}$ is density-connected to another DOA estimate $\hat{\mathbf{q}}$ if there is a DOA estimate $\hat{\mathbf{m}}$ such that both $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are density-reachable from it.

Cluster

A cluster S is a subset of U satisfying:

- $\forall \hat{\mathbf{p}}, \hat{\mathbf{q}}$: if $\hat{\mathbf{p}} \in S$ and $\hat{\mathbf{q}}$ is density-reachable from $\hat{\mathbf{p}}$, then $\hat{\mathbf{q}} \in S$ and
- $\forall \hat{\mathbf{p}}, \hat{\mathbf{q}} \in S$: $\hat{\mathbf{p}}$ is density-connected to $\hat{\mathbf{q}}$.

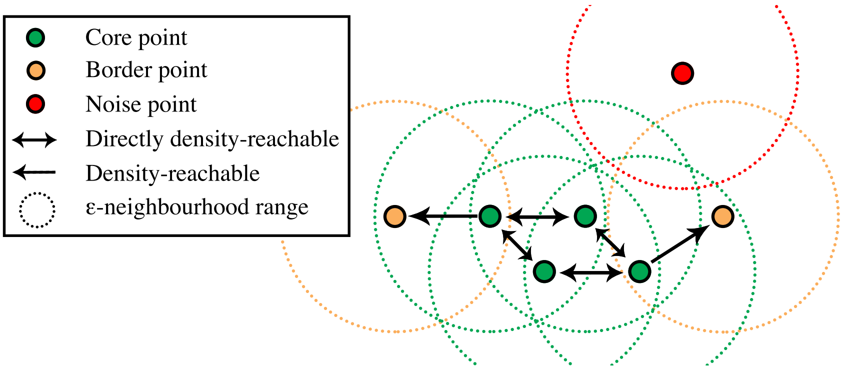


Figure 4.1: DOA estimates labelling by DBSCAN with MinPts=3.

Noise

A subset of DOA estimates in U not belonging to any cluster.

Figure 4.1 illustrates the labelling of an example of several DOA estimates by DBSCAN with MinPts=3. Each core point (green) has at least three DOAs including itself within its ε-radius neighbourhood while the border (orange) and the noise (red) DOAs do not satisfy the core point condition.

Given the user-defined parameters ε and MinPts, the algorithm first detects all the core points. A single cluster is identified in two steps: (1) start from an arbitrary core point and (2) retrieve all points which are density-reachable from it. It then visits the next un-clustered core point and repeats this process until all core points are clustered. The points which do not belong to any cluster are labelled as noise. Pseudocode can be found in [39].

The performance of DBSCAN highly depends on the choice for ε and MinPts pair. In [39], a simple heuristic method is proposed to find the appropriate choice for ε and MinPts from the least dense cluster. Defining k-dist as the distance to the kth nearest neighbour for a point, ε is set to the k-dist of the knee in the graph of sorted k-dist values of all points. Although the tuning is simplified, it is not automatic and the appropriate choice of k depends on the noise domain.

DBSCAN loses reliability with distributions of varying densities as there may not be a value for MinPts, given ε, by which all densities are individually clustered. As

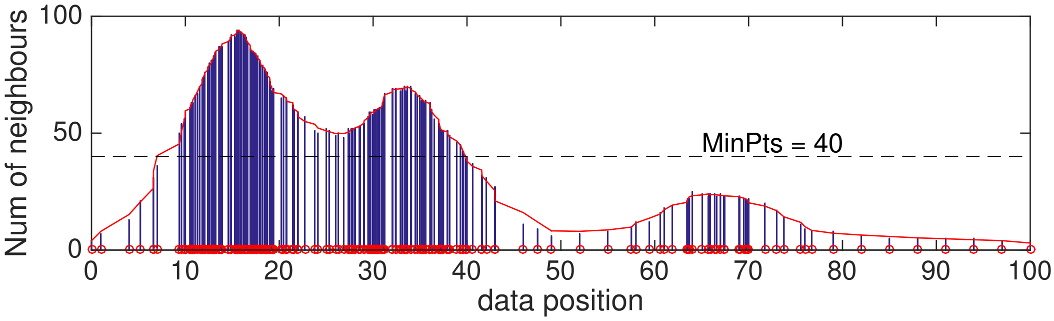


Figure 4.2: Number of neighbours ($|N_6(.)|$) vs data position for an example of 1D data distribution. Data points are marked as red circles on x-axis. The horizontal dashed line, with MinPts as its vertical position, distinguishes the core points.

shown in Fig. 4.2, any choice of MinPts leads either to the erroneous merging of adjacent densities or the missing of the least dense distribution. Mixtures of distributions with widely varying density often occur in DOA estimation especially in multi-source acoustic scenarios where one source is less active or relatively further with respect to the microphone array compared to other sources. Variations of DBSCAN [66, 67, 68, 69, 70] are proposed but all require user engagement for setting parameters.

4.2 Evolutive DBSCAN

In the context of final DOAs extraction from local DOA estimates, due to potential variations among the clusters density, the proposed method runs DBSCAN for various MinPts and iteratively stores the reliable centroids and their associated weights. The distribution of the resulting weighted centroids is shown to be significantly more sparse and less noisy than the distribution of the DOAs. Thus, it allows for more reliable cluster counting.

The definition of density, $|N_\varepsilon()|$, from (4.1) is used. In order to have a consistent metric of density across the entire method, a fixed value of ε is defined. This value must be less than half of the minimum separation of the sources and, since in the example of robot audition sources are normally separated by $> 20°$, $\varepsilon = 10°$ was found to be a reasonable choice. Note that ε is defined as a base unit for density in the proposed algorithm and

Algorithm 1 Evolutive DBSCAN

```
function EVOLUTIVE_DBSCAN(points)
    centroids=[]; %holds alive centroids and weights
    MinPts=max($|N_\varepsilon(.)|$) $-$ 1; cntr=1;
    while (MinPts$\geq$ min($|N_\varepsilon(.)|$) $+$ 1) OR (cntr$\leq$NumIt)
        C=DBSCAN(points,$\varepsilon$,MinPts);
        if isEmpty(centroids) then
            centroids += C(all).centroid; %initialization
            C.dead_members=[]; %all dead members
        else
            C=LABEL_CLUSTERS(C,C_last);
            if anyClusterAlive(C) then
                centroids += C(alive_ones).centroid;
                C=RemoveDead(C,dead_ones);
            end if
        end if
        C_last=C; MinPts -= step; cntr += 1;
    end
    return centroids
end function
```

is not a user defined parameter. Our experiments showed that varying ε in a relatively small range does not significantly change performance as the density at each point changes proportionally to the change in ε.

The density $|N_\varepsilon(.)|$ is calculated for all DOA estimates using (4.1) where the distance between two DOA estimates unit vectors $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ with angular separation $\angle(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ (in radians) is

$$\text{dist}(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = 1 - \cos\left(\angle(\hat{\mathbf{p}}, \hat{\mathbf{q}})\right). \tag{4.2}$$

The choice of ε automatically sets a boundary on the possible values for MinPts in order to have meaningful clustering. The maximum and minimum $|N_\varepsilon(.)|$ in the dataset respectively indicate the upper and lower bounds on MinPts since for MinPts $> \max(|N_\varepsilon(.)|)$ there is no core point and therefore no cluster while for MinPts $\leq \min(|N_\varepsilon(.)|)$ all points are core points grouped as one cluster. This leads to MinPts$\in [\min(|N_\varepsilon(.)|)+1, \max(|N_\varepsilon(.)|)-1]$. The step-size for searching MinPts can be either defined by the user or calculated based on the maximum number of iterations (NumIt) specified by the user. Best resolution is obtained for MinPts steps of 1. Starting from highest to lowest MinPts, at each iteration after a comparison between the current clustering and the previous clustering, the current clusters are labelled as 'dead' or 'alive' each defined as follows:

1) Dead: A cluster is dead if:

- it has a shared member with more than one alive cluster in the last iteration (merge condition) or
- it has a shared member with any previously dead cluster (re-occurrence of a previously merged cluster).

2) Alive: A cluster is alive if it is not dead.

At each iteration the weight and the centroid of the alive clusters are stored where weight is defined as the mean density of the cluster's members. The pseudocode for the main part of the algorithm is presented in Algorithm 1.

4.3 Source counting and DOA extraction

Having obtained M centroids $\{c_i\}_{i=1}^{M}$ and their associated weights $\{w_i\}_{i=1}^{M}$, an autonomous final DBSCAN is performed on the centroids to count the number of detected clusters, L, and their final centroids $\{d_i\}_{i=1}^{L}$ indicating the estimated number of sources and the final DOA estimates respectively, as shown in Fig. 4.3(c). Assuming the distribution of the initial DOA estimates as a mixture of unskewed distributions with additive spatially white noise DOAs, it is expected to have very low spatial variance for the centroids belonging to a repeatingly alive cluster at consecutive iterations. Therefore a smaller value of $\varepsilon_f = 5^\circ$ is defined in the final DBSCAN while MinPts is autonomously determined as follows. Instead of a sorted k-dist graph in [39], a sorted weighted-density graph is built from the centroids $\{c_i\}_{i=1}^{M}$ using their weight $\{w_i\}_{i=1}^{M}$ and their density $\{|N_{\varepsilon_f}(c_i)|\}_{i=1}^{M}$. The advantage of density over k-dist as the metric is the consistency in the behaviour of the sorted density graphs regardless of ε_f, as opposed to sorted k-dist, the behaviour of which highly depends on k. Also using the weights exaggerates the dynamic range and so also the angle of the 'knee' in the graph since the outlier centroids are expected to be from low density clusters of outlier DOA estimates that might have been clustered due to low MinPts at the end of evolutionary process. MinPts is the density at the position of the 'prominent' knee with the lowest weighted density. This is determined as the position of the first peak (excluding the peaks with less than 10% of the highest peak) in its derivative function as shown in

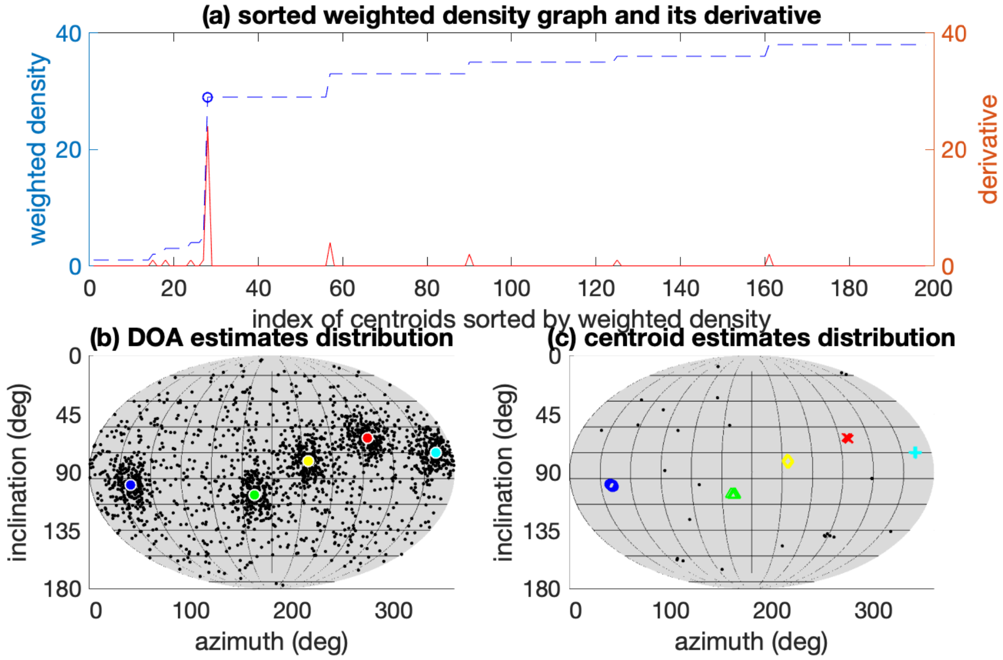


Figure 4.3: (a) Sorted weighted density (blue dashed) graph and its derivative (solid red). Position of the knee marked as blue circle. Distribution of (b) DOA estimates (c) centroid estimates for 5 sources. Final centroids are marked by coloured circles.

Fig. 4.3(a). If $\min(\{|N_{\varepsilon_f}(c_i)|\}_{i=1}^{M}) > 1$, no knee detection is performed and MinPts is the minimum density.

4.4 Evaluation

The algorithm's performance was evaluated using generated and estimated DOAs. Performance metrics SLR and mean error, as defined in 2.10, respectively represent the source counting reliability and DOA estimation accuracy.

For N_s sources, 100 trials per experiment were used. The first true DOA was randomly selected from uniform distribution around the sphere and the subsequent true DOAs placed with Sep separation on the randomly-orientated great circle passing through the first true DOA. Our algorithm was compared with conventional peak picking from the smoothed histogram, original DBSCAN based on k-dist graph [39] as well as Kmeans and its adaptive variations. Peak picking and Kmeans, with $K = N_s$, use prior knowledge of

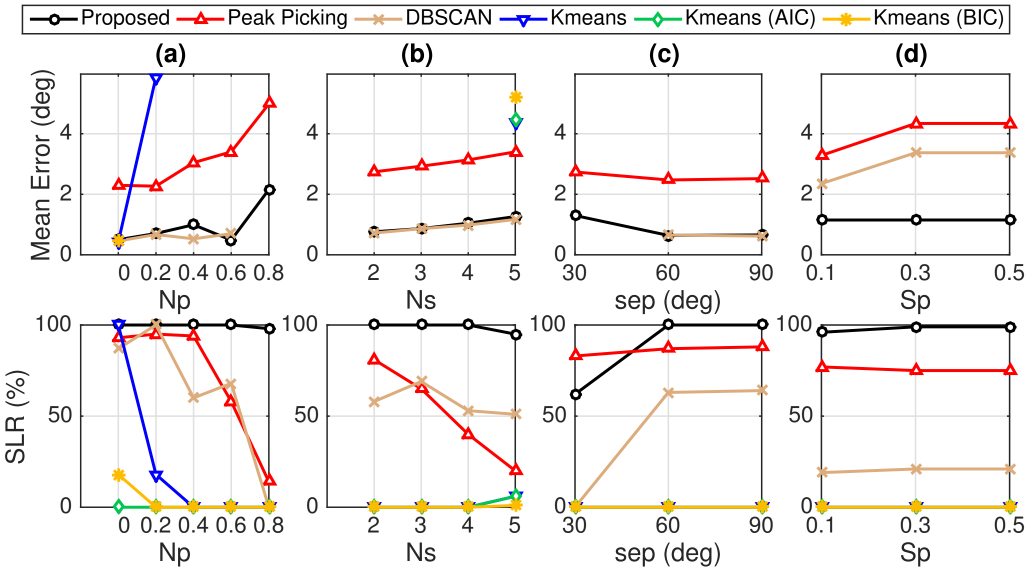


Figure 4.4: Evaluation results using generated DOAs as a function of (a) N_p, (b) N_s, (c) Sep and (d) S_p. Note that the lines and markers for Kmeans and its variations are mostly overlaid by each other as they share zero value for SLR. If SLR is zero, then no mean error is calculated as it is averaged only among the successful localized trials.

the number of sources. Adaptive Kmeans [47] estimates the optimum K by minimising AIC or Bayesian IC (BIC) for $K = \{1, ..., 5\}$. A Gaussian kernel with empirically-chosen standard deviation of 4° was used for peak picking. The original DBSCAN had empirically-chosen $k = 10$ for k-dist graph, which gave the best overall results. NumIt=50 was found a reliable and computationally efficient choice for the proposed Evolutive DBSCAN method.

4.4.1 Evaluation using generated DOAs

The testbed dataset introduced in Section 2.11.4 is used in this evaluation. Figure 4.4 shows that the proposed method significantly outperforms the comparative methods both in mean error and SLR in all cases except for SLR at angular separation $Sep = 30°$. Kmeans for $N_p < 0.4$ and peak picking show better performance than adaptive Kmeans as they use prior knowledge of source numbers. Since AIC considers less penalty for model complexity than model distortion compared to BIC, it has some successful cases of source counting for $N_s = 5$ (Fig. 4.4(b)) while BIC is more successful for $N_s = 2$ with no noise (Fig. 4.4(a)). Due to presence of noise, Kmeans and its adaptive variations very rarely

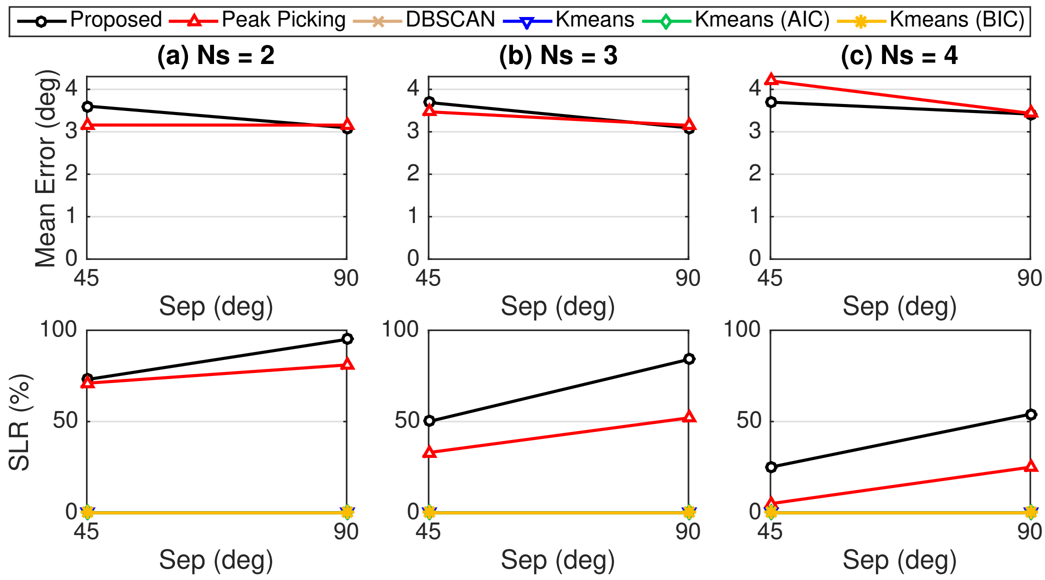


Figure 4.5: Evaluation results using estimated DOAs. Note that the lines and markers for DBSCAN, Kmeans, Kmeans (AIC) and Kmeans (BIC) are completely overlaid by each other as they all share zero value for SLR and therefore are not displayed for the mean error.

localize all the sources successfully which indicates how sensitive distance-based clustering is to noise in general. On the other hand, DBSCAN, due to noise-robustness, outperforms all Kmeans-based methods in our tests.

The results for (b)$N_s = 5$ are used for inferential statistical analysis. There was a statistically significant difference between groups as determined by one-way ANOVA ($F(5, 173) = 268.9, p < 0.001$). A Tukey post hoc test revealed that the estimation error was statistically significantly lower for the proposed method ($1.26° \pm 0.03°, p < 0.001$) compared to peak picking method ($3.40° \pm 0.08°$).

4.4.2 Evaluation using estimated DOAs

The testbed dataset introduced in Section 2.11.2 is used in this evaluation. The assumption of a mixture of unskewed distributions for DOA estimates in multi-source scenarios is realistic for DOA estimators with, for example, the commonly-used DPD [14] test, avoiding the potential skewness which is caused by DOAs at TF bins with simultaneously active multi-source or strong reverberation. Therefore the method of DPD-PIV [42], based on the selection of 25% of the TF bins with largest singular value ratios [46] is used as the DOA estimator.

Figure 4.5 indicates that the proposed method shows an average of 30% improvement in SLR compared to peak picking while maintaining the high accuracy of 3.5° on average. Also shown is the failure of all Kmeans-based methods, due to the presence of noise, and DBSCAN, due to the smooth knee in k-dist graph leading to erroneous trade-off between accurate and noise DOAs.

4.5 Conclusions

The use of density-based clustering is investigated for acoustic source DOAs and an autonomous source enumerator using DBSCAN clustering is proposed. The method runs an evolutive DBSCAN using varying sensitivity to noise and estimates the average density and the centroid of the reliable clusters at each iteration. It then performs a final autonomous DBSCAN clustering on the sparse distribution of reliable centroids without requiring the number of sources. The evaluation using generated and estimated DOAs validates the DOA estimation accuracy of $\leq 4^\circ$, 30% improvement in SLR compared to peak picking and the superior source counting robustness to noise, separation and source numbers for the proposed method compared to the comparative ones.

Chapter 5

Multi-Source Estimation Consistency (MSEC)

SEVERAL existing approaches to MS DOA estimation for speech sources use WDO [36], assuming sparseness of speech in the TF domain, in combination with subspace decomposition [23]. Such methods often follow three stages: (1) SS TF bin detection in which the TF bins predominantly containing a single source are detected; (2) SS DOA estimation where a DOA estimator based on the SS assumption is applied only at the detected SS bins; (3) Multiple source direction estimation using the set of temporal narrowband DOA estimates.

In a SS bin, the observation covariance matrix formed from the microphone array signals is expected to have unit rank. In real-world scenarios, DOA estimation has to be performed in reverberation that is characterized by the additive combination of direct-path propagation and reflections [63]. In such scenarios, SS dominance with unit rank covariance matrix rarely occurs at a TF bin and therefore some form of SS-validity confidence metric is used to detect the more reliable SS bins for use in DOA estimation. Methods such as the coherence test [71], SS Zone (SSZ) detection [72] or DPD test [14] assume the validity of SS assumption over a local TF region in the vicinity of a TF bin of interest and each method defines a specific SS-validity confidence metric. Having obtained the SS validity measure at each bin, two alternative approaches can be used for the selection of reliable bins. The methods in [72][14] identify the SS bins based on a comparison between the SS validity measures and a fixed user-defined threshold whereas the method in [46] selects a user-defined percentage of the TF bins with the strongest SS validity measures. In [71], SS bins are detected using the rank of the correlation matrix at each TF bin. Due to averaging across only local time frames and lack of subspace decomposition in the selection of SS bins, that approach is most effective only for MS DOA estimation in an anechoic environment. In [72], the average of pairwise correlation coefficients between adjacent sensors is used as a SS validity confidence metric, where the correlation averaging is performed only across the local frequencies of each time frame. It does not use subspace decomposition and is therefore prone to noise and multiple coherent sources. In DPD [14], SVD is employed and the SVR of the signals' covariance matrix is used as the SS-validity confidence metric. DPD performs the covariance averaging over adjacent frequencies and time-frames. The latter property, along with the use of subspace decomposition makes DPD robust to reverberation as it aims to find TF bins with not just a dominant SS but also a dominant direct path, ignoring bins containing significant reverberation.

As the number of simultaneously active sources increases, the performance of the previously mentioned methods degrades, although presence of the dominant SS still occurs. This is because the WDO assumption is valid in fewer TF bins and in smaller TF regions as the number of simultaneously active sources increases as shown in Section 5.1.

This Chapter addresses the problem of DOA estimation when WDO assumption is violated due to increase in the number of sources. It first proposes a metric in Section 5.2 and then presents a novel variation of MUSIC algorithm based on the proposed metric in Section 5.6. Figure 5.2 shows an overview of the MS DOA estimation system proposed for the first section. The novelties in this work are: (1) the use of density-based clustering in the context of acoustic DOA estimation, as used in DOAs clustering and source counting units in Fig. 5.2, (2) a novel SS-validity confidence metric for weighting of initial DOA estimates, as used in DOAs weighting unit in Fig. 5.2, and (3) a novel variation of MUSIC algorithm which uses the proposed metric to perform MUSIC individually per source. The proposed Multi-Source Estimation Consistency (MSEC) metric is based on a dynamic MS assumption, as opposed to the SS assumption in conventional approaches. MSEC uses a consistently large TF region where the number of simultaneously active sources within the region is autonomously estimated.

5.1 Problem Analysis

Consider a reverberant environment containing N_s simultaneously active speech sources with uniform angular spacing γ at 1 m distance from a microphone array. Each source represents a different speaker speaking different utterances. The received signal at a microphone in the STFT domain is

$$X(k,\tau) = \sum_{n=1}^{N_s} (D_n(k,\tau) + \sum_{j=1}^{\infty} R_{n,j}(k,\tau)), \tag{5.1}$$

where D_n denotes the direct-path component from source n and $R_{n,j}$ is the component of reflection j from source n. The frequency, k, and time frame, τ, indices are subsequently omitted for notational simplicity.

Let Signal-to-Interference Ratio (SIR) at a TF bin be the ratio of the magnitude of the dominant direct path, $|D_b|$, and the magnitude of the rest of the signals from the mixture of N_s sources, $|X - D_b|$, which includes all other direct paths and reverberations excluding the dominant direct path where

$$b = \underset{n}{\operatorname{argmax}}(|D_n|) \tag{5.2}$$

is the index of the dominant direct path. Figure 5.1 shows in white the TF bins with SIR $\geq 10\,\mathrm{dB}$ in such a scenario for $N_s = \{2,3,4,5\}$ and $\gamma = 50°$. It can be clearly seen that with the increase of N_s, the number of the bins and the size of the TF regions with valid WDO assumption decreases. For SS bin detectors based on a fixed-size analysis TF window, this leads to increasing failure of the SS assumption validity and consequent performance degradation for the SS bin detectors that rely on this assumption.

One solution [37] to this problem is the use of a dynamic MS assumption over a

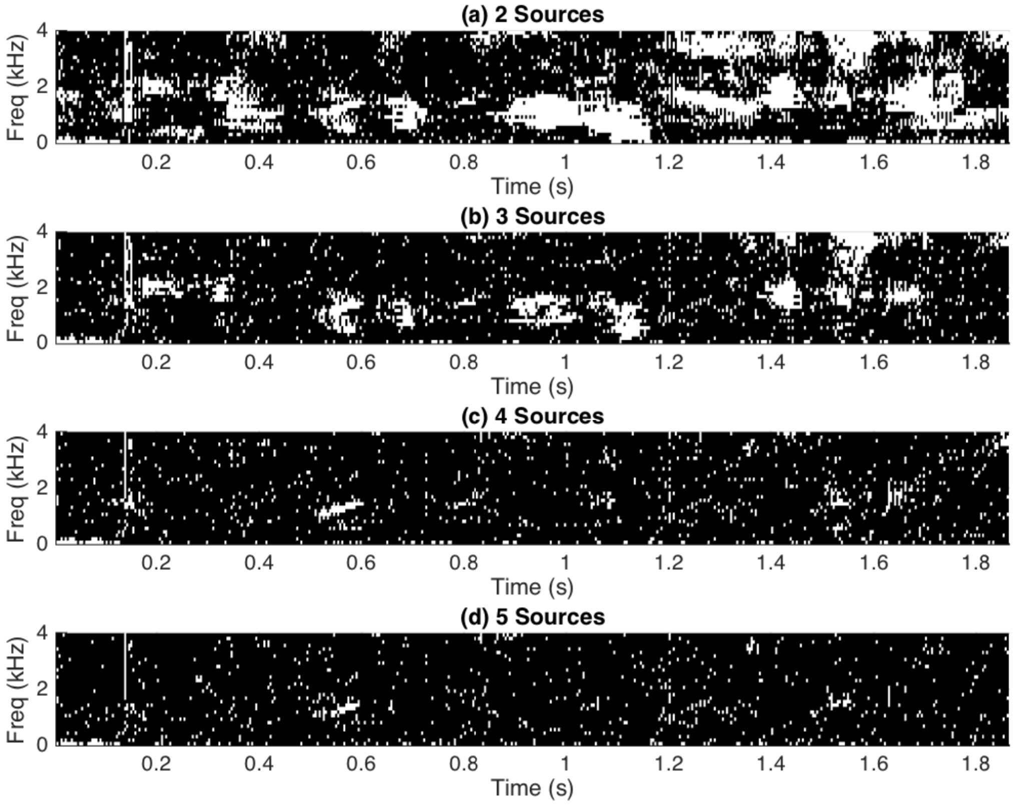


Figure 5.1: Illustration showing in white the TF bins with SIR$\geq$ 10 dB considering the signal as the dominant direct path and the interference as the reverberant signals mixture of (a) 2, (b) 3, (c) 4 and (d) 5 sources with $T_{60} = 0.4$ s.

fixed-size TF region where the number of active sources within the processing TF region is autonomously estimated. For such techniques, estimation of the optimum number of sources remains a challenge. In [37], the authors propose the use of the AIC [38] to find the optimum number of eigenvectors spanning the signal space for the MS assumption. Although this approach overcomes the problem of N_s estimation, it loses reliability with noisy observations.

The use of temporal narrowband DOA estimation based on the SS assumption in a MS scenario is expected to be relatively accurate at the TF bins containing one significantly dominant direct-path component and inaccurate otherwise. The direction of the error in erroneous DOA estimates is determined by the relative phase and amplitude of the impinging plane waves as shown in [17]. Such variance of directional displacements in

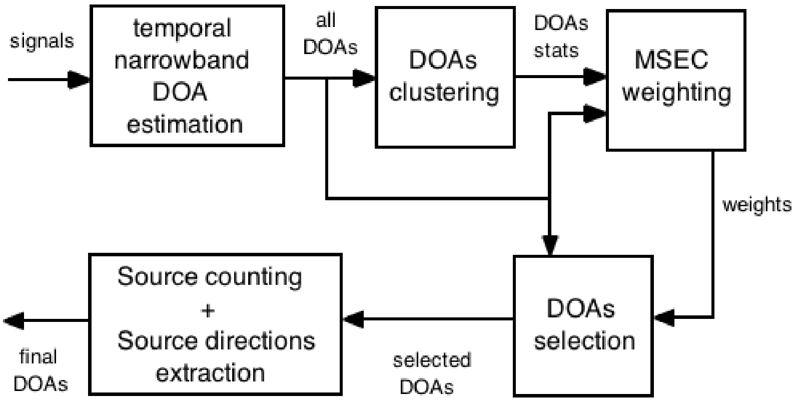


Figure 5.2: Block diagram of the proposed system for MS DOA estimation.

DOA estimates at non-SS bins results in spatially inconsistent erroneous DOAs whereas, in practical scenarios, DOA estimates at SS bins are expected to have spatial consistency if the sources are stationary or only slowly moving over time. In [73] and [74], the authors propose the use of diffuseness of DOA estimates which is based on SS assumption and suffers from the previously-stated problem of SS-based metrics as the number of sources increases. We therefore investigate the use of spatial consistency of SS-based DOA estimates under the MS assumption. We also investigate how to estimate the number of active sources over a TF region using this approach, as well as the validity of the SS assumption at a TF bin.

5.2 MSEC

Assuming that the initial DOAs (one per TF bin) are provided by any chosen temporal narrowband SS DOA estimation procedure, a new SS validity confidence metric is proposed based on spatial consistency of initial DOA estimates and a dynamic MS assumption. Two alternative distance- and density-based clustering techniques for the dynamic MS assumption are introduced and discussed. The architecture of the proposed system is illustrated in Fig. 5.2.

In order to increase the distinctness of densities between accurate and inaccurate initial DOAs for the purpose of robust estimation of the number of active sources, we consider all initial DOA estimates from the previous T frames. Therefore, at each frame

τ, we consider the set of DOA estimates $U(\tau)$ including all initial DOAs from frame τ to $\tau - T$, defined as

$$U(\tau) = \{\hat{\mathbf{u}}(t,k) : \forall k,\, t \in \{\tau, \tau-1, \ldots, \tau-T\}\}, \tag{5.3}$$

where $\hat{\mathbf{u}}(t,k)$ is the estimated DOA unit vector at time frame t and frequency k and T is a fixed user-defined temporal window length.

To quantify spatial consistency of the multi-modal distribution of DOA estimates from $U(\tau)$, an adaptive distance-based clustering technique such as K-means and a density-based noise-robust clustering technique such as DBSCAN [39] are used as two alternative approaches. In the following, τ, t and k are omitted for brevity where unambiguous.

5.2.1 Adaptive K-means Clustering

To find the optimum number of clusters, the AIC is calculated as [38]

$$\mathrm{AIC} = -2O + 2v, \tag{5.4}$$

in which $-O$, the negative maximum log-likelihood of the data, represents a measure of distortion and v, the number of parameters of the model, represents a measure of model complexity.

For K-means with a given K, the first term in (5.4) is replaced with Residual Sum of Squares (RSS) of the clustering giving [75]

$$\mathrm{AIC}(K) = \mathrm{RSS}(K) + 2VK, \tag{5.5}$$

where RSS(.) is the sum of squared angular distances of each member to its cluster centroid and V denotes the number of dimensions of the centroid which leads to VK parameters for K clusters. Note that with the increase of K, $\mathrm{RSS}(K)$ decreases while $2VK$ increases, which makes $\mathrm{AIC}(K)$ a penalty factor for a given model where its minimum gives the best clustering with the minimum number of clusters.

Having performed K-means for $K = \{1, \ldots, K_{max}\}$ on the set of DOAs U with random initializations, using (5.5), the optimum number of clusters, K_c, is chosen as

$$K_c = \arg\min_{K} \left[\mathrm{AIC}(K)\right]. \tag{5.6}$$

5.2.2 MSEC weighting

Having performed clustering on data set $U(\tau)$ by either adaptive K-means or DBSCAN, we obtain the estimated number of clusters $K_c(\tau)$, the clusters $\{C_i(\tau)\}_{i=1}^{K_C(\tau)}$ and the centroids unit vector $\{\hat{\mathbf{c}}_i(\tau)\}_{i=1}^{K_C(\tau)}$ where i is the cluster index. As a representative of the spread of DOA estimates within each cluster, the average member-to-centroid angular distance $\Theta_i(\tau)$ is calculated for each cluster as

$$\Theta_i(\tau) = \frac{1}{|C_i(\tau)|} \sum_{k \in C_i(\tau)} \angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau)), \tag{5.7}$$

where $\angle(.)$ denotes the angle in degrees between two vectors.

The MSEC weight for each DOA estimate is determined from two factors, the cluster weight and the member weight. For each DOA estimate, the cluster weight, which represents the normalized measure of concentration in its associated cluster, is

$$\psi(\tau, k) = 1 - \frac{\Theta_i(\tau)}{\pi},\ k \in C_i(\tau), \tag{5.8}$$

and the member weight, which represents the normalized measure of closeness to its associated centroid, is

$$\lambda(\tau, k) = 1 - \frac{\angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau))}{\pi},\ k \in C_i(\tau). \tag{5.9}$$

The MSEC weight in the TF domain is then formed as

$$w(\tau, k) = \sqrt{\psi(\tau, k)\lambda(\tau, k)}. \tag{5.10}$$

A special case of MSEC with $T = 0$ and $K_{max} = 1$ is proposed and evaluated in [48], which is based on the SS assumption within a time-frame.

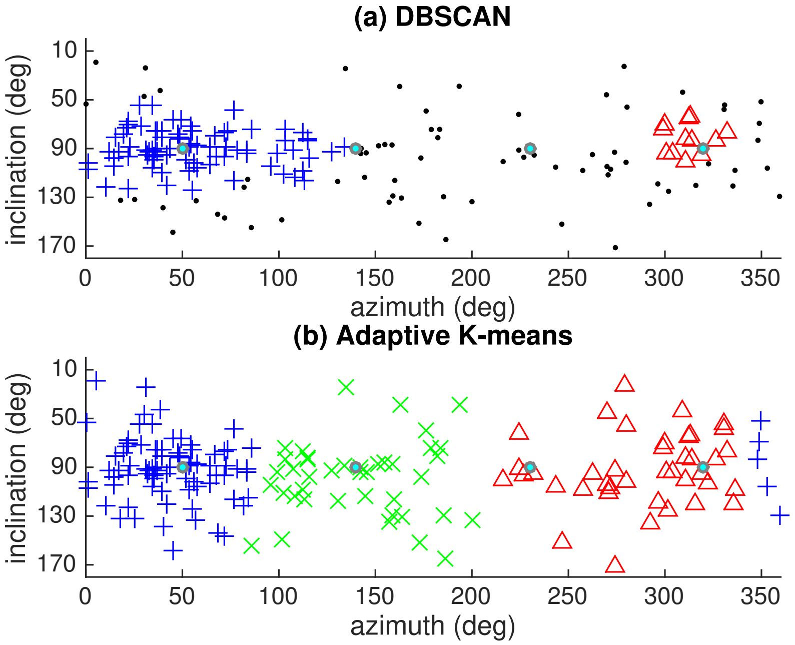


Figure 5.3: An example of DOA estimates from 5 consecutive time-frames clustered by (a) DBSCAN with $(\varepsilon, \text{MinPts}) = (20°, 10)$ and (b) adaptive K-means with $K_{max} = 4$. The colours and markers indicate the clusters while the black dots in (a) are the noise DOAs. The true source directions are marked as cyan filled circles.

Figure 5.3 displays DBSCAN and adaptive K-means clusterings of an example distribution of initial DOA estimates for 5 consecutive frames ($T = 4$). This illustrates that DBSCAN identifies and ignores the noise DOA estimates due to the use of a static definition of cluster density while adaptive K-means assigns every DOA estimate to a cluster.

Figure 5.4 shows a scatter plot of the normalized MSEC weights versus the normalized accuracy of the initial DOAs used in the example of Fig. 5.3. It can be seen that the noise-robust DBSCAN-MSEC has only weighted strongly the DOAs that have > 0.8 normalized accuracy, and zero-weighted the inaccurate DOAs with < 0.8 normalized accuracy.

Having weighted all the DOA estimates in the TF domain, only the estimates with the $P\%$ strongest weights are selected. The choice of P is investigated and discussed in Section 5.7.1.

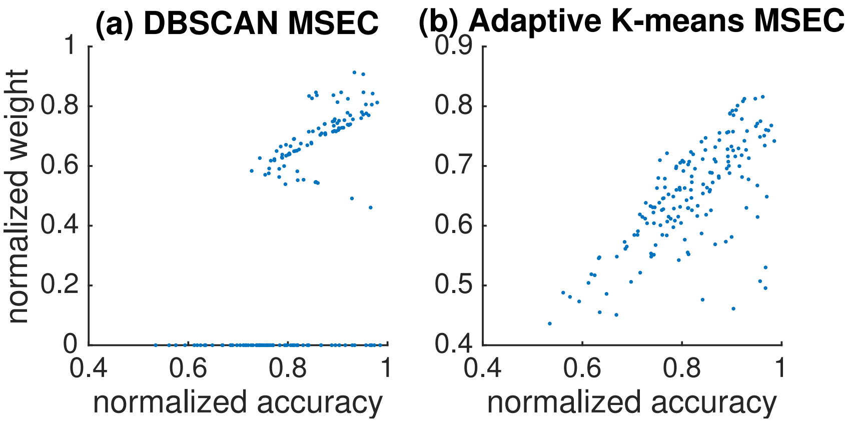


Figure 5.4: Normalized weight vs normalized accuracy for MSEC using (a) DBSCAN and (b) adaptive K-means for the example of Fig. 5.3.

5.3 MSEC metric Evaluations

The performance of the proposed metric is first evaluated using recorded anechoic speech convolved with simulated room impulse responses for SMA in the presence of reverberation and sensor noise. Performance using real speakers in a reverberant room is considered in Section 5.4. The evaluation is performed for a varying number of sources and angular separation. The DPD method is used as a baseline for comparison. Without loss of generality, the inclination of sources is fixed at 90°, for simulated data, so as to place them in the same horizontal plane as the microphone array for clarity of systematic evaluation of the effect of source separation. However, inclination is varied in the experimental verification using real data in Section 5.4. The testbed dataset introduced in Section 2.11.2 is used for this evaluation.

Any narrowband method can be used for the DOA estimator but for fast computation, the efficient PIV [16] method was used in these test as an example SS DOA estimator to obtain the initial DOA estimates.

The approximate covariance matrix $\mathbf{R_a}$ in (2.21) used by DPD had an averaging window with $J_\tau = 6$ and $J_k = 4$ over time and frequency respectively giving 500 Hz and 32 ms of window size in the TF domain based on our frequency and time resolution.

MSEC has a temporal window size of $T = 4$ frames in (5.3), which is chosen to be small enough to decompose the problem of N_s sources into $N < N_s$ sources over the interval and wide enough to form distinguishable densities for consistent DOAs. For clusterings used in variations of MSEC, adaptive K-means has $K_{max} = 4$ with random initialization per K and DBSCAN has $\varepsilon = 10°$ and MinPts= 10 which is approximately 5% of the number of the estimates in dataset $U(\tau)$. These values for the setting parameters of the evaluated methods are empirically chosen.

A uniform weighting strategy, in which all DOA estimates are selected, is also included in the evaluation as a reference. For the purpose of evaluating the performance of the weighting metrics only, a fixed selection percentage of $P = 25\%$ is empirically suggested [46] and used for DPD and both variations of MSEC. Therefore DPD and MSEC both select an equal number of DOA estimates, which is the top 25% DOAs with the highest weights while uniform weighting selects all DOA estimates. The error (in degrees) for each selected DOA estimate is calculated as the angular distance between the estimate and the nearest true DOA.

5.3.1 Accuracy of the selected DOAs

In this section the accuracy of the DOA estimates selected by the weights is evaluated. Figure 5.5 shows the mean error of the DOA estimates selected by each method for $\triangle\phi = \{45°, 90°\}$ and incremental $N_s = \{2, 3, 4\}$. It can be seen that MSEC variations select significantly more accurate DOA estimates compared to DPD and uniform weighting which validates the advantage of MS over SS assumption in MS scenario. DBSCAN-based MSEC has 92% to 129% mean accuracy improvement in these tests compared to DPD due to the dynamic MS assumption and noise-robustness. It can also be seen that as N_s increases the mean accuracy of the uniform and DPD weights improves. This is due to the decrease in the least possible error as the number of sources increases.

Figure 5.8 shows the top view and side view of the normalized smoothed histogram of DOA estimates selected by each method for an example experimental trial. The performance benefits of MSEC are shown and can be explained by observing the distinctness and sharpness of the peaks. It can be seen that MSEC variations have well defined peaks around each of the source positions especially for the fourth source (from the left) where

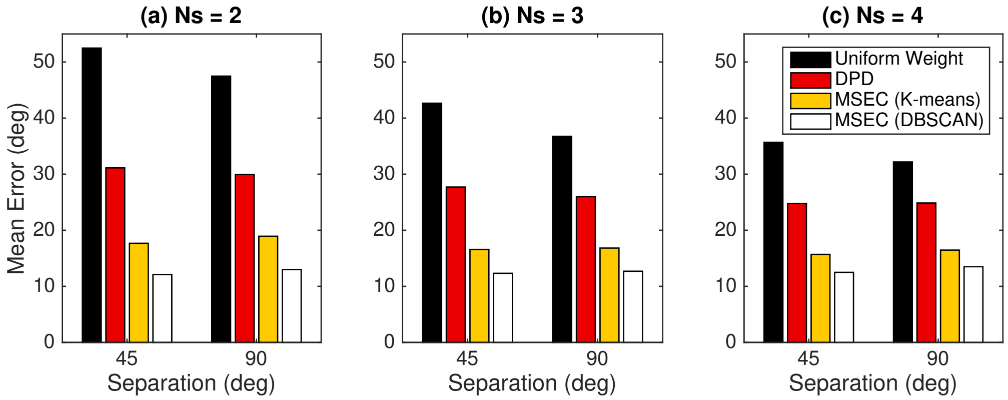


Figure 5.5: The overall mean error of the DOAs for varying separation and incremental number of sources.

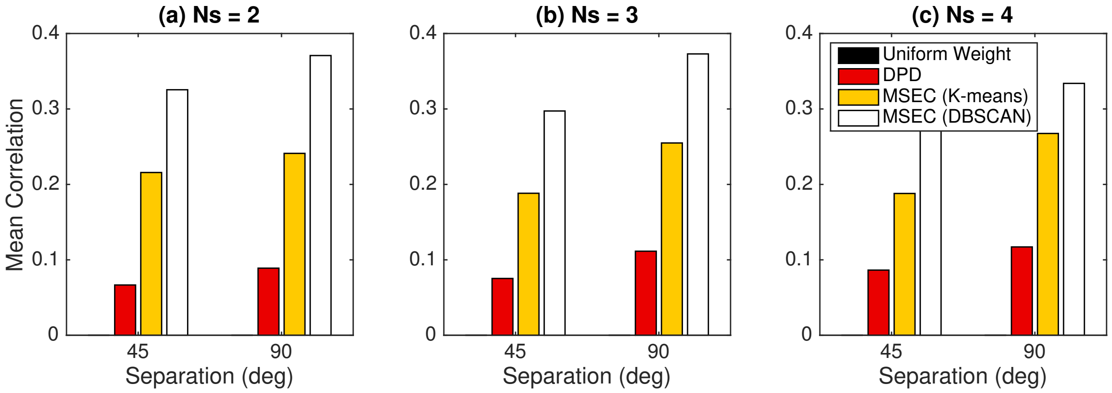


Figure 5.6: The overall mean correlation between the normalized weight and accuracy for varying separation and incremental number of sources.

DPD fails due to the oversize processing TF region at the TF bins with a significantly dominant fourth source resulting in selection of inaccurate DOA estimates. The reason for such failure is visualised and further discussed in the TF domain in Section 5.3.3.

5.3.2 Correlation between weights and DOA estimates accuracy

Figure 5.6 shows the mean correlation between the normalized weights and the normalized accuracy, as defined in 2.10, of their DOA estimate. DPD weights show low correlation with accuracy. On the other hand, MSECs are, at least by a factor of 4, more linearly correlated with DOA estimate accuracy. This is due to two reasons. (1) MSEC is calcu-

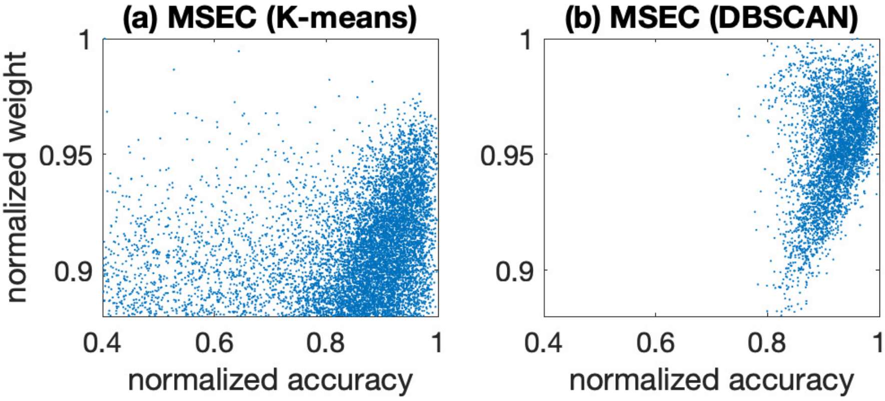


Figure 5.7: Distribution of the normalized weights and their DOA estimate accuracy for an example trial with $(N_s, \triangle\phi) = (2, 90°)$.

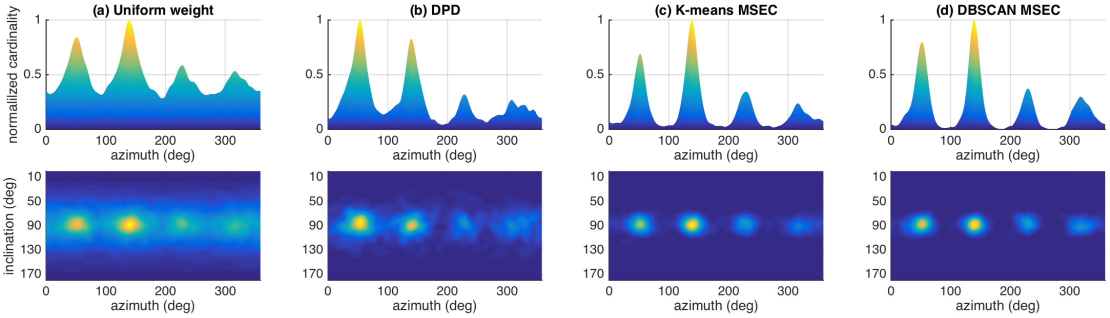


Figure 5.8: The side view (top row) and the top view (bottom row) of the normalized smoothed histogram of the selected DOA estimates using (a) Uniform weights including all DOAs, (b) DPD, (c) adaptive K-means MSEC and (d) DBSCAN MSEC for an example trial with $(N_s, \triangle\phi) = (4, 90°)$.

lated using the DOA estimates and is therefore expected to be directly impacted by DOA accuracy unlike DPD which uses eigenbeams. (2) The MSEC metric is calculated in the spatial domain using angular distances which has the same unit and nature as the DOA estimate accuracy whereas the DPD metric uses the SVR of the eigenbeams.

Figure 5.7 illustrates a scatter plot of the selected normalized weights versus normalized accuracy of their DOA estimates for K-means and DBSCAN-based MSEC for an example trial. It can be seen that DBSCAN-based weighting has significantly fewer inaccurate DOA estimates which are falsely weighted high compared to K-means. This is due to two reasons. (1) DBSCAN is a noise-robust clustering technique and is more capable of ignoring the inaccurate DOA estimates. (2) The outcome clustering of K-means is stochastic for each run because of random initialization and dependency of the outcome on the initialization, while DBSCAN does not require initialization and its outcome is therefore deterministic. During an experimental analysis it was observed that different trials of K-means on the same dataset with the same choice of K sometimes led to inconsistent clusterings and therefore inconsistent estimation of $K_c(\tau)$. Such inconsistent behaviour can sometimes lead to erroneous clustering and so erroneous weighting.

5.3.3 Effect of weighting on counting and direction estimation of sources

In this section the performance of each SS-validity confidence metric is evaluated in the context of source direction estimation and source counting using evolutive DBSCAN presented in Section 4.2. In Section 4.4 it is shown that the evolutive DBSCAN outperforms the conventional histogram peak picking as well as adaptive K-means and original DBSCAN techniques and is therefore chosen as our source counting and source direction extraction technique in this Section. The choice of NumIt=50 was empirically found to be a good trade-off between reliability and computational efficiency for our proposed evolutive DBSCAN. MSEC based on K-means is excluded from the evaluation in this section since DBSCAN-MSEC has a better performance as shown in the previous sections. The two performance metrics SLR and mean error, as defined in 2.10, are used for performance evaluation of the source counting and DOA estimation.

Figure 5.9 shows the mean error and SLR of DPD and DBSCAN-MSEC, abbreviated to MSEC in this section, for varying $\triangle\phi$ and N_s. It can be seen that MSEC noticeably outperforms DPD in all cases. In terms of DOA estimation accuracy, although MSEC and DPD perform very closely, MSEC slightly leads by 1° at 45° separation with 4 sources. In terms of source counting accuracy, MSEC significantly leads especially for $\triangle\phi = 45°$ as N_s increases. MSEC also shows strong robustness to separation and number of sources as its SLR drops only to 75% while DPD's SLR is reduced to 20% with the decrease in $\triangle\phi$ and increase in N_s. Such results match with the observation in Fig. 5.8. It is seen that the peaks of the multi-modal distributions, which affect the accuracy of DOA estimation, remain approximately at the same position for DPD and MSEC while the sharpness and

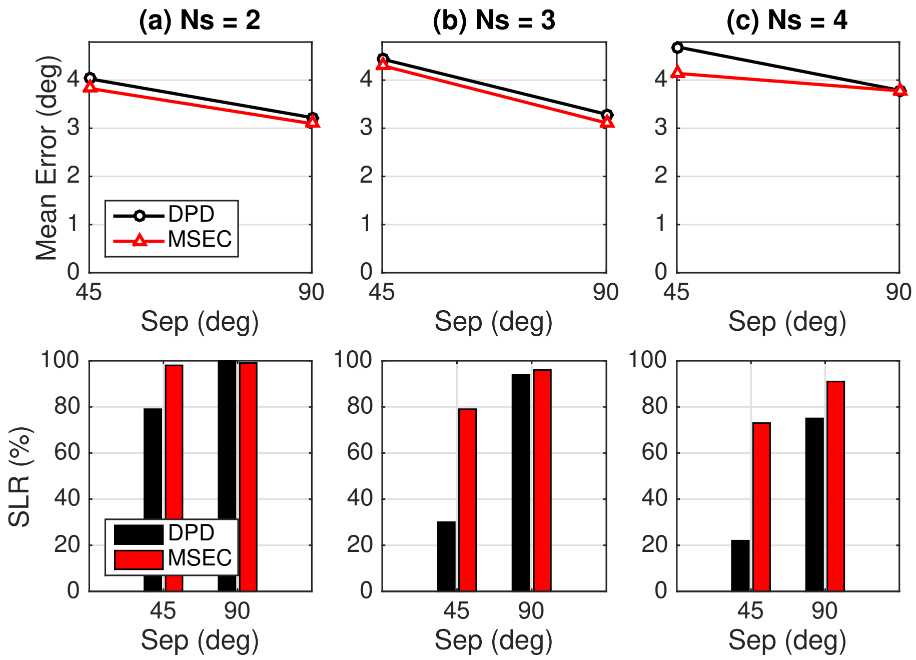


Figure 5.9: Mean error (top row) and SLR (bottom row) for (a) 2, (b) 3 and (c) 4 sources with varying source separation.

distinctness of the peaks, which affect the source counting, are significantly different.

An independent-samples t-test was also conducted to compare mean error for DPD and MSEC methods. There was not a significant difference in the scores for DPD (mean=4.69°, standard deviation=1.32°) and MSEC (mean=4.14°, standard deviation=1.11°) methods; t(93)=1.95, p = 0.054. However, MSEC generally performs better than DPD in SLR.

Figure 5.10 shows the TF bins with the top P = 25% strongest MSEC and DPD weights as well as the bins with PIV DOA estimates, which have $\leq 10°$ error and are considered as accurate DOAs, for an example trial. As shown in Fig. 5.10(c), accurate DOAs occur at varying-size TF regions and even at isolated TF bins. It can be clearly seen that MSEC has been more successful in detecting varying-size TF regions and isolated TF bins due to dynamic MS assumption over relatively large analysis window-size compared to DPD, which is based on the SS assumption over a small analysis window-size.

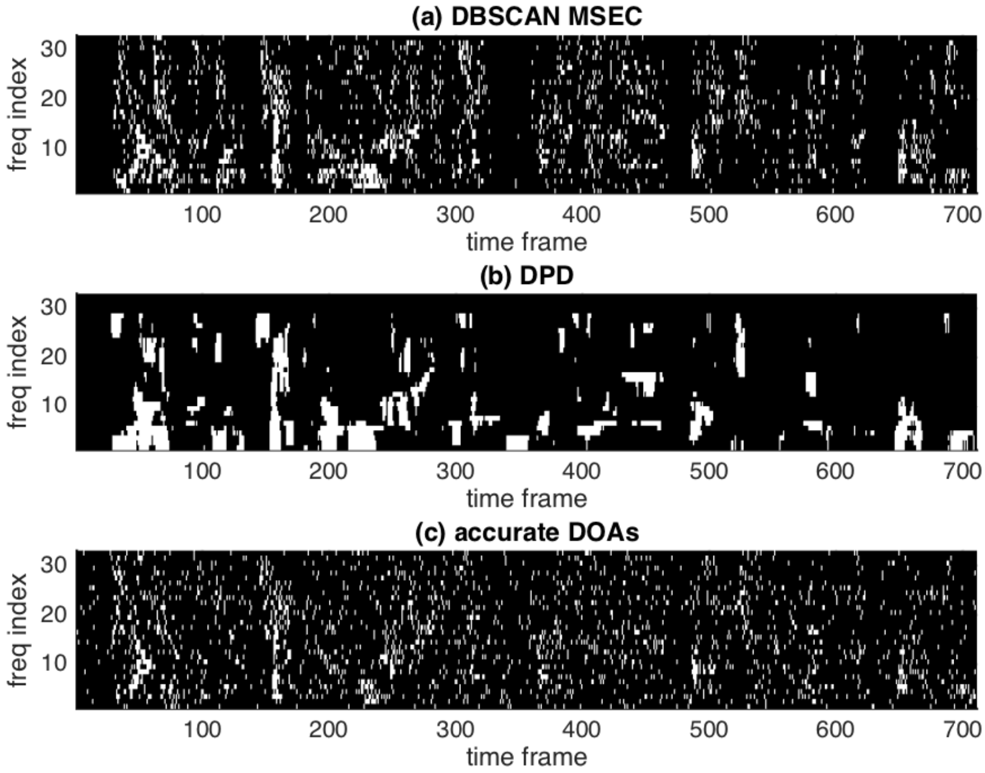


Figure 5.10: TF bins with top $P = 25\%$ strongest (a) DBSCAN-MSEC weights, (b) DPD weights and (c) $\leq 10°$ DOA error for $(N_s, \triangle\phi) = (3, 90°)$.

5.4 MSEC Experimental Verification using Real-world Data

In this section the performance of each method is evaluated using real recordings in a reverberant room introduced in Section 2.11.3. Figure 5.11 shows the normalized smoothed histograms for uniform weighting using all DOA estimates, DPD and DBSCAN-MSEC using $P = 25\%$ of the DOA estimates with the strongest weights, where DOA estimates are obtained using PIVs [16]. Due to only approximate knowledge of the ground-truth position of sources and array in the physical room, accurate numerical estimation error cannot be obtained. The approximate mean estimation error for all methods is 4°. All methods successfully estimate peaks corresponding to all four sources due to wide separation of sources. Therefore the measure of 'peak strength' as introduced in 2.10 is used. Table 5.1 presents the peak strength of each peak for all methods.

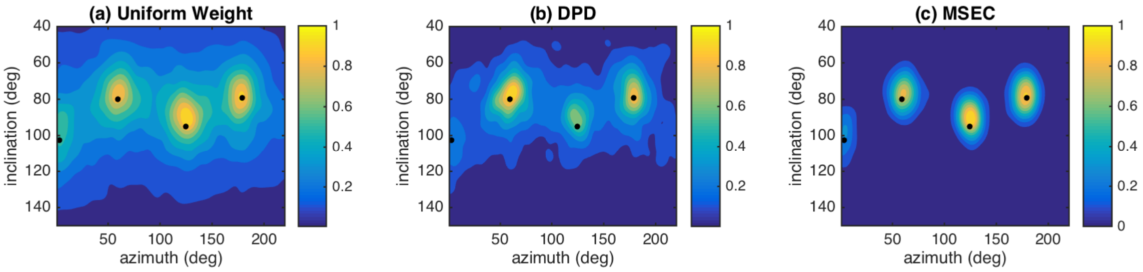


Figure 5.11: Zoomed normalized smoothed histograms for uniform weighting (all DOA estimates), DPD and DBSAN-MSEC (both based on $P = 25\%$) using real recording. The black dot represents the approximate true DOA.

Peak	Uniform Weight	DPD	MSEC
1	2.08	2.94	6.04
2	1.96	2.59	6.03
3	1.82	1.75	4.97
4	0.99	0.67	4.31
Mean	1.71	1.99	5.33

Table 5.1: Peak Strength of each peak for all methods

The smoothed histograms in Fig. 5.11 and the peak strengths in Table 5.1 show that MSEC metric significantly outperforms DPD and uniform weighting using real recordings and serves towards validation of the evaluation results based on simulation in the previous section.

5.5 MSEC metric Conclusion

A confidence metric for validity of SS assumption in a TF bin has been proposed using spatial consistency of initial DOA estimates. It employs adaptive K-means based on AIC or noise-robust DBSCAN clusterings to group spatially consistent initial DOA estimates, which are derived by a SS-based DOA estimator. Each DOA estimate is weighted using its distance-to-centroid and cluster's spread and finally the ones with the strongest weights are selected to be used in source counting and source direction estimation. The proposed metric is based on MS assumption over a relatively large TF region compared to conventional metrics, which are based on SS assumption over a small-size TF region. A novel use of density-based DBSCAN clustering in the context of source localization has also been used to propose an autonomous evolutionary method for source counting and final source direction estimation. The evaluations using simulations and real recordings show that our proposed metric significantly improves the performance of source counting, compared to the baseline and the state-of-the-art metrics, and provides at least the same accuracy as the state-of-the-art for source direction estimation.

5.6 MSEC-MUSIC

In DPD-MUSIC, the covariance matrix is calculated over a local TF region centred on a bin which indicates the SS-assumption over a local TF region. Unlike DPD, MSEC provides an estimate of the SS-bins assigned to each speaker across the entire TF domain. The idea in this work is to remove the TF-domain regional limitation in DPD-MUSIC by replacing DPD with MSEC as the pre-processing stage to improve the quality of covariance matrix used in the MUSIC algorithm, particularly for the case of multiple simultaneously active speech sources.

Having obtained the selected DOA estimates using MSEC, for the purpose of robust clustering, the potential outlier DOAs are removed if the average cardinality over a spatial window of 1 × 1 degree (azimuth × inclination) centred on DOA estimate is below a threshold γ. Applying K-means with $K = N$ on DOA estimates after outlier removal, we obtain the clusters $\{C_n\}_{n=1}^{N}$ and the centroids unit vectors $\{\hat{\mathbf{c}}_n\}_{n=1}^{N}$. The MSEC covariance matrix for source n is formed using the SS-bins across all TF domain that are assigned to source n

$$\mathbf{R}_{MSEC}^{n} = \frac{1}{|C_n|} \sum_{(\tau,k)\in C_n} \mathbf{a_{lm}}(\tau,k) \mathbf{a_{lm}}^{H}(\tau,k), \tag{5.11}$$

where $|C_n|$ indicates the number of members in cluster C_n. Using SVD on $\mathbf{R}_{MSEC}^{n}$ as in (2.22), the MSEC-MUSIC spectrum for each source is given as

$$P_{MSEC-MUSIC}^{n}(\Omega) = \frac{1}{\| \left(\mathbf{U_v}^{n}\right)^{H} \mathbf{Y_{lm}}^{*}(\Omega)\|^{2}}, \tag{5.12}$$

in which the global peak indicates the estimated DOA for that source.

5.7 MSEC-MUSIC Evaluations

An evaluation of methods was conducted using simulation with the same testbed materials as in Section 2.11.2. MSEC was performed on initial DOA estimates obtained by the PIV [16] DOA estimator. We empirically chose $K_{max} = 4$, $T = 4$ frames, $\gamma = 0.3$ for the average cardinality threshold in outlier removal. DPD test had $J_\tau = 6$ and $J_k = 4$ as the size of its averaging window in the TF domain in (2.21). We empirically chose the threshold in (2.28) as $\epsilon = 6$ which also matches the recommended value in the original paper [14]. The MUSIC spectrum in (2.30) and (5.12) was calculated with 1° resolution across azimuth and inclination (360×181). Incoherent DPD-MUSIC was excluded from our evaluation since studies in [49] show that incoherent DPD-MUSIC fails in the case of low angular separation of sources as the two peaks associated with two adjacent sources can be merged into one peak over summation of MUSIC spectra which causes the second highest peak to be detected far from the sources.

The original DPD test is based on absolute selection due to comparison of SVR with a fixed threshold. This results in reduction of selected TF bins as the number of sources increases for DPD unlike MSEC which is based on relative selection of top M% best DOA estimates. For the purpose of fairness in the selection process in our evaluation, we also include an alternative DPD-MUSIC based on relative selection which selects the TF bins with the top M% SVR, η_{DPD}.

5.7.1 Effect of selection percentage

In this section, we evaluate the effect of selection percentage for MSEC and DPD-MUSIC with relative selection in order to find the optimum M for both methods.

Figure 5.12 shows the mean estimation error as a function of selection percentage M for $N = \{2, 3, 4, 5\}$ and $\triangle\phi = 45°$. As we can see in Fig. 5.12, low ($\leq 10\%$) and high ($> 50\%$) values of M respectively cause underestimation and overestimation of number of bins which both result in high estimation error. As expected, the optimum M increases with increasing N. We can also observe that the average performance of MSEC, compared to DPD, is more dependent on the M since the value of M directly affects the quality of the covariance matrix in (5.11) which is the input to SVD of MUSIC unlike DPD in which the covariance matrix calculation in (2.21) is independent of M. According to these findings, the value $M = 25\%$ is selected for both MSEC and relative-DPD.

5.7.2 Overall Evaluation

In this section, we evaluate the best performing approaches of MSEC-MUSIC ($M = 25\%$), absolute ($\epsilon = 6$) and relative ($M = 25\%$) coherent DPD-MUSIC for $N = \{2, 3, 4, 5\}$ and widely varying $\triangle\phi$.

As can be seen in Fig. 5.13, in all cases of N for all methods the performance of DOA estimation improves as $\triangle\phi$ decreases below 30°. Since the spatial resolution of Y_{lm} depends on the maximum SH order L, below a certain $\triangle\phi$ multiple sources active in a TF bin are considered as a single source spatially between the true sources and therefore that bin is selected as a SS-bin. In such cases, the lower the angular separation of sources is, the lower the estimation error will be. For separation of $\triangle\phi \geq 30°$, both DPD-MUSIC methods in our experiments lose accuracy and robustness to N and $\triangle\phi$ unlike MSEC-MUSIC which shows relatively strong robustness. As expected, relative DPD shows higher robustness to N as it uses a dynamic selection process unlike absolute DPD with static selection. In

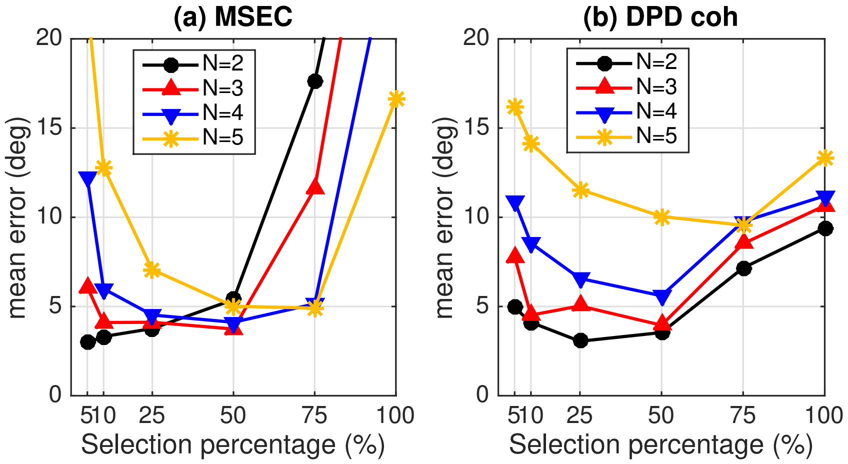


Figure 5.12: Mean estimation error as a function of M for MSEC- and relative DPD-MUSIC for varying N and $\triangle\phi = 45°$.

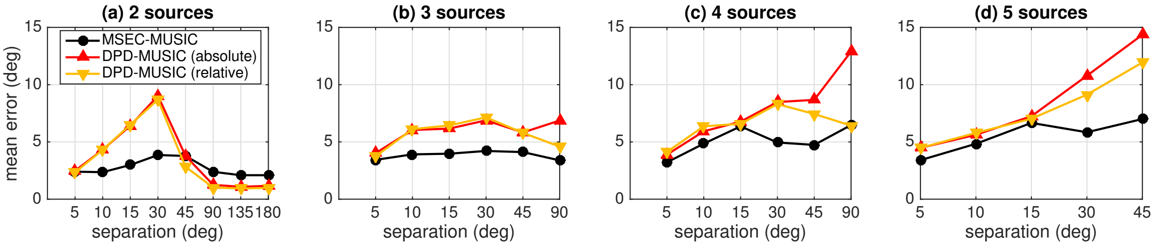


Figure 5.13: Mean error of MSEC-MUSIC, relative and absolute DPD-MUSIC for varying N and $\triangle\phi$.

overall MSEC-MUSIC, due to global consideration of SS-bins, shows stronger robustness to source separation and number of sources as it varies from 2.4° to 6.5° compared to DPD-MUSIC which is based on local consideration of SS-bin and changes from 2.2° to 15° mean estimation error.

The results for (d) 5 sources with 45° separation are used for inferential statistical analysis. There was a statistically significant difference between groups as determined by one-way ANOVA ($F(2, 297) = 16.99, p < 0.001$). A Tukey post hoc test revealed that the mean estimation error was statistically significantly lower for MSEC-MUSIC ($7.00° \pm 0.91°, p < 0.001$) compared to relative DPD-MUSIC ($11.94° \pm 0.91°$).

5.8 MSEC-MUSIC Conclusions

A DOA estimation method has been proposed for multiple active sources. The method exploits a variant of multi-source clustering of speaker-dominant time frequency bins to make an improvement on the computation of the spatial covariance matrix used in the MUSIC algorithm. The effectiveness of this approach has been tested for multiple simultaneously active speech sources in a simulated acoustic environment with 0.4 s reverberation time, and using a spherical microphone array. The simulation shows that our technique MSEC-MUSIC significantly outperforms the state-of-the-art DPD-MUSIC with less than 6.5° mean estimation error, 4° and 2.5° robustness to number of sources and source separation respectively for up to 5 sources with widely varying source separations in the presence of realistic reverberation and sensor noise. As a conclusion, our work indicates that estimation of a global covariance matrix per speaker, compared to clustering of local signal spaces derived from local covariance matrices, leads to a more accurate global signal space per speaker.

5.9 MSEC conclusions

As an overall conclusion, this Chapter employs spatial consistency of the initial narrowband DOA estimates over time and frequency to weight the reliability of each DOA. This metric can be used to select the reliable initial DOAs from which final DOAs are extracted. In addition, an alternative approach is also proposed in which the reliable TF bins selected using the proposed metric are clustered and grouped per source and a MUSIC algorithm is applied for each source individually. MSEC metric outperforms the baseline and the state-of-the-art metrics with accuracy of $< 4°$ for DOA estimation and improvement from 20% to $> 70\%$ on source detection compared to the state-of-the-art metric. The MSEC-MUSIC also outperforms the state-of-the-art DPD-MUSIC with $< 6.5°$ accuracy, 4° and 2.5° robustness to number of sources and source separation.

Chapter 6

Dual-Intensity Vectors (DIVs)

THE degradation factors for the performance of DOA estimators are typically reverberation, sensor/environmental noises and sources' activity, loudness and movement. In this chapter we are mainly interested in challenging scenarios in which the sources are stationary and simultaneously active with different average loudness where one source is mostly and significantly masked by others. This is an often-occurring scenario where the number of sources exceeds two and one or more sources are at further distance than others.

The narrowband DOA estimation methods either provide analytical closed-form solution or are based on steering vectors (open-form). In analytical-based methods, the DOA is directly calculated from the signals whereas the steering-based methods combine the signals with the steering vector of the array for all the possible directions to find the direction with the most reliable outcome. PIV can be named as a well-known analytical-based method. Although analytical approaches such as PIV are computationally fast due to avoiding many possible directions and estimating the DOA directly instead, they suffer from a limitation, due to using low-order eigenbeams, which results in low spatial resolution.

On the other hand, steering-based methods such as SRP and MUSIC benefit from their formulation's extendibility to high order spatial information and so high spatial accuracy but they can be computationally expensive if they are applied as a narrowband estimator per TF bin.

SS-based narrowband DOA estimators only result in accurate DOA if the WDO assumption is valid at the bin. For narrowband MS scenarios where multiple sources are simultaneously active in a bin, SS-based DOA estimator fails due to violation of this assumption. In the most optimistic cases where subspace decomposition is used, the SS-based methods estimate the DOA of the most dominant source and still fail in detecting a quiet source if the source rarely has the chance of domination in a bin as shown later in this chapter.

A solution to this problem is the use of MS narrowband DOA estimation. A well-known example of it, is MSMUSIC [25, 23] where the noise subspace is defined by the eigenvectors excluding the first N such vectors which represent the signal subspace, assuming the presence of N simultaneously active sources in the bin. The DOAs per bin are obtained as the top N peaks in the MUSIC spectrum. This approach suffers from high computational cost of steering and multi-peak detection per TF bin. The single peak detection in the SS-based approaches is relatively fast as it aims for the global maximum in a wrapped space whereas multi-peak detection requires multiple 2D-neighbouring check for peak detection, which is significantly more computationally expensive than the single peak detection.

The motivation behind this chapter is to study and propose a relatively low-computational and analytical narrowband MS DOA estimation algorithm which uses low-order spatial information to directly estimate single, two or more DOAs per TF bin. Two alternative methods for an analytical narrowband MS DOA estimation, named as Multi-Source Pseudointensity Vectors (MSPIVs) and Dual-Intensity Vectors (DIVs), are proposed.

The first method, MSPIV, is based on de-mixing a mixture of N sources into a multiple SS scenario on each of which PIV is performed. The second method DIV is based on a two-source assumption and analytically derives up to two DOAs from the mixture signals. A subspace version of the second method is also proposed in order to improve its robustness to noise and scenarios with more than two sources.

6.1 Multi-Source PIVs

The idea for the first proposed method is to extend the concept of SSPIV to multi-source in a TF bin rather than the dominant source only. Assuming N active sources in a bin, MSPIV for each source $n \in \{1...N\}$ is

$$\mathbf{I}^{(n)}_{\mathbf{mspiv}}(\tau, k) = \frac{1}{2}\Re\left\{ a_{00}(\tau, k)^* \begin{bmatrix} D_{-x}(\tau, k, \mathbf{U}_n) \\ D_{-y}(\tau, k, \mathbf{U}_n) \\ D_{-z}(\tau, k, \mathbf{U}_n) \end{bmatrix} \right\}, \tag{6.1}$$

where $\mathbf{U}_n$ is the n^{th} column of the matrix $\mathbf{U}$ obtained in (2.22). Note that $\mathbf{I}^{(1)}_{\mathbf{mspiv}}$ is equivalent to $\mathbf{I}_{\mathbf{sspiv}}$ in (2.25). The MSPIV DOAs are obtained using (2.19) for each $\mathbf{I}^{(n)}_{\mathbf{mspiv}}$.

6.2 DIVs

So far all variations of PIV either take the input assuming as a single source scenario (PIV) or decompose the input into multiple single source scenarios (SSPIV and MSPIV). In DIV the input is assumed to be the mixture of two sources and then both DOAs are analytically derived as shown in Appendix C.

6.3 Narrowband illustrative validation

Consider two simultaneously active sources in a TF bin with DOAs $(\varphi_1, \theta_1) = (30, 30)^\circ$ and $(\varphi_2, \theta_2) = (120, 70)^\circ$ and no reverberation. Let the Sources Ratio (SR) denote the ratio of their amplitudes $(10\log_{10}(\frac{|S_1|}{|S_2|}))$ in decibels. The phases of S_1 and S_2 are random.

Figure 6.1 illustrates the SRP spectra and the DOAs estimated by PIV, SRP and the proposed method for such a scenario with varying SR=$\{10, 5, 0\}$ dB and additive white Gaussian sensor noise giving SNR=$\{30, 20, 10\}$ dB. Table 6.1 presents the associated mean errors for PIV and DIV. It can be seen that the decrease of SR (through the columns per each row), due to the increase in the violation of WDO assumption, results in more error

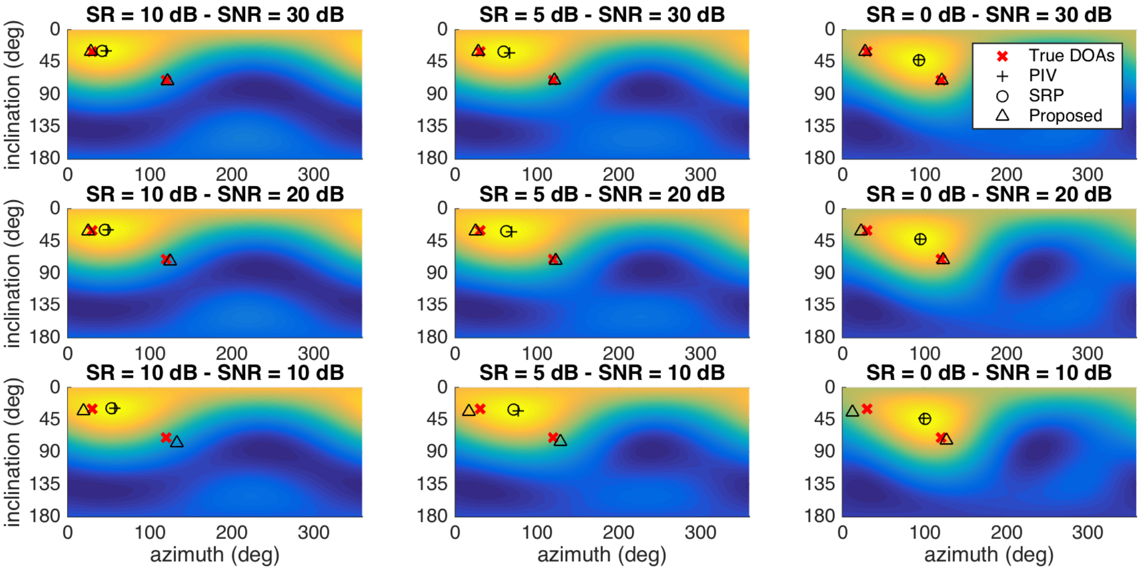


Figure 6.1: The SRP spectra for two sources with varying sources ratio of SR=$\{10, 5, 0\}$ dB (columns) and SNR=$\{30, 20, 10\}$ dB (rows). The true DOAs are marked as red cross ($\times$) and the estimates by PIV, SRP and the proposed method are marked as black plus ($+$), circle ($\circ$) and triangle ($\triangle$) respectively.

PIV , DIV	SR = 10 dB	SR = 5 dB	SR = 0 dB
SNR = 30 dB	8.2° , 1.4°	18.5° , 1.0°	35.9° , 1.0°
SNR = 20 dB	9.4° , 4.0°	19.7° , 3.2°	35.1° , 3.1°
SNR = 10 dB	13.6° , 10.3°	23.7° , 8.8°	32.1° , 8.7°

Table 6.1: Mean error of narrowband PIV and DIV, in black and red respectively, for varying SNR and SR as shown in Figure 6.1.

for PIV and SRP estimates whereas the proposed method shows strong robustness to SR due to the dual source assumption.

Across all cases, the proposed method also maintains the relatively high accuracy of 1° to 10° mean error (with respect to the closest true DOA) compared to PIV whose mean error dramatically varies between 8° to 36°. However, as the SNR reduces from 30 dB to 10 dB, through the rows per each column, PIV estimates have approximately 5° change of mean error while the proposed method shows an approximate 9° increase in the mean error. The proposed method shows less robustness to noise as it estimates twice number of parameters from noisy observations compared to PIV and therefore its estimates are more affected by the change of SNR.

6.4 Subspace DIVs

In order to improve DIV's robustness to noise, the mixture of signal subspaces $\mathbf{U}_1$ and $\mathbf{U}_2$, obtained using EVD in (2.22), are formed as

$$\mathbf{U}_{1,2} = \mathbf{U}_1 + \mathbf{U}_2, \tag{6.2}$$

which represents the de-noised eigenbeams mixture of the two most dominant sources. Instead of $\{a_{lm}\}_{l=0}^{1}$, the first four elements of the vector $\mathbf{U}_{1,2}$ are treated as the inputs to DIV formulation to give SSDIV.

6.5 Wideband illustrative validation

Consider three simultaneously active male talkers for a duration of 2 s with DOAs $(\varphi_1, \theta_1) = (190, 70)^\circ$, $(\varphi_2, \theta_2) = (290, 110)^\circ$ and $(\varphi_3, \theta_3) = (30, 150)^\circ$ all 1 m away from the array in an anechoic environment. Random white Gaussian noise with SNR= 10 dB is added to the sensors. The sources 1 and 3 have an equal average loudness level while source 2 has SIR=-26 dB (10% of the loudness of sources 1 and 3).

Figure 6.2 illustrates $|a_{00}|$ for each source in isolation before the mixing (a,b and c) and for the mixture of sources (d) in the STFT domain. Figure 6.2(e) displays the ID of the dominant source in the STFT domain where dominance is defined as the maximum contribution of $|a_{00}|$ in the mixture. It can be seen that source 2 is rarely dominant compared to the other sources due to its low-level loudness and overlap with others. Hence very few bins, compared to other sources, with accurate DOA estimates belonging to source 2 are expected from PIV and SSPIV.

Figure 6.3 illustrates the smoothed 2D histograms of DOAs estimated by PIV, SSPIV, MSPIV, SSDIV and MSMUSIC. Both MSPIV and MSMUSIC were applied using $N = 3$. Note that PIV and SSPIV estimate a single DOA per TF bin whereas MSPIV estimates three DOAs per bin. SSDIV and MSMUSIC respectively estimate up to two and three DOAs per bin. As expected, PIV and SSPIV almost fail to localize source 2

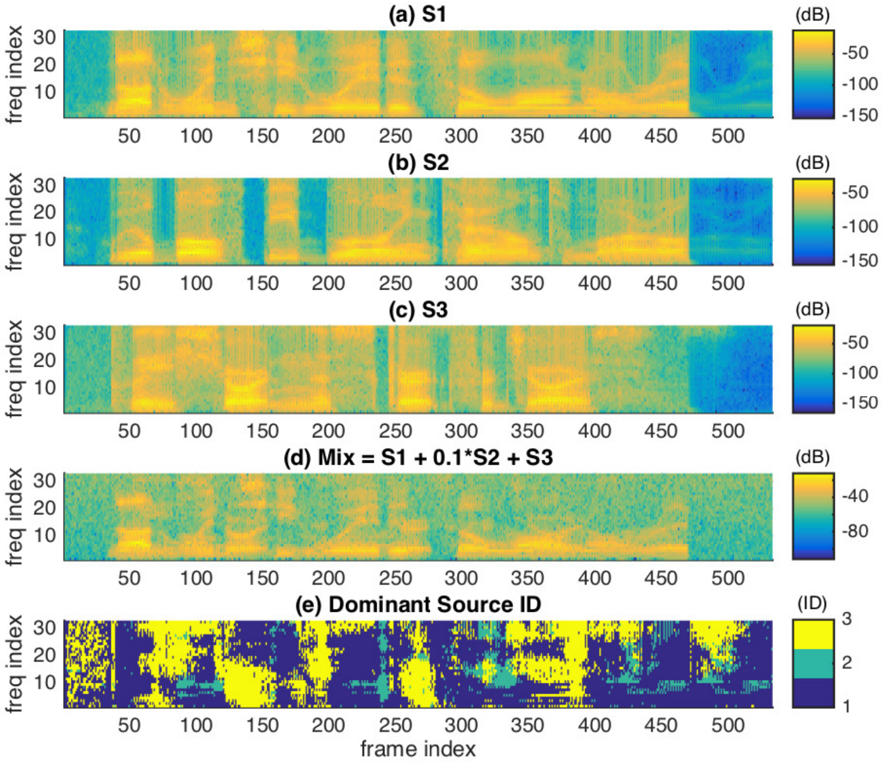


Figure 6.2: Magnitude of the omnidirectional (zeroth order) eigenbeam for (a) S_1, (b) S_2, (c) S_3 and (d) the mixture of three sources. Part (e) shows the ID of the dominant source where dominance is defined as the maximum magnitude contribution in the mixture.

since there are relatively very few bins where source 2 is dominant as shown in Fig. 6.2(e). On the other hand, both proposed methods estimate enough DOAs belonging to source 2, which form a significant peak around it in the histogram. This is due to successful estimation of DOAs belonging to source 2 even in TF bins where it is not dominant. MSMUSIC shows slightly better localization of source 2 compared to PIV and SSPIV as it is based on a multi-source assumption but is outperformed by our two proposed methods. It can also be seen that MSMUSIC results in extra peaks at the opposite directions of true DOAs. Considering MUSIC as equivalent to beamforming since it provides spatial selectivity, this phenomenon is the result of estimating a peak at the main lobe and a second peak at the rear lobe of the beam pattern in the MUSIC spectrum in the TF bins where only a single source is present or significantly dominant.

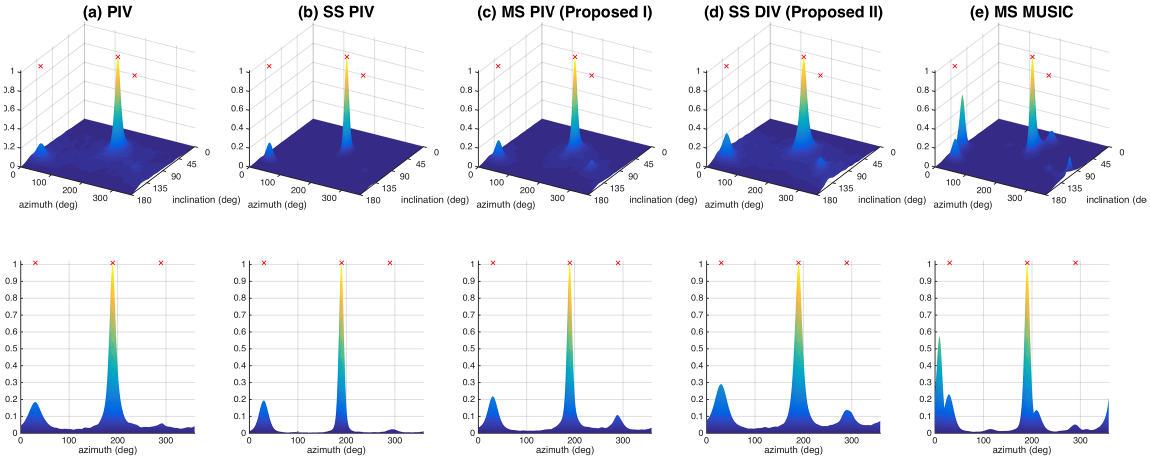


Figure 6.3: Corner view (top row) and the side view (bottom row) of the normalized smoothed 2D histograms of DOAs estimated by (a) PIV, (b) SSPIV, (c) MSPIV, (d) SSDIV and (e) MSMUSIC. The true DOAs are marked by red cross (×)

6.6 Evaluations

The performance of the proposed (MSPIV and SSDIV) and comparative (PIV, SSPIV and MSMUSIC) methods is evaluated using the dataset introduced in Section 2.11.5. Some segments (with overall duration of 1 second) of source 2 were cut off in time to make inconsistency of activity as illustrated in Figure 6.4(b). Note that the sources were mixed with a mixing coefficient of 1 for sources 1, 3 and 4 and 0.2 (SIR=-24 dB) for source 2 so that source 2 is strongly masked by others. A sampling frequency of 8 kHz was used with 50% overlapping time-frames of 8 ms duration. The sample covariance matrix in (2.21) had $J_\tau = 6$ and $J_k = 4$ as the size (number of bins) of the averaging windows over time and frequency respectively. This gives 32 ms and 500 Hz window-size in the TF domain based on our time and frequency resolution. MSMUISC and MSPIV assume $N = N_s = 4$ active sources per bin. Hence MSMUSIC and MSPIV estimate four DOAs whereas PIV and SSPIV estimate one DOA and SSDIV estimates up to two DOAs per bin. Only the eigenbeams for up to SH order $L = 1$ (first four eigenbeams) are used for each method in order to exclude the impact of high-order ($l > 1$) harmonics in the performance since PIV and SSPIV only use up to the first SH order.

Figure 6.4 illustrates the STFT representation of each source in isolation, (a) to

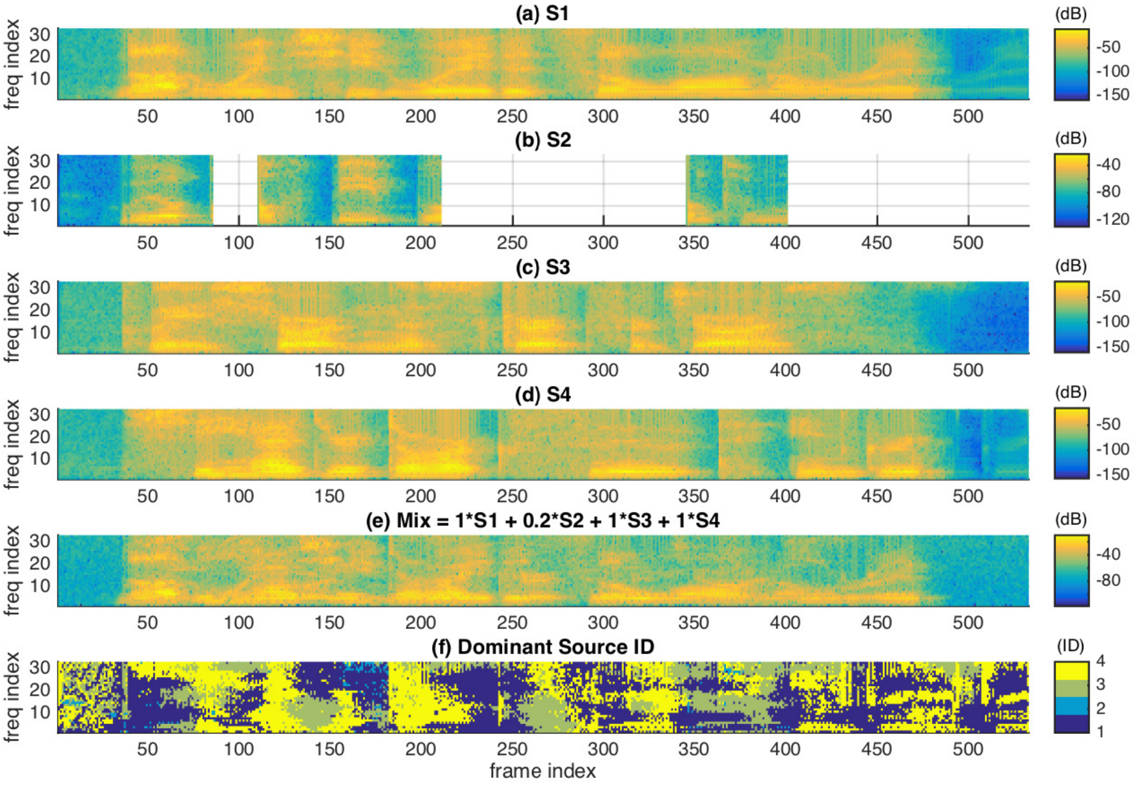


Figure 6.4: Magnitude of the omnidirectional (zeroth order) eigenbeam for (a) S_1, (b) S_2, (c) S_3, (d) S_4 and (e) the mixture of four sources. Part (f) shows the ID of the dominant source where dominance is defined as the maximum magnitude contribution in the mixture.

(d), the mixture (e) as well as the ID of the dominant source in each TF bin (f). It can be seen from Fig 6.4(f) that source 2 is rarely dominant which makes this situation a challenging scenario for SS-based DOA estimators.

In each trial, the top $N_s = 4$ peaks in the smoothed histogram were selected as the estimated DOAs. In order to avoid any ambiguity due to data association uncertainty in our results, best case data association was used to obtain the mean estimation error using (2.31).

Figure 6.5 presents the mean error of each method for SNR=$\{10, 20, 30, \text{Inf}\}$ dB. MSMUSIC fails with high error due to the presence of erroneous peaks of the rear lobe, as explained before. This is caused as a fixed number of peaks are selected per bin for each method. Other techniques do not suffer from the rear lobe problem since they are

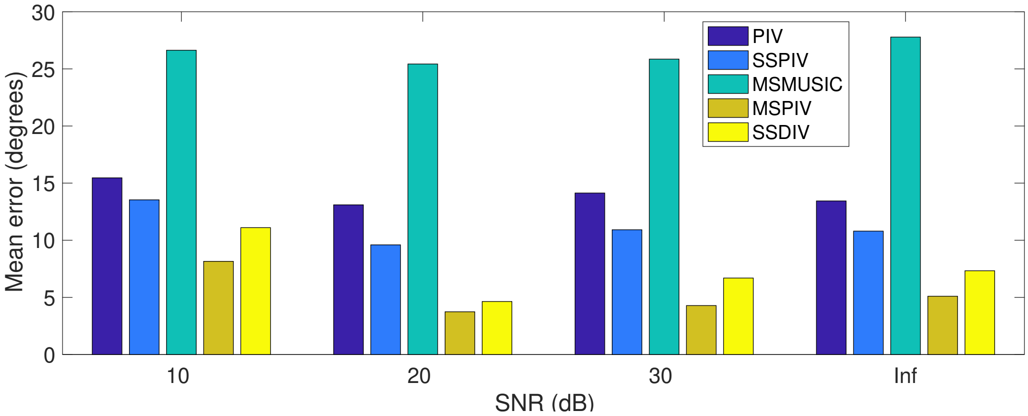


Figure 6.5: The mean error of PIV, SSPIV, MSMUSIC, MSPIV and SSDIV for varying SNR.

not steering-based. Note that DIV (and SSDIV) estimate two identical DOAs in case of a SS scenario, which makes it a robust algorithm for a SS scenario although it is a MS DOA estimator. The proposed MSPIV and SSDIV both significantly outperform the baseline techniques PIV and SSPIV with almost double accuracy due to the utilization of the narrowband MS assumption. MSPIV shows 1° to 4° better accuracy than SSDIV since it assumes more accurate number of sources per bin. Note that although some MSPIV estimates can be erroneous in TF bins with < 4 active sources, the irregular directions of the errors avoid the formation of a prominent peak in the histogram. The same phenomenon happens for SSDIV in TF bins with < 2 active sources.

The results for SNR=20 dB are used for inferential statistical analysis. There was a statistically significant difference between groups as determined by one-way ANOVA ($F(4, 383) = 36.19, p < 0.001$). A Tukey post hoc test revealed that the mean estimation error was statistically significantly lower for MSPIV ($5.73° \pm 1.45°, p < 0.001$) and SSDIV ($6.63° \pm 1.46°, p < 0.001$) compared to PIV method ($15.09° \pm 1.46°$).

6.7 Computational Complexity

This section evaluates the computational complexity of each technique. Table 6.2 presents the approximate number of real multiplications and MATLAB$^{\text{TM}}$ average running time (evaluated for comparison purposes all on the same computer, MacBook Pro with 2.6 GHz

	PIV	SSPIV	MSMUSIC	MSPIV	SSDIV
Approx. # real ×	50	50	10^6	$50N$	100
Average running time (s)	0.203	1.024	329.738	1.682	1.241

Table 6.2: Computationality of each method

Intel Core i5 processor and 8 GB 1600 MHz DDR3 memory) for each method. Note that the number of real multiplications were calculated per bin for narrowband DOA estimator only excluding the covariance matrix calculation, EVD operation and peak picking.

The PIV and MSMUSIC have the lowest and the highest computational complexity respectively as expected. SSDIV has approximately double the number of real multiplications than PIV and SSPIV while MSPIV has N times more multiplications than PIV as it performs PIV N times where N is the assumed number of active sources per bin. MSMUSIC has extremely high computational cost due to steering to all directions per bin (180 × 360 directions) excluding even the high computational cost of multi-peak picking per bin.

6.8 Conclusions

Two narrowband analytical DOA estimators based on the MS assumption have been proposed. Using EVD, MSPIV decomposes a MS scenario into multiple SS scenarios and SSDIV decomposes the MS scenario into a mixture of the two most dominant sources. The first approach performs SS-based PIV on each SS scenario whereas the second one directly estimates two DOAs from a two-source scenario. An evaluation was performed for a challenging scenario of one source being strongly masked by three other simultaneously active sources in an anechoic environment with noisy sensors. The results show an improvement of almost double the accuracy on average at the cost of double computational complexity compared to the conventional SS-based DOA estimators. They also significantly outperform the steering-based and MS-based DOA estimator, MSMUSIC with much less computational complexity.

Chapter 7

Thesis Conclusions

IN this chapter a summary and conclusions of the thesis are presented. Section 7.1 highlights its main achievements and Section 7.3 outlines some suggestions for future research.

7.1 Summary of thesis achievements

The aim of this thesis was to propose a number of techniques addressing three challenges: (1) Computationally efficient DOA estimation with maintained accuracy as high-order steering-based methods. (2) Successful source detection and DOA estimation for violated WDO assumption where sources are masked due to either increasing number of sources (covered in Chapter 5) or short/quiet activity (covered in Chapter 6). (3) Autonomous source counting method that is also reliable for extreme conditions of DOAs spatial distribution. Each proposed technique belongs to a block in the chain of wideband MS DOA estimation system shown in Figure 1.2. The main achievements are as follows:

AIV: A narrowband DOA estimation technique has been proposed which improves the accuracy and robustness to noise, reverberation and multi-source of PIV by utilization of high order harmonics and an efficient optimization of a cost function. The evaluations using simulation and real recordings prove that the proposed method outperforms the baseline and performs multi-thousand times faster than the state-of-the-art with less than

1.5° accuracy loss. This technique can be utilized in the narrowband DOA estimation block in Figure 1.2. *(Chapter 3)*

Evolutive DBSCAN: A source counting method has been proposed which employs density-based noise-robust DBSCAN clustering in an evolutionary framework. The results using generated and estimated DOAs show that the proposed technique outperforms the conventional histogram peak picking as well as the original DBSCAN and adaptive Kmeans based on AIC and BIC with $\leq 4°$ accuracy and more than 30% improvement in source counting. This technique can be employed in the last block in Figure 1.2. *(Chapter 4)*

MSEC: A SS-validity confidence metric for detection and selection of SS bins is proposed which uses the estimation consistency of initial DOA estimates based on adaptive MS assumption per time frame. The evaluations using simulations and real recordings validate the high accuracy of $< 4°$ for DOA estimation and improvement from 20% to $> 70\%$ on source detection for our proposed metric compared to the state-of-the-art metric. Using the proposed metric, a variation of MUSIC DOA estimator is proposed. It is shown that the proposed MSEC-MUSIC improves the quality of the covariance matrix for the subspace decomposition and consequently the DOA accuracy in a MS scenario. The MSEC confidence metric can be used in the post-processing block in Figure 1.2. *(Chapter 5)*

SSDIV and MSPIV: Two narrowband DOA estimators are proposed based on multi-source and two-source assumption. The first one decomposes a MS scenario into multiple SS scenarios and performs a fast analytical SS DOA estimator. The second one directly and efficiently estimates up to two DOAs per bin using low-order harmonics. The evaluation and illustrative validation show a double accuracy improvement compared to the baseline SS and MS DOA estimators. These techniques can be utilized in the narrowband DOA estimation block in Figure 1.2. *(Chapter 6)*

As a conclusion, this thesis addresses the problem of wideband MS DOA estimation using SMAs and proposes multiple solutions, each for an operating block in the system of solution shown in Figure 1.2. Three narrowband DOA estimators (AIV, MSPIV and

SSDIV) are proposed for the narrowband DOA estimation operating block based on single-, two- or multi-source assumption. A metric (MSEC) for selection of the reliable initial narrowband DOA estimates is proposed for the DOA selection operating block. And finally an autonomous source counting and source DOA extraction (Evolutive DBSCAN) is proposed for the final operating block in the chain. For each block, the evaluation results show that the proposed methods outperform the baseline and the state-of-the-art techniques in accuracy and/or computationality.

7.2 Impact and applications

Chapter 3 provides a computationally efficient narrowband DOA estimator with high accuracy and robustness close to computationally expensive state-of-the-art. The outcome method can be applied in scenarios where both accuracy and computation are important such as real-time robot audition or teleconferencing. Although with the advances in CPU and GPU processing times computationally expensive methods will eventually find their ways into real-time applications, there will be a need and preference for computationally low cost methods in some areas such as hearing aids, due to limitation on the device size and processing capabilities.

Chapter 4 presents an autonomous source counting method with reliable performance even on radical conditions. This method can have a significant impact on applications where no *priori* knowledge of the scene or sources are available such as robots in a battlefield or unknown environments.

Chapters 5 and 6 provide solutions for scenarios where WDO assumption is violated. For example, in teleconferencing or meeting diarization, there are often talkers with various loudness and/or multiple talkers who are simultaneously speaking which causes temporal and spectral masking or loss in size of TF regions with valid SS assumption. In such scenarios, MSEC and DIV can make noticeable improvement on the accuracy of localization and source detection.

7.3 Future research directions

The potential further improvements or contribution of each proposed technique are presented as follow:

AIV: A subspace variation of AIV can be studied in which the input is the first set of eigenvectors obtained from the EVD of the covariance matrix. It can improve the robustness to noise and reverberations. In addition, AIV can be extended to the narrowband MS assumption by applying it separately on each set of eigenvectors, similar to MSPIV.

Evolutive DBSCAN: The proposed source counting technique can be used and evaluated in the context of source tracking. Temporal counting of the number of active sources is a challenging task for which evolutive DBSCAN can be employed. A computationally efficient alternative of evolutive DBSCAN can also be designed using a recursive approach. Rather than performing the entire DBSCAN at each iteration, the additional core points in the new iteration can be detected.

DIV: The proposed DIV has been validated in an anechoic scenario. A reverberation-robust approach for DIV can be studied in the future. In addition, DIV can be employed and evaluated in the context of source separation as it also extracts the sources' signals in addition to sources DOAs from the mixture of two sources. The effect of various combinations of two eigenvectors for the signals' subspace can also be studied.

Bibliography

[1] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing.* Berlin, Germany: Springer-Verlag, 2016.

[2] B. Rafaely, *Fundamentals of Spherical Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer, 2015.

[3] I. Balmages and B. Rafaely, “Open-sphere designs for spherical microphone arrays,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 727–732, 2007.

[4] G. W. Elko and J. Meyer, “Spherical microphone arrays for 3D sound recordings,” in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., 2004, ch. 3, pp. 67–89.

[5] E. Fisher and B. Rafaely, “Near-field spherical microphone array processing with radial filtering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 256–265, 2011.

[6] B. Rafaely, “Analysis and design of spherical microphone arrays,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[7] B. Rafaely, B. Weiss, and E. Bachmat, “Spatial aliasing in spherical microphone arrays,” *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007.

[8] B. Rafaely, “The spherical-shell microphone array,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 740–747, May 2008.

[9] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, 2007.

[10] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[11] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, Jan. 2010, ch. 11.

[12] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[13] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 2–16, 2010.

[14] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[15] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.

[16] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[17] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudo-intensity vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[18] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, September 2015.

[19] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[20] H. Teutsch and W. Kellermann, "EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Philadelphia, PA, USA, Mar. 2005, pp. iii/89–iii/92.

[21] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Proc. Audio Eng. Soc. (AES) Convention*, New York, NY, USA, Oct. 2003, pp. 1–9.

[22] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Orlando, FL, USA, May 2002, pp. 1949–1952.

[23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[24] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 984–995, 1989.

[25] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[26] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[27] C. Chen, R. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[28] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1539–1551, October 2015.

[29] J. L. Yuxiang Hu and X. Qiu, "A maximum likelihood direction of arrival estimation method for open-sphere microphone arrays in the spherical harmonic domain," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 791–794, 2015.

[30] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Nice, France, Jul. 2014.

[31] 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. European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[32] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *IWAENC*, Seattle, WA, USA, September 2008.

[33] W. Zhang and B. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, 2010.

[34] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a laplacian mixture model," *Digital Signal Processing*, vol. 18, no. 1, pp. 66–76, January 2011.

[35] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angle distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 4629–4632.

[36] S. Rickard and Z. Yilmaz, “On the approximate W-disjoint orthogonality of speech,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 2002, pp. 529–532.

[37] H. Wang and M. Kaveh, “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, Aug. 1985.

[38] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[39] M. Ester, H. P. Krigel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial database with noise,” in *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, WA, 1996, pp. 226–231.

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

[41] D. Levin, E. A. P. Habets, and S. Gannot, “On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields,” *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[42] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, “Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[43] B. D. van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[44] S. Hafezi, A. H. Moore, and P. A. Naylor, “Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain,” *IEEE/ACM Transactions*

on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1956–1968, October 2017.

[45] ——, "Robust source counting for DOA estimation using density-based clustering," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July 2018.

[46] ——, "Multiple DOA estimation based on estimation consistency and spherical harmonic multiple signal classification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Kos, Greece, September 2017, pp. 1280–1284.

[47] ——, "Multi-source estimation consistency for improved multiple direction-of-arrival estimation," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, March 2017.

[48] ——, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017.

[49] ——, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, September 2016.

[50] ——, "3D acoustic source localization in the spherical harmonic domain based on optimized grid search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Shanghai, China, March 2016.

[51] ——, "Modelling source directivity in room impulse response simulation for spherical microphone arrays," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, Sep. 2015.

[52] S. Hafezi. Room Impulse Response for Directional source (RIRD) generator. [Online]. Available: http://www.ee.ic.ac.uk/sap/rirdgen/

[53] ——. Spherical Microphone array Impulse Response for Directional source (SMIRD) generator. [Online]. Available: http://www.ee.ic.ac.uk/sap/smirdgen/

[54] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999.

[55] B. Rafaely, "Plane-wave decomposition of the pressure on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004.

[56] M. J. Crocker and F. Jacobsen, "Sound intensity," in *Handbook of Acoustics*, M. J. Crocker, Ed. Wiley-Interscience, 1998, ch. 106, pp. 1327–1340.

[57] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132.

[58] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[59] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing. Berlin Heidelberg: Springer, 2016.

[60] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, Corpus LDC93S1, 1993.

[61] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Technical Report, Jun. 1987.

[62] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Dec. 2011. [Online]. Available: http://www.itu.int/rec/T-REC-P.56-201112-I/en

[63] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation.* Springer, 2010.

[64] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Work-*

shop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2009, pp. 221–224.

[65] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D DOA estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 96–100.

[66] A. Ram, A. Sharma, A. S. Jalall, R. Singh, and A. Agrawal, "An enhanced density based spatial clustering of applications with noise," in *IEEE International Advance Computing Conferece (IACC)*, Patiala, India, March 2009.

[67] P. Liu, D. Zhou, and N. Wu, "Vdbscan: Varied density based spatial clustering of applications with noise," in *Proc. of IEEE International Conference on Service Systems and Service Management*, Chengdu, China, 2007, pp. 1–4.

[68] C. Xiaoyun, M. Yufang, Z. Yan, and W. Ping, "Gmdbscan: Multi-density dbscan cluster based on grid," in *IEEE International Conference on e-Business Enginerring (ICEBE)*, 2008.

[69] Z. Xiong, R. Chen, Y. Zhang, and X. Zhang, "Multi-density dbscan algorithm based on density levels partitioning," *Journal of Information and Computational Science*, vol. 9, pp. 2739–2749, October 2012.

[70] O. Uncu, W. A. Gruver, D. B. Kotak, D. Sabaz, Z. Alibhai, and C. Ng, "Gridbscan: Grid density-based spatial clustering of applications with noise," in *IEEE Intl. Conf. on Systems, Man and Cybernetics*, Taipei, Taiwan, October 2006.

[71] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," vol. 123, pp. 2136–2147, 2008.

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

sures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 2625–2628.

[73] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 285–288.

[74] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[75] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval.* Cambridge, UK: Cambridge University Press, 2008.

Appendix A

Theoretical error of AIV

For the formulation in this Section, (τ, k) are omitted for notational simplicity. In a noise-free scenario as in (2.10), consider $\tilde{a}_{lm} = SY_{lm}^{*}(\Omega_u)$ as the clean eigenbeams of the direct path. For an arbitrary look direction Ω, the clean eigenbeam error function is

$$\begin{aligned}\tilde{E}_{lm}(\Omega) = \tilde{a}_{lm} - SY_{lm}^{*}(\Omega) \\ = SY_{lm}^{*}(\Omega_u) - SY_{lm}^{*}(\Omega_u) g_{lm}^{*}(\Omega_u, \triangle\Omega) \\ = SY_{lm}^{*}(\Omega_u)\left(1 - g_{lm}^{*}(\Omega_u, \triangle\Omega)\right), \quad \text{(A.1)}\end{aligned}$$

and by using (2.2)

$$g_{lm}(\Omega_u, \triangle\Omega) = \frac{P_{lm}(\cos(\theta_u + \triangle\theta))}{P_{lm}(\cos(\theta_u))} e^{im\triangle\varphi}, \quad \text{(A.2)}$$

where $\Delta\Omega = \angle(\Omega, \Omega_u)$ and $\angle(.)$ denotes the angle between two directions in degree.

Now assume the noisy scenario where the noisy cost function in (3.4) is decomposed into clean $\tilde{E}_{lm}$ and the noise eigenbeam n_{lm} resulting in

$$\begin{aligned}\Psi(\Omega) = \sum_{lm} \mid \tilde{E}_{lm}(\Omega) - n_{lm} \mid^2 = \tilde{\Psi}(\Omega) + C_n \\ + 2\sum_{lm} \mid \tilde{E}_{lm}(\Omega) \mid\mid n_{lm} \mid \cos(\Gamma_{lm}(\Omega)), \quad \text{(A.3)}\end{aligned}$$

where $\tilde{\Psi}(\Omega) = \sum\limits_{lm} \mid \tilde{E}_{lm}(\Omega) \mid^2$ is the noise-free cost function, $C_n = \sum\limits_{lm} \mid n_{lm} \mid^2$ is a noise-based constant, $\sum\limits_{lm} = \sum\limits_{l} \sum_{m=-l}^{l}$ and

$$\begin{aligned}\Gamma_{lm}(\Omega) &= \angle \tilde{E}_{lm}(\Omega) - \angle n_{lm} \\ &= \angle S + \angle Y_{lm}^{*}(\Omega_u) + \angle\left(1 - g_{lm}^{*}(\Omega_u, \triangle\Omega)\right) - \angle n_{lm}, \quad \text{(A.4)}\end{aligned}$$

where $\angle(.)$ denotes the phase of complex number.

The derivative of the noisy cost function in (A.3) is

$$\Psi^{'}(\Omega) = \tilde{\Psi}^{'}(\Omega) + 2\sum_{lm} \mid n_{lm} \mid \left(\mid \tilde{E}_{lm}(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega)\right)\right)^{'}, \quad \text{(A.5)}$$

where $(.)^{'} = \frac{d}{d\Omega}(.)$ is the derivative operator.

For $\triangle\theta \approx 0$ in (A.2) $g_{lm}^{*}(\Omega_u, \triangle\Omega) = e^{-im\triangle\varphi}$, which results in

$$\left(1 - g_{lm}^{*}(\Omega_u, \triangle\Omega)\right) = 2\sin\left(\frac{m\triangle\varphi}{2}\right) e^{i\left(\frac{\pi}{2} - \frac{m\triangle\varphi}{2}\right)}. \quad \text{(A.6)}$$

Substituting (A.6) into (A.1)

$$\tilde{\Psi}^{'}(\Omega) = \sum_{lm} 2m \mid SY_{lm}^{*}(\Omega) \mid^{2} \sin\left(m\triangle\varphi\right), \quad \text{(A.7)}$$

and

$$\begin{aligned}&\left(\mid \tilde{E}_{lm}(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega)\right)\right)^{'} = \\ &\qquad m \mid SY_{lm}^{*}(\Omega) \mid \cos\left(\Gamma_{lm}(\Omega) - \frac{m\triangle\varphi}{2}\right). \quad \text{(A.8)}\end{aligned}$$

At the optimal look direction Ω_s, we have $\Psi^{'}(\Omega_s) = 0$, which by substituting (A.7) and (A.8) into (A.5) gives

$$\begin{aligned}&\sum_{lm} \mid SY_{lm}^{*}(\Omega_u) \mid^{2} m\sin\left(m\triangle\varphi_s\right) = \\ &\qquad \sum_{lm} \mid n_{lm} \mid\mid SY_{lm}^{*}(\Omega_u) \mid m\sin\left(\Lambda_{lm}(S, \Omega_u, n_{lm}) - m\triangle\varphi_s\right), \quad \text{(A.9)}\end{aligned}$$

where $\Lambda_{lm}(S, \Omega_u, n_{lm}) = \angle S + \angle Y^*_{lm}(\Omega_u) - \angle n_{lm}$ is the combined phase from the direct path and the noise eigenbeams.

Simplifying (A.9) gives

$$\sum_{j=0}^{L} \sqrt{A_j^2 + B_j^2} \sin\left(j \triangle \varphi_s - \arctan\left(\frac{B_j}{A_j}\right)\right) = 0, \tag{A.10}$$

where

$$A_j = \frac{1}{\mu} \sum_{l=0}^{L} \sum_{|m|=j} \mid m \mid\mid \tilde{a}_{lm} \mid^2 (\mu + \cos(\Lambda_{lm})), \tag{A.11}$$

and

$$B_j = \frac{1}{\mu} \sum_{l=0}^{L} \sum_{|m|=j} m \mid \tilde{a}_{lm} \mid^2 \sin(\Lambda_{lm}), \tag{A.12}$$

where $\mu^2 = \frac{|SY^*_{lm}(\Omega_u)|^2}{|n_{lm}|^2} = \frac{|\tilde{a}_{lm}|^2}{|n_{lm}|^2}$ is the SNR for spatially white noise (equal noise level across all microphones) and is fixed for all (l, m). Note that (S, Ω_u, n_{lm}) are omitted from Λ_{lm} for notational simplicity.

Appendix B

Table of Gradients for AIV

$l(m)$	$Y^*_{lm}(\theta,\varphi)$	$\nabla\{\mid Y^*_{lm}(\Omega)\mid^2\}$
0(0)	$\sqrt{\frac{1}{4\pi}}$	0
1(-1)	$\sqrt{\frac{3}{8\pi}}\sin(\theta)e^{i\varphi}$	$(\frac{3}{8\pi})\sin(2\theta)\hat{\theta}$
1(0)	$\sqrt{\frac{3}{4\pi}}\cos(\theta)$	$(\frac{-3}{4\pi})\sin(2\theta)\hat{\theta}$
1(1)	$-\sqrt{\frac{3}{8\pi}}\sin(\theta)e^{-i\varphi}$	$(\frac{3}{8\pi})\sin(2\theta)\hat{\theta}$
2(-2)	$\sqrt{\frac{15}{32\pi}}\sin^2(\theta)e^{2i\varphi}$	$(\frac{15}{32\pi})4\sin^3(\theta)\cos(\theta)\,\hat{\theta}$
2(-1)	$\sqrt{\frac{15}{8\pi}\frac{1}{2}}\sin(2\theta)e^{i\varphi}$	$(\frac{15}{8\pi})\sin(2\theta)\cos(2\theta)\,\hat{\theta}$
2(0)	$\sqrt{\frac{5}{16\pi}}(3\cos^2(\theta)-1)$	$(\frac{-15}{8\pi})\sin(2\theta)\big(3\cos^2(\theta)-1\big)\,\hat{\theta}$
2(1)	$-\sqrt{\frac{15}{8\pi}\frac{1}{2}}\sin(2\theta)e^{-i\varphi}$	$(\frac{15}{8\pi})\sin(2\theta)\cos(2\theta)\,\hat{\theta}$
2(2)	$\sqrt{\frac{15}{32\pi}}\sin^2(\theta)e^{-2i\varphi}$	$(\frac{15}{32\pi})4\sin^3(\theta)\cos(\theta)\,\hat{\theta}$
3(-3)	$\sqrt{\frac{35}{64\pi}}\sin^3(\theta)e^{3i\varphi}$	$(\frac{35}{64\pi})3\sin(2\theta)\sin^4(\theta)\hat{\theta}$
3(-2)	$\sqrt{\frac{105}{32\pi}}\sin^2(\theta)$ $\times\cos(\theta)\,e^{2i\varphi}$	$(\frac{105}{32\pi})\sin(2\theta)\sin^2(\theta)$ $\times(2-3\sin^2(\theta))\hat{\theta}$
3(-1)	$\sqrt{\frac{21}{64\pi}}\sin(\theta)$ $\times(5\cos^2(\theta)-1)e^{i\varphi}$	$(\frac{21}{64\pi})(4-5\sin^2(\theta))$ $\times(4-15\sin^2(\theta))\sin(2\theta)\hat{\theta}$
3(0)	$\sqrt{\frac{7}{16\pi}}(5\cos^3(\theta)$ $-3\cos(\theta))$	$(\frac{7}{16\pi})3(2-5\sin^2(\theta))$ $\times(-4+5\sin^2(\theta))\sin(2\theta)\hat{\theta}$
3(1)	$-\sqrt{\frac{21}{64\pi}}\sin(\theta)$ $\times(5\cos^2(\theta)-1)e^{-i\varphi}$	$(\frac{21}{64\pi})(4-5\sin^2(\theta))$ $\times(4-15\sin^2(\theta))\sin(2\theta)\hat{\theta}$
3(2)	$\sqrt{\frac{105}{32\pi}}\sin^2(\theta)$ $\times\cos(\theta)\,e^{-2i\varphi}$	$(\frac{105}{32\pi})\sin(2\theta)\sin^2(\theta)$ $\times(2-3\sin^2(\theta))\hat{\theta}$
3(3)	$-\sqrt{\frac{35}{64\pi}}\sin^3(\theta)e^{-3i\varphi}$	$(\frac{35}{64\pi})3\sin(2\theta)\sin^4(\theta)\hat{\theta}$

Table B.1: Gradient of the first part of the cost function for up to the third order.

$l(m)$	$\nabla\left\{\mid Y^*_{lm}(\Omega)\mid\cos\left(\lambda_{lm}-\angle Y^*_{lm}(\Omega)\right)\right\}$
0(0)	0
1(-1)	$\sqrt{\frac{3}{8\pi}}\left\{\cos(\theta)\cos\left(\lambda_{1(-1)}-\varphi\right)\hat{\theta}+\sin(\theta)\sin\left(\lambda_{1(-1)}-\varphi\right)\hat{\varphi}\right\}$
1(0)	$-\frac{1}{2}\sqrt{\frac{3}{\pi}}\sin(\theta)\cos\left(\lambda_{10}\right)\hat{\theta}$
1(1)	$-\sqrt{\frac{3}{8\pi}}\left\{\cos(\theta)\cos\left(\lambda_{1(1)}+\varphi\right)\hat{\theta}-\sin(\theta)\sin\left(\lambda_{1(1)}+\varphi\right)\hat{\varphi}\right\}$
2(-2)	$\sqrt{\frac{15}{32\pi}}\left\{\sin(2\theta)\cos\left(\lambda_{2(-2)}-2\varphi\right)\hat{\theta}+2\sin^2(\theta)\sin\left(\lambda_{2(-2)}-2\varphi\right)\hat{\varphi}\right\}$
2(-1)	$\sqrt{\frac{15}{8\pi}}\left\{\cos(2\theta)\cos\left(\lambda_{2(-1)}-\varphi\right)\hat{\theta}+\frac{1}{2}\sin(2\theta)\sin\left(\lambda_{2(-1)}-\varphi\right)\hat{\varphi}\right\}$
2(0)	$-\sqrt{\frac{5}{16\pi}}3\sin(2\theta)\cos\left(\lambda_{20}\right)\hat{\theta}$
2(1)	$-\sqrt{\frac{15}{8\pi}}\left\{\cos(2\theta)\cos\left(\lambda_{2(1)}+\varphi\right)\hat{\theta}-\frac{1}{2}\sin(2\theta)\sin\left(\lambda_{2(1)}+\varphi\right)\hat{\varphi}\right\}$
2(2)	$\sqrt{\frac{15}{32\pi}}\left\{\sin(2\theta)\cos\left(\lambda_{2(2)}+2\varphi\right)\hat{\theta}-2\sin^2(\theta)\sin\left(\lambda_{2(2)}+2\varphi\right)\hat{\varphi}\right\}$
3(-3)	$\sqrt{\frac{35}{64\pi}}\{3\sin^2(\theta)\cos(\theta)\cos\left(\lambda_{3(-3)}-3\varphi\right)\hat{\theta}$ $+3\sin^3(\theta)\sin\left(\lambda_{3(-3)}-3\varphi\right)\hat{\varphi}\}$
3(-2)	$\sqrt{\frac{105}{32\pi}}\{\sin(\theta)(2-3\sin^2(\theta))\cos\left(\lambda_{3(-2)}-2\varphi\right)\hat{\theta}$ $+2\sin^2(\theta)\cos(\theta)\sin\left(\lambda_{3(-2)}-2\varphi\right)\hat{\varphi}\}$
3(-1)	$\sqrt{\frac{21}{64\pi}}\{\cos(\theta)(4-15\sin^2(\theta))\cos\left(\lambda_{3(-1)}-\varphi\right)\hat{\theta}$ $+\sin(\theta)\,(5\cos^2(\theta)-1)\sin\left(\lambda_{3(-1)}-\varphi\right)\hat{\varphi}\}$
3(0)	$\sqrt{\frac{7}{16\pi}}3\sin(\theta)(-4+5\sin^2(\theta))\cos\left(\lambda_{3(0)}\right)\hat{\theta}$
3(1)	$-\sqrt{\frac{21}{64\pi}}\{\cos(\theta)(4-15\sin^2(\theta))\cos\left(\lambda_{3(1)}+\varphi\right)\hat{\theta}$ $-\sin(\theta)\,(5\cos^2(\theta)-1)\sin\left(\lambda_{3(1)}+\varphi\right)\hat{\varphi}\}$
3(2)	$\sqrt{\frac{105}{32\pi}}\{\sin(\theta)(2-3\sin^2(\theta))\cos\left(\lambda_{3(2)}+2\varphi\right)\hat{\theta}$ $-2\sin^2(\theta)\cos(\theta)\sin\left(\lambda_{3(2)}+2\varphi\right)\hat{\varphi}\}$
3(3)	$-\sqrt{\frac{35}{64\pi}}\{3\sin^2(\theta)\cos(\theta)\cos\left(\lambda_{3(3)}+3\varphi\right)\hat{\theta}$ $-3\sin^3(\theta)\sin\left(\lambda_{3(3)}+3\varphi\right)\hat{\varphi}\}$

Table B.2: Gradient of the second part of the cost function for up to the third order.

Appendix C

DIV Derivation

For the formulation in this section, (τ, k) are omitted for notational simplicity. In a TF bin, consider two simultaneously active sources with plane-wave signals S_1 and S_2, assuming $S_1 \neq -S_2$ (no opposite phase with equal amplitude), and their DOAs (θ_1, φ_1) and (θ_2, φ_2) respectively. Writing (2.10) in a noise-free situation for up to $L = 1$ gives

$$\bar{a}_{0(0)} = S_1 + S_2, \tag{C.1}$$

$$\bar{a}_{1(-1)} = S_1 \sin(\theta_1) e^{i\varphi_1} + S_2 \sin(\theta_2) e^{i\varphi_2}, \tag{C.2}$$

$$\bar{a}_{1(0)} = S_1 \cos(\theta_1) + S_2 \cos(\theta_2), \tag{C.3}$$

$$\bar{a}_{1(+1)} = S_1 \sin(\theta_1) e^{-i\varphi_1} + S_2 \sin(\theta_2) e^{-i\varphi_2}, \tag{C.4}$$

where

$$\bar{a}_{lm} = \frac{a_{lm}}{\sqrt{\frac{(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}}} \tag{C.5}$$

are the eigenbeams compensated by the square root component in (2.2) for notational simplicity.

Using (C.1) and (C.3)

$$S_2 = \bar{a}_{0(0)} - S_1, \tag{C.6}$$

$$S_1 = \frac{\bar{a}_{1(0)} - \bar{a}_{0(0)} \cos(\theta_2)}{\cos(\theta_1) - \cos(\theta_2)}. \tag{C.7}$$

Summing and subtracting (C.2) with (C.4) respectively gives

$$B^{+} = S_1 \sin(\theta_1)\cos(\varphi_1) + S_2 \sin(\theta_2)\cos(\varphi_2), \tag{C.8}$$

$$B^{-} = S_1 \sin(\theta_1)\sin(\varphi_1) + S_2 \sin(\theta_2)\sin(\varphi_2), \tag{C.9}$$

where $B^{+} = (\bar{a}_{1(-1)} + \bar{a}_{1(+1)})/2$ and $B^{-} = (\bar{a}_{1(-1)} - \bar{a}_{1(+1)})/2i$.

Using (C.6) and (C.7), substituting S_1 and S_2 both as a function of θ_1 and θ_2 into (C.8) and (C.9) results in two complex equations with four real unknowns (θ and φ for sources 1 and 2). Splitting the real and imaginary parts of the two complex equations (C.8) and (C.9) gives four real equations which can be simplified into

$$\begin{bmatrix} E_\theta & F_\theta \\ H_\theta & J_\theta \end{bmatrix} \begin{bmatrix} \cos(\varphi_1) \\ \cos(\varphi_2) \end{bmatrix} = \begin{bmatrix} D_\theta \\ G_\theta \end{bmatrix}, \tag{C.10}$$

$$\begin{bmatrix} E_\theta & F_\theta \\ H_\theta & J_\theta \end{bmatrix} \begin{bmatrix} \sin(\varphi_1) \\ \sin(\varphi_2) \end{bmatrix} = \begin{bmatrix} K_\theta \\ L_\theta \end{bmatrix}, \tag{C.11}$$

where

$$\begin{cases} E_\theta = & \sin(\theta_1)\left(\Re\left\{\bar{a}_{1(0)}\right\} - \Re\left\{\bar{a}_{0(0)}\right\}\cos(\theta_2)\right), \\ F_\theta = & \sin(\theta_2)\left(\Re\left\{\bar{a}_{0(0)}\right\}\cos(\theta_1) - \Re\left\{\bar{a}_{1(0)}\right\}\right), \\ H_\theta = & \sin(\theta_1)\left(\Im\left\{\bar{a}_{1(0)}\right\} - \Im\left\{\bar{a}_{0(0)}\right\}\cos(\theta_2)\right), \\ J_\theta = & \sin(\theta_2)\left(\Im\left\{\bar{a}_{0(0)}\right\}\cos(\theta_1) - \Im\left\{\bar{a}_{1(0)}\right\}\right), \\ D_\theta = & \Re\left\{B^{+}\right\}\left(\cos(\theta_1) - \cos(\theta_2)\right), \\ G_\theta = & \Im\left\{B^{+}\right\}\left(\cos(\theta_1) - \cos(\theta_2)\right), \\ K_\theta = & \Re\left\{B^{-}\right\}\left(\cos(\theta_1) - \cos(\theta_2)\right), \\ L_\theta = & \Im\left\{B^{-}\right\}\left(\cos(\theta_1) - \cos(\theta_2)\right). \end{cases} \tag{C.12}$$

Using linear matrix algebra

$$\begin{cases} \cos(\varphi_1) = & \frac{JD-FG}{EJ-FH} \\ \sin(\varphi_1) = & \frac{JK-FL}{EJ-FH} \end{cases}, \tag{C.13}$$

$$\begin{cases} \cos(\varphi_2) = & \frac{EG-HD}{EJ-FH} \\ \sin(\varphi_2) = & \frac{EL-HK}{EJ-FH} \end{cases}. \tag{C.14}$$

Note that index $(.)_\theta$ is omitted from now on for notational simplicity. Using the Pythagorean identity $(\sin^2(\varphi) + \cos^2(\varphi) = 1)$ for (C.13) and (C.14) gives a system of two equations

$$\begin{cases} (JD - FG)^2 + (JK - FL)^2 & = (EJ - FH)^2, \\ (EG - HD)^2 + (EL - HK)^2 & = (EJ - FH)^2, \end{cases} \tag{C.15}$$

where unknowns are θ_1 and θ_2.

Substituting E, F, H, J, D, G, K, L from (C.12) into (C.15) and then solving them for θ_1 results in a quadratic equation

$$C_2 \cos^2(\theta_1) - C_1 \cos(\theta_1) + C_0 = 0, \tag{C.16}$$

where

$$\begin{cases} C_2 = & \vartheta_R A_I^2 + \vartheta_I A_R^2 - 2\chi A_I A_R + (a_R A_I - a_I A_R)^2, \\ C_1 = & 2\left(\vartheta_R A_I a_I + \vartheta_I A_R a_R - \chi\left(a_R A_I + a_I A_R\right)\right), \\ C_0 = & \vartheta_R a_I^2 + \vartheta_I a_R^2 - 2\chi a_I a_R - (a_R A_I - a_I A_R)^2, \end{cases} \tag{C.17}$$

and

$$\begin{cases} \vartheta_R = & \left(\Re\{B^+\}\right)^2 + \left(\Re\{B^-\}\right)^2, \\ \vartheta_I = & \left(\Im\{B^+\}\right)^2 + \left(\Im\{B^-\}\right)^2, \\ \chi = & \Re\{B^+\}\Im\{B^+\} + \Re\{B^-\}\Im\{B^-\}, \end{cases} \tag{C.18}$$

where $a_R = \Re\left(\bar{a}_{1(0)}\right)$, $a_I = \Im\left(\bar{a}_{1(0)}\right)$, $A_R = \Re\left(\bar{a}_{0(0)}\right)$ and $A_I = \Im\left(\bar{a}_{0(0)}\right)$.

The two solutions for (C.16) give the inclination of sources

$$\begin{cases} \theta_1 = & \arccos(\frac{C_1+\sqrt{\triangle}}{2C_2}), \\ \theta_2 = & \arccos(\frac{C_1-\sqrt{\triangle}}{2C_2}), \end{cases} \tag{C.19}$$

where

$$\triangle = C_1^2 - 4C_2C_0. \tag{C.20}$$

Note that since inclination is within $[0, \pi]$ there is no sign ambiguity for θ_1 and θ_2.

Having calculated inclinations θ_1 and θ_2, then E, F, H, J, D, G, K, L are calculated using (C.12). The tangent of azimuths are calculated using (C.13) and (C.14)

$$\begin{cases} \tan(\varphi_1) = & \frac{JK-FL}{JD-FG}, \\ \tan(\varphi_2) = & \frac{EL-HK}{EG-HD}. \end{cases} \tag{C.21}$$

Since the azimuths are within $[0, 2\pi]$, the sign of cosine of azimuths from (C.13) and (C.14) are used to solve the sign ambiguity of φ_1 and φ_2 giving

$$\begin{cases} \varphi_1 = & \arctan\left(\frac{JK-FL}{JD-FG}\right) + \left(1 - \text{sgn}(\frac{JD-FG}{EJ-FH})\right)\frac{\pi}{2}, \\ \varphi_2 = & \arctan\left(\frac{EL-HK}{EG-HD}\right) + \left(1 - \text{sgn}(\frac{EG-HD}{EJ-FH})\right)\frac{\pi}{2}, \end{cases} \tag{C.22}$$

where sgn(.) denotes the sign operator.

Note that such analytical solution may have no or only one valid answer if $\triangle \leq 0$ in (C.20) or $|\cos(\theta_1)| > 1$ in (C.16).

Special case of equal inclinations

In case of equal inclinations for both sources, $\theta_1 = \theta_2$, (C.7) is undefined and therefore a different solution is required.

Let $\theta_1 = \theta_2 = \theta$, then (C.3) can be rewritten as

$$\bar{a}_{1(0)} = \cos\left(\theta\right)\left(S_1 + S_2\right). \tag{C.23}$$

Using (C.1) and (C.23)

$$\theta = \theta_1 = \theta_2 = \arccos\left(\Re(\frac{\bar{a}_{1(0)}}{\bar{a}_{0(0)}})\right). \tag{C.24}$$

Having obtained θ, (C.8) and (C.9) can be rewritten as

$$\bar{B}^+ = S_1 \cos\left(\varphi_1\right) + S_2 \cos\left(\varphi_2\right), \tag{C.25}$$

$$\bar{B}^- = S_1 \sin\left(\varphi_1\right) + S_2 \sin\left(\varphi_2\right), \tag{C.26}$$

where $\bar{B}^{\pm} = B^{\pm}/\sin(\theta)$.

From (C.25), (C.26) and (C.1)

$$S_1 = \frac{\bar{a}_{0(0)} \cos\left(\varphi_2\right) - \bar{B}^+}{\cos\left(\varphi_2\right) - \cos\left(\varphi_1\right)}, \tag{C.27}$$

$$S_2 = \frac{\bar{B}^+ - \bar{a}_{0(0)} \cos\left(\varphi_1\right)}{\cos\left(\varphi_2\right) - \cos\left(\varphi_1\right)}. \tag{C.28}$$

Substituting S_1 and S_2, from (C.27) and (C.28), into (C.26) and then simplifying it results in

$$\bar{m} X_x + \bar{n} Y_x + \bar{q} = Z_x, \tag{C.29}$$

where the known coefficients are

$$\begin{cases} \bar{m} = & \frac{-(\bar{a}_{1(-1)}+\bar{a}_{1(+1)})\sqrt{\left(\bar{a}_{0(0)}\right)^2-\left(\bar{a}_{1(0)}\right)^2}}{\bar{a}_{1(-1)}\bar{a}_{1(+1)}-\left(\bar{a}_{0(0)}\right)^2+\left(\bar{a}_{1(0)}\right)^2}, \\ \bar{n} = & \frac{\bar{a}_{1(-1)}\bar{a}_{1(+1)}+\left(\bar{a}_{0(0)}\right)^2-\left(\bar{a}_{1(0)}\right)^2}{\bar{a}_{1(-1)}\bar{a}_{1(+1)}-\left(\bar{a}_{0(0)}\right)^2+\left(\bar{a}_{1(0)}\right)^2}, \\ \bar{q} = & \frac{(\bar{a}_{1(-1)}-\bar{a}_{1(+1)})^2}{2(\bar{a}_{1(-1)}\bar{a}_{1(+1)}-\left(\bar{a}_{0(0)}\right)^2+\left(\bar{a}_{1(0)}\right)^2)}, \end{cases} \tag{C.30}$$

and the unknowns are

$$X_x = \cos(\varphi_1) + \cos(\varphi_2), \tag{C.31}$$

$$Y_x = 1 + \cos(\varphi_1)\cos(\varphi_2), \tag{C.32}$$

$$Z_x = \sin(\varphi_1)\sin(\varphi_2). \tag{C.33}$$

Splitting the real and imaginary parts of (C.29) result into two real equations which, using linear matrix algebra, give

$$\begin{bmatrix} X_x \\ Y_x \end{bmatrix} = \begin{bmatrix} \bar{m}_R & \bar{n}_R \\ \bar{m}_I & \bar{n}_I \end{bmatrix}^{-1} \begin{bmatrix} Z_x - \bar{q}_R \\ -\bar{q}_I \end{bmatrix}, \tag{C.34}$$

where $(.)_R = \Re(.)$ and $(.)_I = \Im(.)$.

On the other hand Pythagorean identity ($\sin^2(\varphi) + \cos^2(\varphi) = 1$) for (C.31), (C.32) and (C.33) gives

$$Y_x^2 - X_x^2 = Z_x^2. \tag{C.35}$$

Substituting X_x and Y_x as a function of Z_x from (C.34) into (C.35) results in a quadratic equation

$$B_2(Z_x - \bar{q}_R)^2 + B_1(Z_x - \bar{q}_R) + B_0 = 0, \tag{C.36}$$

where

$$\begin{cases} B_2 = & (\bar{m}_I^2 - \bar{n}_I^2) - \bar{d}^2, \\ B_1 = & \bar{q}_I(\bar{m}_R\bar{m}_I - \bar{n}_R\bar{n}_I) - 2\bar{q}_R\bar{d}^2, \\ B_0 = & \bar{q}_I^2(\bar{m}_R^2 - \bar{n}_R^2) - \bar{q}_R^2\bar{d}^2, \end{cases} \tag{C.37}$$

and $\bar{d} = \bar{m}_R\bar{n}_I - \bar{m}_I\bar{n}_R$. Then Z_x is obtained as

$$Z_x = \sin(\varphi_1)\sin(\varphi_2) = \frac{-B_1 \pm \sqrt{\triangle_Z}}{2B_2} + \bar{q}_R, \tag{C.38}$$

where $\triangle_Z = (B_1)^2 - 4B_2B_0$. Having obtained Z_x, then X_x and Y_x are obtained using

(C.34). Using (C.31) and (C.32), the cosines of the azimuths are

$$\begin{cases} \cos(\varphi_1) = & (X_x + \sqrt{\triangle_\varphi})/2, \\ \cos(\varphi_2) = & (X_x - \sqrt{\triangle_\varphi})/2, \end{cases} \tag{C.39}$$

where $\triangle_\varphi = (X_x)^2 - 4(Y_x - 1)$. In order to overcome the problem of sign ambiguity of φ (where $\varphi \in (-\pi, \pi]$ or the ambiguity between φ and $2\pi - \varphi$ where $\varphi \in [0, 2\pi)$) having its cosine, the following procedure can be done. Substituting the cosine of the azimuths in (C.39) into (C.27) and (C.28) gives S_1 and S_2 which along with (C.26) and (C.33) can be used to obtain the correct sign for φ_1 and φ_2.