Parametric Models and Algorithms for Direction-of-Arrival Estimation

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in
Electronics and Communication Engineering

by

Ruchi Pandey
2018802008
ruchi.pandey@research.iiit.ac.in

Advisor: Dr. Santosh Nannuru

International Institute of Information Technology Hyderabad
500 032, India

July 2024

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis titled ***Parametric Models and Algorithms for DOA Estimation*** by Ruchi Pandey has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Santosh Nannuru

Acknowledgements

I want to take this opportunity to express my sincere gratitude and appreciation to several incredible people in my life. Without their tremendous guidance, mentoring, help, care, and love, I could not have been what I am today.

I extend my heartfelt gratitude to the universe that provided me the opportunity to embark on this journey of learning and growth under the excellent guidance of my advisor, Prof. Santosh Nannuru, at the Signal Processing and Communications Research Center (SPCRC), IIIT Hyderabad. His constant inspiration, support, and motivating presence have been instrumental throughout my doctoral pursuit, and my gratitude toward him knows no bounds. He serves as my unwavering support system, akin to an umbrella fostering an environment where I can absorb knowledge, learn, and thrive. Words fall short of expressing my profound thanks to him; he embodies an outstanding advisor, an exceptional individual, and an exemplary researcher. I credit my current standing to his invaluable mentorship.

I sincerely thank Dr. Santosh for facilitating collaborations with distinguished researchers and express my gratitude to Prof. Peter Gerstoft (UCSD) for his timely, invaluable guidance. Special thanks to Dr. Huy Phan, a former professor at Queen Mary College London, for his insightful mentorship during the online internship. Alongside my collaborators, I am thankful to the professors at SPCRC for giving valuable feedback during white Friday presentations, allowing me to participate in various events, and for their continual guidance. I reserve a special acknowledgement for Prof. Sachin, whose support has made the lab environment warm and welcoming, making the lab feel like a second home.

I am profoundly grateful to the International Institute of Information Technology Hyderabad for providing excellent infrastructure and research environments. Special thanks to the Student Travel Grant Committee (STGC) at IIIT-H for providing travel grants to attend multiple conferences. I am also thankful to the IEEE Signal Processing Society (SPS) for providing a travel grant for presenting our work at ICASSP 2022 and to the European Association for Signal Processing (EURASIP) for supporting my presentations at EUSIPCO 2022 and EUSIPCO 2023.

My sincere thanks to all my SPCRC colleagues and friends. Reflecting on my formative years, my friends played an integral role in boosting my confidence and imparting invaluable lessons. I extend my heartfelt thanks to Zakir, Anish, Sai Ganesh, Krishna, Sudeepini, Purva Dii, Madhuri Ma'am, Shiva Anna, Anoushka, Ayu, Shreyas, Souradeep, Karthik, Sumanth, Rajashekar, Om, Sasanka, Usha, and Spanddhana for the joyful and memorable times at IIIT Hyderabad. Special acknowledgement to Dr. Jyoti Maggu and Dr. Shelly Vishwakarma for their mentorship across various facets of life. The best moments at IIIT-H campus were possible due to these special individuals: Ayush, Kali, Nilesh, and

Rishi, who became my second family. Their support, love, pampering, shared joy, laughter, and tears enriched my journey. I am grateful to God for blessing me with Ayush, an incredibly special person whose unwavering support has been a constant throughout and beyond this academic journey.

Finally, this thesis is a testament to the boundless love, support, and understanding of my family: my father, Mr. Pravin Pandey; my mother, Mrs. Shubhada Pandey; my elder sister, Mrs. Richa Tiwari; and my dearest brother, Mr. Aditya Pandey. Their constant motivation and encouragement in tough times have shaped my journey. They celebrated every small victory and showed immense love and courage in supporting me through rejections. This thesis is as much theirs as mine, and I dedicate it with profound love and gratitude to my parents and late grandparents.

Ruchi Pandey

Abstract

The ability to selectively focus on desired sounds in noisy environments poses a significant challenge with broad applications, including smart devices, driver assistance systems, smart homes, video conferencing, drones, and hearing aids. Acoustic source localization involves identifying the position of a sound source amidst various factors like reflections, reverberation, and background noise. While extensively studied, acoustic source localization remains an active area of research due to its diverse applications. Existing localization algorithms face several challenges that limit their effectiveness and practicality. These challenges include reliance on narrowband models, computational efficiency, adaptability to non-stationary targets, robustness against noise and reverberations, high-resolution localization, and distinguishing between correlated sources. Overcoming these challenges is crucial for the development of advanced localization algorithms that enhance accuracy, efficiency, and reliability in practical scenarios.

This thesis is divided into two main parts. Firstly, a comprehensive performance analysis is conducted to evaluate various localization algorithms using real-world datasets, aiming to gain a deep understanding of their capabilities. Secondly, a novel technique called trajectory localization (TL) is proposed, which enables accurate estimation of complex trajectories of multiple moving sources simultaneously, eliminating the need for tracking filters.

The technical contributions of this thesis include experimental validation of existing localization algorithms and the development of wideband signal models and algorithms on real-world recordings. Deep learning architecture is introduced that incorporates direction of arrival (DOA) derivatives for improving the temporal continuity of DOA, hence resulting in smoother source trajectories. Next, we develop parametric models and algorithms for joint localization and tracking tasks and explore various trajectory localization algorithms. The effectiveness of the proposed algorithms is demonstrated through their application to real-world recordings in challenging scenarios. Moreover, the proposed models and algorithms have the potential to extend beyond sound waves and be applied to other data types, such as radio waves, expanding their impact across various applications.

Contents

List of Figures

List of Tables

Abbreviations

ANM	atomic norm minimization
BP	basis pursuit
CBF	conventional beamforming
CNN	convolutional neural networks
CRNN	convolutional recurrent neural network
CS	compressive sensing
DCASE	detection and classification of acoustic scenes and events
DNN	deep neural network
DOA	direction of arrival
FFNN	feedforward neural networks
GCC	generalized cross-correlation
GCC-PHAT	generalized cross-correlation with phase transform
LASSO	least absolute shrinkage and selection operator
LOCATA	localization and tracking
MAE	mean absolute error
MAP	maximum-a posteriori
MMV	multiple measurement vector
MSE	mean squared error
MUSIC	multiple signal classification
NOMP	Newtonized orthogonal matching pursuit
NSP	nullspace property
OMP	orthogonal matching pursuit
RIP	restricted isometric property
RMSE	root mean squared error
RNN	recurrent neural networks
SBL	sparse Bayesian learning
SELD	sound event localization and detection
SFW	sliding Frank-Wolfe
SMV	single measurement vector

SNR	signal-to-noise ratio
SRP	steered response power
SRP-PHAT	steered response power with phase transform
SSL	sound source localization
Std Dev	standard deviation
STFT	short-time Fourier transform
TL	trajectory localization
ULA	uniform linear array
VAD	voice activity detector

Symbols

α	trajectory parameter
$\mathbb{E}[\cdot]$	expectation measure
d	inter-sensor spacing
γ	variance of source amplitude
σ	variance of noise
$\mathbf{\Gamma}$	covariance matrix of source amplitude
$\mathbf{Y}$	observation/measurement matrix
$\mathbf{A}_{\text{sv}}$	sensing matrix
$\mathbf{A}$	dictionary matrix with steering vectors as columns
$\mathbf{a}$	steering vector
$\mathbf{N}$	additive noise matrix
$\mathbf{n}$	additive noise vector
$\mathbf{X}$	source amplitude matrix
$\mathbf{x}$	source amplitude vector
L	number of snapshots
K	number of sources
λ	wavelength
f	frequency
$\Delta\phi$	phase difference
τ	lag
N	number of sensors in array
M	number of angles on predefined angular grid
$\mathbf{S_y}$	empirical sample covariance matrix
$\mathbf{S}_{\mathbf{y}}^{\text{true}}$	sample covariance matrix
$\mathbf{E_s}$	signal subspace
$\mathbf{E_n}$	noise subspace
θ_k	DOA at k^{th} direction/angle
θ^l	DOA at l^{th} snapshot
P_{d}	probability of detection

P_{ds}	probability of detection for static sources
P_{dm}	probability of detection for moving sources
TP	true positive
FN	false negative
TP_s	true positive for static source
TP_m	true positive for moving source
FN_s	false negative for static source
ν	fundamental frequency of sinusoidal signals
μ_l	measure
$\mathcal{M}$	set of complex measure
FN_m	false negative for moving source
$\boldsymbol{\Psi}$	continuous trajectory DOA space
$\mathcal{W}$	set of trajectory parameters for multiple sources
$\boldsymbol{\omega}_k$	vector of parameters defining k^{th} source trajectory
$\tilde{\boldsymbol{X}}_k$	a diagonal matrix of k^{th} source amplitude across snapshots

List of Related Publications

[P1] Ruchi Pandey, Santosh Nannuru, and Aditya Siripuram, **"Sparse Bayesian learning for acoustic source localization"**, in proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[P2] Ruchi Pandey and Santosh Nannuru, **"Parametric Models for DOA Trajectory Localization"**, in proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[P3] Ruchi Pandey, Santosh Nannuru, and Peter Gerstoft, **"Experimental Validation of Wideband SBL Models for DOA Estimation"**, in proceedings of *European Signal Processing Conference (EUSIPCO)*, 2022.

[P4] Ruchi Pandey, Shreyas Jaiswal, Huy Phan, and Santosh Nannuru, **"Improving trajectory localization accuracy via DOA estimation"**, in proceedings of *European Signal Processing Conference (EUSIPCO)*, 2023.

[P5] Ruchi Pandey and Santosh Nannuru, **"Grid-free algorithms for direction-of-arrival trajectory localization"**, in *The Journal of the Acoustical Society of America (JASA)*, 2024.

Related co-author publications:

[P6] Shreyas Jaiswal, Ruchi Pandey, and Santosh Nannuru, **"Deep Architecture for DOA Trajectory Localization"**, in proceedings of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Chapter 1

Introduction

Acoustic source localization involves determining the position of a sound source in the presence of noise, which is further affected by factors like reflections, reverberation, and background interference. The ability to concentrate on desired sounds in noisy environments is a challenging task and finds broad applications in areas such as smart devices [1], advanced driver assistance for detecting other vehicles [2], smart home to interact with the speaker [3], video conferencing for better meeting experience [4], drones to recognize activity in their environment [5], and hearing aid to improve the focus on the desired sound source [6]. Microphone arrays are commonly used in audio-based devices to capture sound signals from multiple directions.

Acoustic source localization refers to the process of determining the exact spatial coordinates of a sound source. In this thesis, we consider localization as direction-finding problem where the aim is to find the direction of arrival (DOA). In DOA estimation problem, the angle or direction from which a sound wave arrives at the sensor array is estimated. Localization provides estimates of source locations at specific time instants. Along with localization, it is typically desired that the sources are tracked over time. This is important because sources move around spatially; connecting the localization results over time provides a complete source trajectory. Despite being an extensively studied problem in array signal processing, acoustic source localization remains a critical and active area of research due to its diverse applications.

The existing localization algorithms encounter multiple challenges that limit their effectiveness and practicality. These challenges include reliance on wideband models that assume a flat spectrum, and failing to capture the rich frequency spectrum in natural audio signals. It is crucial to incorporate coherent modelling and processing across multiple frequencies to overcome this limitation. Additionally, computational efficiency is a key concern, demanding fast and hardware-friendly algorithms for real-time applications. Adapting to non-stationary targets and accommodating changes in data is essential for robust localization. Moreover, algorithms must be robust against noise and reverberations to operate accurately in noisy environments. High-resolution localization is vital when sources are in close proximity, necessitating precise localization of each source. Lastly, accurately distinguishing between correlated sources poses a challenge that needs to be addressed. Overcoming these challenges is crucial for developing advanced localization algorithms and enhancing accuracy, efficiency, and reliability in

various practical scenarios.

This chapter presents a general overview along with the background and motivation of the research carried out in this thesis. It is followed by presenting the research scope, problems, and objectives of this thesis. Towards the end, the contribution of the proposed research to the field and the author's contribution to the mentioned papers are summarised.

1.1 Objective and scope of the thesis

Accurate localization and tracking of sound sources in dynamic and noisy environments are crucial with the increasing prevalence of smart devices equipped with microphone arrays. This thesis aims to investigate and analyze various localization algorithms on real-world datasets to understand their performance comprehensively. As audio signals have a rich frequency spectrum, the thesis aims to enhance existing localization algorithms by exploring many wideband models and algorithms and their applicability in real-world scenarios. Additionally, the research focuses on deep neural network (DNN) based polyphonic sound event localization and detection (SELD) problems [7], which combine detection and localization tasks and have practical applications. By incorporating DOA and DOA derivatives, the proposed model seems to improve the temporal continuity of DOA estimates and suppress sudden changes in DOA.

Furthermore, the thesis tackles two significant challenges in sound source localization and tracking. Firstly, existing methods assume that the DOA of sound sources remains constant within short intervals, limiting their effectiveness in capturing fast-moving sources. Secondly, grid-based approaches commonly used in traditional methods suffer from reduced accuracy when resolving non-grid-aligned source parameters. To overcome these limitations, this research focuses on developing novel parametric models and algorithms for joint localization and tracking tasks, eliminating the explicit need for tracking filters to obtain smoother trajectories. The signal models and algorithms will be developed to accurately estimate complex trajectories of multiple moving sources simultaneously. The research focuses on general trajectory models capable of capturing linear and nonlinear motion. We introduced gridless algorithms for DOA trajectory estimation to overcome the limitations of grid-based algorithms. The performance evaluation of the proposed algorithms is carried out to demonstrate that by directly estimating source trajectories, the algorithms will offer higher resolution, improved noise robustness, and faster processing. The research will ultimately contribute to achieving more accurate source trajectories without the explicit need for tracking filters, thereby advancing the sound source localization and tracking field. In this thesis, we do not specifically address issues such as reverberation, reflections, estimating source numbers, or considering varying source numbers.

The developed signal models and algorithms will not be limited to sound waves but can be extended to other data types, such as radio waves. This versatility will expand the impact of the research across various applications, including radar-based tracking, autonomous vehicles, and robot navigation [8–11].

Ultimately, this research has the potential to contribute to a safer and more interconnected world where smart devices can effectively perceive and track sources, leading to improved performance in numerous applications.

1.2 Contributions

The following are the main technical contributions of this thesis:

- This thesis investigates and analyzes the performance of various existing localization algorithms such as conventional beamforming (CBF), multiple signal classification (MUSIC), and sparse Bayesian learning (SBL) using the real-world recordings from localization and tracking (LOCATA) dataset [12]. We demonstrate the effectiveness of SBL as a promising method for DOA estimation, which addresses real-world challenges included in LOCATA. The performance analysis shows that the compressive sensing (CS) algorithm of SBL outperforms CBF and MUSIC in all the considered tasks. The work establishes SBL as a robust approach for DOA estimation in challenging scenarios.

- Building upon the findings of the above analysis, the thesis focuses on wideband DOA estimation using SBL algorithms. We address the limitations of existing wideband SBL algorithms by proposing a realistic signal model that considers the change in source variance across the frequency range. Three wideband SBL variants (SBL1, SBL2, and SBL3) are applied and evaluated, along with wideband versions of CBF and MUSIC. Through simulations and experiments using the LOCATA dataset, we demonstrate that SBL3, which incorporates a shared colored spectrum, performs best across different signal models and array configurations. This work presents an improved understanding of wideband SBL algorithms and their applicability in real-world scenarios. The findings reveal the effectiveness of an intermediate model that allows the spectrum to vary with the frequency band, which accurately balances sparsity and power spectrum for real-world signals.

- Next, we explore Deep Neural Network (DNN) based methods to show the significance of predicting DOA derivatives alongside DOA for enhancing localization performance. We propose a new model, which combines DOA and their derivatives, and compare it with the existing which predicts only DOA. Our experiments using the TAU-NIGENS Spatial Sound Events 2021 dataset highlight the improvement achieved by considering both DOA and derivatives, especially under low signal-to-noise ratio (SNR) conditions. This study emphasizes the importance of incorporating higher-order derivatives in sound event localization and detection tasks, opening avenues for future research.

- We introduce novel parametric signal models: polynomial and bandlimited, to identify DOA trajectories, which capture the dynamic motion of a source within a block. Instead of estimating

DOA, the work focuses on estimating trajectory parameters. The grid-based TL-CBF, TL-OMP, and TL-SBL algorithms are developed to estimate DOA trajectories. Grid-based methods face challenges in resolving non-grid-aligned source parameters, leading to reduced localization accuracy. Gridless algorithms overcome these limitations by estimating parameters in continuous trajectory space, improving real-world performance. We propose two gridless algorithms: i) Sliding Frank-Wolfe (SFW), which solves the Beurling LASSO problem, and ii) Newtonized Orthogonal Matching Pursuit (NOMP), which improves over OMP using cyclic refinement. Furthermore, we extend our analysis to include wideband processing. The results present the impact of SNR, number of snapshots, resolution limits, grid step size, and computational complexity. The study highlights the potential of parametric trajectories to eliminate the need for tracking filters and improve both localization and tracking performance.

- We apply the proposed algorithms to recordings from LOCATA. These recordings present challenging scenarios, such as near-field sources, ambient noise from nearby roads, and multiple moving sources and arrays. We are particularly interested in observing how the proposed TL algorithms in this thesis can be applied to showcase their superior performance in such scenarios.

Furthermore, in conjunction with the aforementioned contributions, the author has also participated in developing data-driven methods for trajectory localization tasks. Specifically, the utilization of the U-Net architecture has facilitated the estimation of linear trajectory parameters [13].

1.3 Structure of the thesis

The thesis is structured into the following chapters. Chapter 2 offers an overview of existing literature on DOA estimation, including traditional methods, CS methods, DNN based methods, and gridless localization methods are discussed.

Chapter 3 focuses on DOA estimation algorithms applied to real-world datasets, specifically LOCATA and detection and classification of acoustic scenes and events (DCASE). Experimental validation of CBF, MUSIC and SBL on LOCATA is conducted, followed by performance analysis of wideband SBL algorithms using different assumptions about the source spectrum. Additionally, we analyze the DNN-based architecture for SELD tasks using the DCASE dataset.

In Chapter 4, we introduce a novel concept of trajectory localization for joint localization and tracking tasks. We propose several grid-based and gridless trajectory localization (TL) algorithms, where trajectories are estimated for block array data instead of assuming a constant DOA. The analysis is further extended to include multi-frequency signals. We also apply the proposed algorithms to recordings from LOCATA.

Finally, Chapter 5 concludes the thesis by summarizing the research findings, providing a final perspective on the study, and laying the groundwork for future studies.

Chapter 2

Background and literature review

This chapter discusses the mathematical representation of DOA estimation problem and explains various localization algorithms. The discussion starts with traditional methods such as CBF, MUSIC, generalized cross-correlation (GCC), generalized cross-correlation with phase transform (GCC-PHAT), and steered response power with phase transform (SRP-PHAT) [14–16]. Next, the DOA estimation problem is formulated as a sparse recovery problem, and CS based methods like basis pursuit (BP), orthogonal matching pursuit (OMP), and SBL are explained [17, 18]. The chapter also delves into deep learning-based approaches, which can be broadly categorized into classification and regression tasks [19]. Finally, the chapter concludes by discussing gridless localization methods, including atomic norm minimization (ANM) based methods, Newtonized orthogonal matching pursuit (NOMP), and sliding Frank-Wolfe (SFW) [20–22].

2.1 Direction of arrival estimation

In array signal processing, source localization aims to find the directional information of the source of interest with respect to receiver array [14]. The source localization is often considered a parameter estimation problem, where the key parameter of interest is the DOA [23].

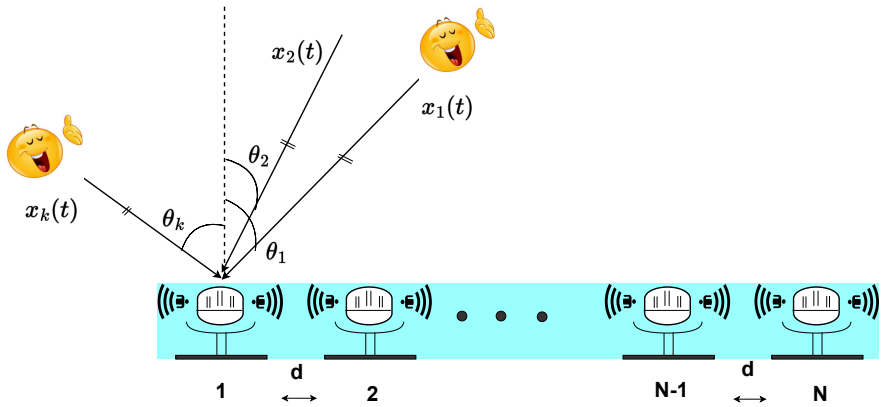

Figure 2.1 DOA estimation using N-element uniform linear array

Fig. 2.1 illustrates an N-element uniform linear array (ULA), where d represents the inter-sensor spacing between two consecutive sensors. Under the far-field assumption (when the distance between the source and array is much greater than the inter-sensor spacing), the incoming wavefronts from sources can be considered as plane waves impinging on the array [15]. The direction from which the propagating wave impinges on the array is called the DOA. As the incident wavefront reaches each sensor with different time delays, these delays correspond to phase differences in the frequency domain. Properly arranging these time delays or phase differences can accurately reconstruct the signal received from a source (the array output). The array response or steering vector represents the relative phase differences between each element in the sensor array, and for ULA, it is expressed as

$$\mathbf{a}(\theta_k) = \begin{bmatrix} 1 \\ e^{j\ 2\pi \frac{d}{\lambda} \sin(\theta_k)} \\ e^{j\ 2\pi \frac{2d}{\lambda} \sin(\theta_k)} \\ \vdots \\ e^{j\ 2\pi \frac{(N-1)d}{\lambda} \sin(\theta_k)} \end{bmatrix}. \tag{2.1}$$

Here '1' represents the reference signal (received from sensor 1). The phase difference between two consecutive sensors is $\Delta\phi = 2\pi \frac{d}{\lambda} \sin(\theta_k)$, where λ is the wavelength and θ_k is the angle made by the direction of propagation corresponding to k^{th} source with the normal to the linear array (see Fig. 2.1). Note that (2.1) represents the steering vector for the ULA. The expression of the steering vector differs across array structures (rectangular, circular, and spherical arrays) and also varies for 2D and 3D localization.

Let there be K sources with complex amplitudes $(x_1, x_2, \ldots, x_k)$ impinging on a ULA with N sensors from different directions $(\theta_1, \theta_2, \ldots, \theta_k)$ as shown Fig. 2.1. The frequency domain representation of the narrowband single measurement vector (SMV) can be given as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{a}(\theta_k)\, \mathbf{x}_k + \mathbf{n}_k = \mathbf{A}_{\text{sv}}\, \mathbf{x} + \mathbf{n} \tag{2.2}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ e^{j\,2\pi \frac{d}{\lambda} \sin(\theta_1)} & e^{j\,2\pi \frac{d}{\lambda} \sin(\theta_2)} & \ldots & e^{j\,2\pi \frac{d}{\lambda} \sin(\theta_K)} \\ \vdots & \vdots & \ldots & \vdots \\ e^{j\,2\pi \frac{(N-2)d}{\lambda} \sin(\theta_1)} & e^{j\,2\pi \frac{(N-2)d}{\lambda} \sin(\theta_2)} & \ldots & e^{j\,2\pi \frac{(N-2)d}{\lambda} \sin(\theta_K)} \\ e^{j\,2\pi \frac{(N-1)d}{\lambda} \sin(\theta_1)} & e^{j\,2\pi \frac{(N-1)d}{\lambda} \sin(\theta_2)} & \ldots & e^{j\,2\pi \frac{(N-1)d}{\lambda} \sin(\theta_K)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{K-1} \\ x_K \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{N-1} \\ n_N \end{bmatrix}$$

Here $\mathbf{y} \in \mathbb{C}^N$ is the received measurement vector, $\mathbf{A}_{\text{sv}} \in \mathbb{C}^{N \times K}$ is the steering matrix whose columns are the steering vectors corresponding to the K different angles (to be estimated). The $\mathbf{x} \in \mathbb{C}^K$ denotes the complex amplitude of source signals, and $\mathbf{n} \in \mathbb{C}^N$ accounts for the additive noise. The

complex source amplitudes $\mathbf{x}$ and noise $\mathbf{n}$ are modeled as random Gaussian and assumed to be independent. It can be seen from (2.1) and (2.2) that the equations are nonlinear with respect to the direction of arrivals, and the model is a narrowband model (single frequency). In the DOA estimation problem, we are interested in solving the equations in (2.2) and finding the angles using the measurements. Let us consider the signals are received at different timestamps, then a frequency domain representation of multiple measurement vector (MMV) model is given as

$$\mathbf{Y} = \mathbf{A}_{\text{sv}}\,\mathbf{X} + \mathbf{N} \tag{2.3}$$

Here $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2 .... \mathbf{y}_L]$ and $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2 .... \mathbf{n}_L]$ are measurement and noise matrix with $N \times L$ dimension. Here L denotes the total number of snapshots. $\mathbf{A}_{\text{sv}}$ is steering matrix $(N \times K)$ whose columns are steering vectors corresponding to K source angles. The source signals $\mathbf{X}$ have dimensions $K \times L$, where l^{th} column represents the source amplitude from all K directions for l^{th} snapshot. The MMV helps to get better localization accuracy. Figure 2.2 illustrates a block diagram that provides a comprehensive overview of these methods, which are broadly classified into four groups.

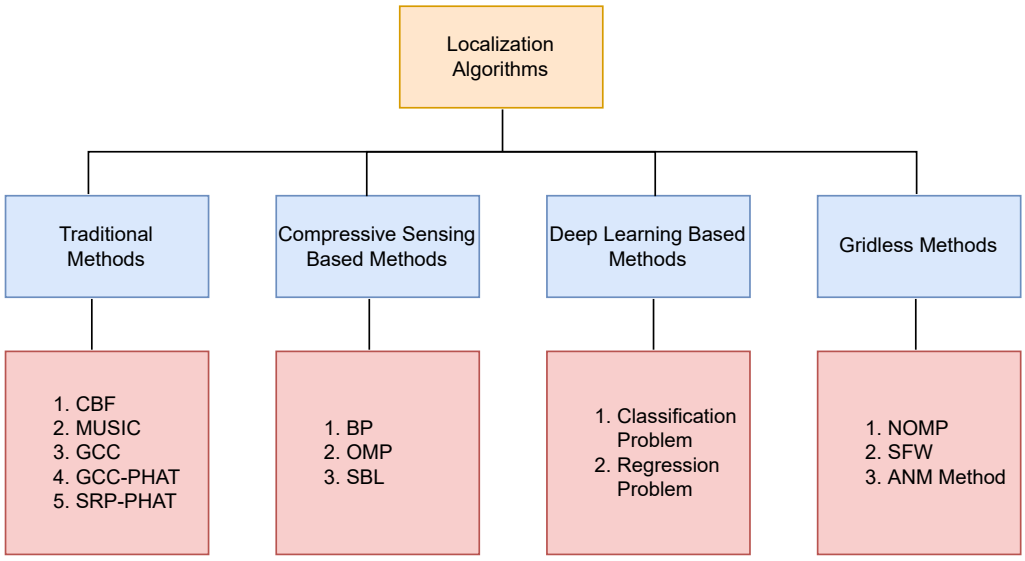

Figure 2.2 Localization algorithms broadly divided into four categories

2.2 Traditional methods

The discussion begins with the straightforward method of beamforming, followed by the high-resolution subspace-based approach called MUSIC. In the context of time-domain localization, the discussion covers the GCC, GCC-PHAT and, subsequently, the SRP-PHAT algorithms.

2.2.1 Conventional Beamforming (CBF)

CBF or spatial filtering is the simplest DOA estimation method [24]. This technique involves computing the power spectrum across various angles on a predefined angular grid and then finding the peaks from the spectrum to localize the source by amplifying the output in a specific direction while attenuating signals from all other directions [14]. Spatial weights are assigned to steering vectors during the angular power spectrum computation. In other words, the received signals are combined coherently when the steering angle aligns with the true DOA, resulting in a correlation peak. As depicted in Figure 2.3, the computed angular power spectrum exhibits multiple peaks corresponding to each source.

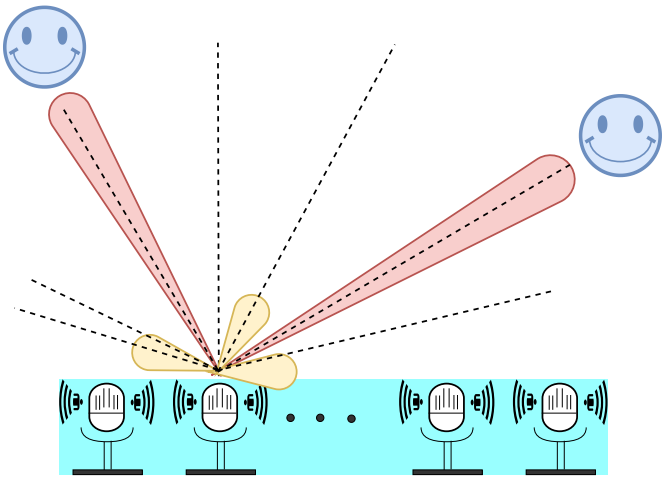

Figure 2.3 Conventional beamforming

Let M be the total number of potential DOAs on a predefined discrete angular grid $\theta \in [-90, 90]^\circ$. The steering vectors are computed for all M candidate angles. The source signal $\mathbf{X}$ has dimensions $M \times L$, where l^{th} column represents the source amplitude from all M directions for l^{th} snapshot. For l^{th} snapshot, the source amplitude in θ direction can be obtained by computing the correlation between observation and all the steering vectors corresponding to M angles. The CBF power spectrum can be computed using the below equation

$$
\begin{aligned}
\mathbf{P}_{\text{CBF}}(\theta) &= \frac{1}{L}\sum_{i=1}^{L} |\mathbf{a}^H(\theta)\, \mathbf{y}_i|^2 \\
&= \frac{1}{L}\sum_{i=1}^{L} \mathbf{a}^H(\theta)\, \mathbf{y}_i \mathbf{y}_i^H\, \mathbf{a}(\theta) \\
&= \mathbf{a}^H(\theta)\, \mathbf{S}_{\mathbf{y}}\, \mathbf{a}(\theta) \\
\text{where } \mathbf{S}_{\mathbf{y}} &= \frac{1}{L}\sum_{l=1}^{L} \mathbf{y}_l\, \mathbf{y}_l^H.
\end{aligned}
\tag{2.4}
$$

Here $\mathbf{a}(\theta)$ is the array steering vector (2.1) corresponding to a source located at θ angle. $\mathbf{S_y}$ is the empirical data covariance matrix computed using L snapshots. The true data covariance is given as $\mathbf{S}_{\mathbf{y}}^{\text{true}} = E\{\mathbf{y}\mathbf{y}^H\}$. Incorporating both azimuth (horizontal angle) and elevation (vertical angle) angles into the steering vector allows for more precise localization in two-dimensional space. Although CBF is computationally efficient and robust to noise, it suffers from low resolution and many sidelobes; hence, it is difficult to localize two nearby sources [17, 24].

2.2.2 Multiple Signal Classification (MUSIC)

MUSIC is a high-resolution subspace-based method for DOA estimation [25]. In this approach, the estimated sample covariance matrix (**Sy**) is decomposed into two orthogonal subspaces: the signal subspace and the noise subspace. While the specific signal components remain unknown, the sum of the signal and noise components is given as $\mathbf{S}_{\mathbf{y}}^{\text{true}}$.

$$\begin{aligned}\mathbf{S}_{\mathbf{y}}^{\text{true}} &= E\{\mathbf{y}\mathbf{y}^H\} \\ &= E\{\mathbf{A}_{\text{sv}}\,\mathbf{X}\mathbf{X}^H\,\mathbf{A}_{\text{sv}}^H\} + E\{\mathbf{N}\mathbf{N}^H\} \\ &= \mathbf{A}_{\text{sv}}\mathbf{S}_{\text{x}}\mathbf{A}_{\text{sv}}^H + \sigma^2\,\mathbf{I}. \end{aligned} \tag{2.5}$$

Here $\mathbf{S}_{\text{x}}$ is $(K \times K)$ signal covariance matrix, K is the number of source signals. The eigenvalue decomposition is applied to the empirical sample covariance matrix ($\mathbf{S}_{\text{y}}$) computed using (2.4). The obtained eigenvectors are sorted in the descending order of the eigenvalues. The signal subspace $\mathbf{E_s}$ is constructed by selecting K eigenvectors, and the remaining $(N - K)$ eigenvectors correspond to the noise subspace $\mathbf{E_n}$. The pseudo power spectrum for MUSIC is given by

$$\mathbf{P}_{\text{mu}}(\theta) = \frac{1}{\mathbf{a}^H(\theta)\mathbf{E_n}\mathbf{E_n}^H\mathbf{a}(\theta)}. \tag{2.6}$$

MUSIC uses the orthogonality between the signal and the noise subspaces to locate the maxima in the spectrum [17]. The $\mathbf{a}^H(\theta)$ is orthogonal with the columns of $\mathbf{E_n}$, the value of the denominator is zero (or close to zero when noise is present), and $\mathbf{P}_{\text{mu}}(\theta)$ shows a peak corresponding to source DOA. For localizing multiple sources, the K such peaks will be selected. Although MUSIC is a high-resolution method, its performance gets compromised with fewer snapshots. The covariance matrix must be formed from sufficient snapshots to ensure accurate eigenvalue decomposition, separating the eigenvectors into distinct signal and noise subspaces. Additionally, the number of signal sources must be less than the number of sensors to ensure that the covariance matrix has enough dimensions to distinguish between signal and noise subspaces.

2.2.3 Generalized Cross-Correlation (GCC)

The GCC method is a time domain localization method that computes delay or lags between two received signals from the sensor pair. Let $\mathbf{y}_l(t)$ and $\mathbf{y}_k(t)$ are time domain received signals from sensor

k and l. The $\mathbf{Y}_k(\omega)$ and $\mathbf{Y}_l(\omega)$ are the Fourier transform of $\mathbf{y}_l(t)$ and $\mathbf{y}_k(t)$. The time difference of arrival (TDOA) estimate is the delay or lags $\hat{\tau}_{kl}$ that maximizes the cross-correlation $R_{kl}(\tau)$ between k^{th} and l^{th} sensor and can be expressed as

$$\hat{\tau}_{kl} = \arg\max_{\tau} R_{kl}(\tau) \tag{2.7}$$

$$R_{kl}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_k(\omega)\,\mathbf{Y}_k(\omega)\,G_l'(\omega)\,\mathbf{Y}_l'(\omega)\,e^{j\omega\tau}\,d\omega. \tag{2.8}$$

Here $G_k(\omega)$ and $G_l(\omega)$ are prefilters designed to filter the received sensor signals, aiming to mitigate the impact of channel interference and noise. The signal $\mathbf{Y}_l'(\omega)$ represents the complex conjugate of $\mathbf{Y}_l(\omega)$. Note that when the prefilters $G_k(\omega)$ and $G_l(\omega)$ are set to unity (i.e., no filtering is applied), the expression in (2.8) is equivalent to the standard cross-correlation function. The computed delay $\hat{\tau}_{\text{kl}}$ can be combined with the parameters of known array geometry using the relation $\hat{\tau}_{kl} = \frac{d}{c}\sin(\theta)$. A matrix $\mathbf{D}$ can be constructed to compute the delays using multiple sensors, whose columns are the inter-sensor spacing for each pair of microphones. For N number of sensors, the total possible sensor pairs are given by $N_t = {}^{N}C_2$. The DOA vector $\mathbf{b}(\theta)$ for 1D localization can be obtained using these M computed delays and their relationship can be given as

$$\boldsymbol{\tau} = \frac{1}{c}\mathbf{D}^T\mathbf{b}(\theta), \tag{2.9}$$

where $\mathbf{b}(\theta)$ is a unit vector representing DOA and $\boldsymbol{\tau}$ is the vector containing M delays. The least-square estimate of $\mathbf{b}(\theta)$ is given by

$$\hat{\mathbf{b}}_{\text{LS}}(\theta) = [(\mathbf{D}\mathbf{D}^T)^{-1}\,\mathbf{D}]\,c\,\hat{\boldsymbol{\tau}}(\theta). \tag{2.10}$$

The performance of GCC degrades and often produces inaccurate DOA estimates in low SNR conditions. Also, the GCC method suffers from low resolution and is not well-suited to handle multi-frequency signals. Hence, for better performance in reverberant scenarios, different weighting functions can be used, such as maximum likelihood (ML) weighting function, phase transform (PHAT), Roth processor, smoothed coherence transform (SCOT), and bandpass weighting function [26].

2.2.4 Generalized Cross-Correlation with Phase Transform (GCC-PHAT)

To improve the robustness of GCC function, a phase transform (PHAT) weighting function is applied in (2.8). It whitens the microphone signals by normalizing the cross-spectral density using the magnitude of spectrum [26]. When the PHAT weighing function is used as prefilters in (2.8), the method is called GCC-PHAT. The GCC-PHAT is more robust to noise and amplitude differences than the GCC method. The GCC-PHAT method is computed as

$$R_{kl}(\boldsymbol{\tau}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)|}\,\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)\,e^{j\omega\tau}\,d\omega, \tag{2.11}$$

here $G_k(\omega)G_l(\omega) = \frac{1}{|\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)|}$ is the PHAT weighting function. The PHAT function tends to enhance the effect of frequencies with low power compared to noise power. This can cause the estimates to be corrupted by the effect of uncorrelated noise [27]. In the case of multiple sources, the GCC-PHAT may give erroneous peaks in the presence of reflections. Additionally, each source will be assigned by the maximum lag, but the absolute maximum might not be assigned to the same source all the time, hence resulting in artificial source switching [27, 28]. Hence, estimating the multiple sources using GCC-PHAT often leads to incorrect estimates [28].

2.2.5 Steered Response Power with Phase Transform (SRP-PHAT)

To enhance the performance of GCC-PHAT, steered response power (SRP) is employed, offering improved accuracy, particularly in scenarios with multiple sources or complex wavefront geometries, making it well-suited for real-time applications. SRP operates on a concept akin to a beamformer, where the power of the array output is calculated at various angles, and the angle corresponding to the highest power is considered the true DOA. However, the performance of SRP may suffer in scenarios with strong reflections. Different weighting functions can be utilized to improve its performance in such cases [27, 28] to address this limitation. These techniques enable SRP to handle challenging acoustic environments and deliver more reliable DOA estimates. In SRP-PHAT, a phase transform weighing function is employed in conjunction with the SRP function to mitigate the effects of reflections and reverberation. For SRP-PHAT, a cumulative GCC-PHAT value is calculated across all microphone pairs at each delay (τ_q), which is associated with the candidate DOA (θ_q). The peak of the SRP-PHAT function indicates the location of the source. Let us consider $\mathbf{y}_k$ and $\mathbf{y}_l$ as the signals arriving from the direction θ_q and received by the k^{th} and l^{th} microphones, respectively. The time delay between $\mathbf{y}_k$ and $\mathbf{y}_l$ can be denoted as $\tau_{kl}(\theta_q)$. Then, the estimated GCC-PHAT value for $\mathbf{y}_k$ and $\mathbf{y}_l$, analogous to equation (2.11), can be expressed as

$$\hat{R}^{\text{PHAT}}_{\mathbf{y}_k\mathbf{y}_l}(\tau_{kl}(\theta_q)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)|}\,\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)\,e^{j\omega\tau_{kl}(\theta_q)}\,d\omega. \tag{2.12}$$

Now the SRP-PHAT can be expressed by following

$$P(\theta_q) = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \hat{R}^{PHAT}_{\mathbf{y}_k\mathbf{y}_l}(\tau_{kl}(\theta_q), \tag{2.13}$$

$$P(\theta_q) = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)|}\,\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)\,e^{j\omega\tau_{kl}(\theta_q)}\,d\omega, \tag{2.14}$$

$$P(\theta_q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \frac{1}{|\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)|}\,\mathbf{Y}_k(\omega)\,\mathbf{Y}_l'(\omega)\,e^{j\omega\tau_{kl}(\theta_q)}\,d\omega. \tag{2.15}$$

The N represents the number of sensors. The value of $P(\theta_q)$ can be computed for different angles from a discrete predefined angular grid. The direction, or true DOA, is determined by identifying the angle for which $P(\theta_q)$ has its maximum value. From Equation (2.13), it is evident that SRP-PHAT involves summing all possible combinations of GCC over sensor pairs. Consequently, when the number of points to scan becomes large, the computation complexity of SRP-PHAT increases significantly [16]. To mitigate this computational burden, SRP-PHAT with stochastic region contraction (SRC) is employed [29]. Compared to GCC-PHAT, SRP-PHAT exhibits superior performance, as it is robust to noise and offers improved localization of multiple sources with reduced ambiguity. This makes SRP-PHAT a preferred method for various applications.

2.3 Compressive Sensing (CS) based methods

CS or sparse signal processing is an emerging field that has garnered significant attention across various engineering disciplines, including computer science, information theory, electrical engineering, and applied mathematics [18]. The CS framework addresses sparse inversion problems by utilizing only a few noisy linear measurements while imposing sparsity constraints on unknown signals [30]. In this section, we provide a concise introduction to CS and discuss its application to the DOA estimation problem as a sparse recovery problem. We delve into popular CS methods, including BP, OMP, and SBL, highlighting their significance in addressing the challenges of sparse signal recovery.

2.3.1 CS model for DOA estimation

The DOA estimation problem can be formulated as a sparse recovery (CS) problem. The signal model given in (2.2) is reformulated as a sparse model, which linearly maps the compressible unknown signal $\mathbf{x}$ into the given measurements $\mathbf{y}$, is given as

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}. \tag{2.16}$$

In this CS model (2.16), $\mathbf{y} \in \mathbb{R}^N$ is the measurements received from N sensors, $\mathbf{x} \in \mathbb{R}^M$ represents the source amplitudes in M directions/angles from a predefined angular grid, and $\mathbf{A}$ is a $N \times M$ dictionary matrix whose columns are steering vectors corresponding to each candidate direction/angle from the same predefined DOA grid. For example, let us consider a predefined DOA grid over $[-90, 90]^\circ$ with 1° resolution, then there is $M = 181$ candidate DOA in 1D localization (only azimuth). The sparsity assumption is based on the fact that only a few sources are present among all the candidate angles. Let K be the number of unknown sources ($K < N$), and $M \gg K$. Thus, in the given model, $\mathbf{x}$ is an M-length vector with K-sparsity, and CS methods are employed to estimate the K non-zero values corresponding to K directions or angles.

The dictionary $\mathbf{A}$ maps signal of interest $\mathbf{x}$ into the linear measurements $\mathbf{y}$. Note that unlike system (2.2), above system (2.16) is an undetermined system as $M > N$, i.e., the number of linear equations is

less than the number of unknowns ($\mathbf{x}$). The given system is underdetermined and has an infinite number of solutions. The dictionary $\mathbf{A}$ is assumed to be known and fixed. The CS techniques are used to find the sparsest solution for unknown $\mathbf{x}$ by imposing sparsity on it. The $\mathbf{x}$ is assumed to be K sparse, i.e. only K nonzero values are present in $\mathbf{x}$ and $M \gg K$. In literature, it has been shown that under the sufficient sparsity assumption of underlying signal and the incoherence mapping of the underlying signal into the measurements, CS methods can solve the given underdetermined system [18, 31–33]. The principal assumption of CS lies in the sparsity of the underlying signal ($\mathbf{x}$) and its mapping into fewer measurements using a sensing matrix ($\mathbf{A}$).

2.3.2 Basis Pursuit (BP)

The given model in (2.16) is an underdetermined system, and one common approach would be the least square solution, but the solution will be non-sparse. There are many different ways to find the sparse solution for vector $\mathbf{x}$ such as l_0 and l_1 minimization techniques. Let the l_p norm of a vector $\mathbf{x} \in \mathbb{R}^{\mathrm{m}}$ is defined as

$$||\mathbf{x}||_p = (\sum_{i=1}^{m} |x_i^p|)^{\frac{1}{p}}. \tag{2.17}$$

In equation (2.17), for $p = 0, 1$ and 2, the l_0, l_1, and l_2 norm can be formulated. It can be seen that when $p = 0$, it is known as the l_0 norm (which is not a norm). The l_0 norms count all the nonzero values in the signal. The solution that minimizes the l_0 norm gives the sparse solution.

The l_0 norm minimization gives simply a sparse solution by counting the non-zero entries of vector $\mathbf{x}$, which contains the significant information of the vector $\mathbf{x}$. The problem of l_0 norm minimization can be formulated as given in the below equation

$$\min_{\mathbf{x} \in \mathbb{R}^m} ||\mathbf{x}||_0 \text{ subject to } \mathbf{y = A\,x}. \tag{2.18}$$

The above equation leads to a non-convex optimization problem and is NP hard to solve it. The proof of the above concept can be found in theorem 2.17 in [18]. As the above l_0 minimization is intractable, to solve it further and get the sparse solution CS methods of BP and least absolute shrinkage and selection operator (LASSO) are explained in next subsection. The condition on sparsity and measurements to reconstruct k sparse vector from the observations is well explained in chapter 2 of the excellent book [18].

The l_0 minimization problem is NP-hard and intractable, however, under the assumptions of sufficient sparsity and incoherent columns of $\mathbf{A}$, it can be approximated to l_1 minimization problem [17, 18, 33–35]. The BP provides the solution of $\mathbf{x}$, whose coefficient has the minimum l_1 norm [36].

$$\min_{\mathbf{x}} ||\,\mathbf{x}\,||_1 \text{ subject to } \mathbf{y} = \mathbf{A}\,\mathbf{x}. \tag{2.19}$$

where $||\mathbf{x}||_1 = \sum_{i=1}^{M} |x_i|$ represents the l_1 norm of $\mathbf{x}$. The above problem is a convex optimization problem and can be solved using linear programming methods [35,36]. The BP solution is tractable and can be solved efficiently even for larger dimensions. Due to the convexity, BP converges to the global minima. If the noise $\mathbf{n}$ is included as in model (2.16), the BP formulation can be modified as

$$\min_{\mathbf{x}} ||\,\mathbf{x}\,||_1 \text{ subject to } ||\,\mathbf{A}\,\mathbf{x}\,\text{-}\,\mathbf{y}\,||_2 \leq \eta. \tag{2.20}$$

Here η is the upper bound of the noise norm such that $||\mathbf{n}||_2 \leq \eta$ and $||\mathbf{n}||_2 = (\sum_{i=1}^{m} |n_i^2|)^{\frac{1}{2}}$ represents the l_2 norm of $\mathbf{n}$. It indicates the degree to which noise can be accommodated to yield a sparse solution for the vector $\mathbf{x}$. The mathematical insights about the approximation of l_0 minimization into l_1 minimization can be understood using concepts of nullspace property (NSP) and restricted isometric property (RIP) (refer chapters 4 and 6 of [18]). A regularization parameter λ is introduced in (2.20) to get the balance between the sparsity and the noise tolerance. The unconstrained formulation for (2.20) can be written as

$$\min_{\mathbf{x}} ||\,\mathbf{A}\,\mathbf{x}\,\text{-}\,\mathbf{y}\,||_2^2 + \lambda||\,\mathbf{x}\,||_1 \tag{2.21}$$

The above formulation is known as the LASSO method to solve the system model (2.16). The regularization term λ acts as a weighting parameter between the sparsity and the noise tolerance. The high value of λ leads to a more sparse solution, whereas the low value of λ provides solutions that closely match the observed data. The LASSO problem can be solved using linear or quadratic programming methods using simplex or interior point algorithms [36].

In the DOA estimation problem, the solution of $\mathbf{x}$ can be obtained by solving l_1 minimization problem, given the observation ($\mathbf{y}$) and sensing matrix ($\mathbf{A}$). Further, if the noise level is unknown, then the choice of η is of significant importance and, if chosen incorrectly, can lead to erroneous DOAs. In (2.21), if noise is overestimated (λ is set to be high), it gives a very sparse solution (a few DOAs might be missed). Also, if noise is underestimated (lower value of λ), the sparsity will get compromised (many false DOAs). In [17], the DOA estimation is addressed using the CS-based l_1 minimization technique.

2.3.3 Orthogonal matching pursuit (OMP)

In 2007, Tropp and Gilbert introduced an Orthogonal Matching Pursuit (OMP) method, which expanded upon earlier developments in sparse signal recovery, particularly the matching pursuit technique. OMP is a greedy algorithm and computationally efficient and faster compared to other CS methods [37]. The central idea of OMP is to iteratively identify the basis vector or column of the matrix $\mathbf{A}$ that yields the maximum projection of the measurement vector $\mathbf{y}$. At each iteration, the algorithm finds the index of the column most correlated with the observation, removes that particular column from $\mathbf{A}$, and continues to extract the next significantly correlated columns from the remaining residuals. In the end, K columns are selected, containing information about the $\mathbf{x}$ signal of interest. The residual is guaranteed to be orthogonal to all the columns of $\mathbf{A}$.

In [38], OMP method is used to address the DOA estimation problem. OMP offers the advantage of low computational complexity and can work effectively with a single snapshot. The details about this greedy algorithm can be referred to from [37]. In [38], OMP is applied to address DOA estimation, and its performance is compared against the MUSIC and MVDR methods. The OMP algorithm is summarized below.

Algorithm 1 OMP algorithm

1. $\mathbf{r}_0 = \mathbf{y},\ \mathbf{A}_0 = \emptyset, \Lambda_0 = \emptyset$ and an iteration counter $c = 1$
2. Initialization: $\mathbf{r}_0 = \mathbf{y},\ \mathbf{A}_0 = \emptyset, \Lambda_0 = \emptyset$ and an iteration counter $c = 1$
3. Find the corresponding index λ_c of the optimization problem
   $\lambda_c = \arg\max |\mathbf{A}^H \mathbf{r}_{c-1}|$
4. Augment the index set $\Lambda_c = \Lambda_{c-1} \cup \{\lambda_c\}$ and the matrix of chosen atoms $\mathbf{A}_c = \mathbf{A}[:, \Lambda_c]$
5. Solve the following optimization problem to obtain the signal vector estimate for $\mathbf{A}_c$
   $\mathbf{x}_c = \arg\min_x ||\mathbf{A}_c \mathbf{x} - \mathbf{y}||_2$
6. Calculate the new approximation ($\boldsymbol{\beta}_c$) of $\mathbf{y}$ and the new residual:
   $\boldsymbol{\beta}_c = \mathbf{A}_c\, \mathbf{x}_c$
   $\mathbf{r}_c = \mathbf{y} - \boldsymbol{\beta}_c$
7. Increase c by 1, and return to Step 2) if $c < K$.

2.3.4 Sparse Bayesian learning (SBL)

Let us consider the signal model as (2.16) to derive the SBL update equations for estimating the hyperparameters. The optimization problem given in (2.21) and other l_p minimization techniques can also be derived using the Bayesian framework [39,40]. Let us consider model (2.16) with Gaussian noise assumption, then the underlying optimization problem in (2.19) can be solved using the maximum-a posteriori (MAP) estimator given as

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathrm{p}(\mathbf{x}|\mathbf{y}) \tag{2.22}$$

$$= \arg\max_{\mathbf{x}} \mathrm{p}(\mathbf{y}|\mathbf{x})\, \mathrm{p}(\mathbf{x}) \tag{2.23}$$

$$= \arg\min_{\mathbf{x}} \ -\log \mathrm{p}(\mathbf{y}|\mathbf{x}) - \log \mathrm{p}(\mathbf{x}) \tag{2.24}$$

$$= \arg\min_{\mathbf{x}} \, ||\, \mathbf{y} - \mathbf{A}\, \mathbf{x}\, ||_2^2 + \lambda \sum_{i=1}^{M} g(|x_i|), \tag{2.25}$$

where g is a strictly concave function and often leads to the sparse solution with maximum N nonzero values. The different choices of $g(.)$ lead to different levels of sparsity. For the LASSO framework, a Laplacian prior $\mathrm{p}(\mathbf{x}) = \frac{a}{2} e^{-a|\mathbf{x}|}$ and $\mathrm{p}(\mathbf{y}|\mathbf{x})$ to be a Gaussian likelihood leads to the sparse solution for $\mathbf{x}$. Also, if both the prior and likelihood are assumed to be Gaussian, it leads to the l_2 norm regularized problem. In a similar manner, different l_1 minimization techniques can be formulated with the correct choice of prior.

Let the $g(x_i) = |\mathbf{x}_i|^p$ with $p \leq 1$, then the MAP estimate can be given as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} || \mathbf{y} - \mathbf{A}\,\mathbf{x} ||_2^2 + \lambda \sum_{i=1}^{M} (|x_i|^p). \tag{2.26}$$

The drawback of the MAP estimator is that if the prior is very sparse, i.e., $p \sim 0$, many local minima will be obtained, converging to sub-optimal local minima leads to convergence error. Also, if the prior is not sparse enough, i.e., $p \sim 1$, the obtained global minimum does not lead to the sparsest solution, resulting in the structural error. A hierarchical Bayesian approach is used to avoid these errors, where the prior is parameterized using hyperparameter γ. These hyperparameters are used to tie various parameters into fewer parameters, leading to fewer minima. Sparse Bayesian learning is a technique which estimates the information by computing the posterior distribution given by

$$\hat{\gamma} = \arg\max_{\gamma} \mathrm{p}(\gamma|\mathbf{y}) = \arg\max_{\gamma} \int \mathrm{p}(\mathbf{y}|\mathbf{x})\,\mathrm{p}(\mathbf{x}|\gamma)\,\mathrm{p}(\gamma)\,dx \tag{2.27}$$

In SBL, instead of solving a MAP problem, the hyperparameter γ is estimated and further used to get the posterior distribution of $\mathbf{x}$ given by $\mathrm{p}(\mathbf{x}|\mathbf{y};\hat{\gamma})$. The solution for the above equation with different sparse priors is not tractable, so some assumptions are needed to solve the problem. Let Gaussian scale mixture be used to represent different sparse priors over $\mathbf{x}$ with different scale mixing density $\mathrm{p}(\gamma_i)$ given by

$$\mathrm{p}(\mathbf{x}_i) = \int \mathrm{p}(\mathbf{x}_i|\gamma_i)\,\mathrm{p}(\gamma_i)\,d\gamma_i = \int \mathcal{N}(x_i; 0, \gamma_i)\,\mathrm{p}(\gamma_i)\,d\gamma_i. \tag{2.28}$$

The optimal value of hyperparameter γ can be estimated as

$$\hat{\gamma} = \arg\max_{\gamma} \mathrm{p}(\gamma|\mathbf{y}) = \arg\max_{\gamma} \mathrm{p}(\mathbf{y}|\gamma)\,\mathrm{p}(\gamma) \tag{2.29}$$

$$= \arg\min_{\gamma} \log|\boldsymbol{\Sigma}_{\mathbf{y}}| + \mathbf{y}^T\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} - 2\sum_{i=1}^{M} \log \mathrm{p}(\gamma_i), \tag{2.30}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}} = \sigma^2\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^H$ is the covariance matrix and $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_M) = \mathrm{diag}(\boldsymbol{\gamma})$ is the diagonal matrix with all the γ's in the diagonal. For the correct values of γ, the posterior distribution can be computed as

$$\mathrm{p}(\mathbf{x}|\mathbf{y};\hat{\boldsymbol{\gamma}}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \tag{2.31}$$

here $\boldsymbol{\mu}_{\mathbf{x}} = E[\mathbf{x}|\mathbf{y};\hat{\boldsymbol{\gamma}}] = \hat{\boldsymbol{\Gamma}}\mathbf{A}^H(\sigma^2\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Gamma}}\mathbf{A}^H)^{-1}\mathbf{y}$ and $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathrm{Cov}[\mathbf{x}|\mathbf{y};\hat{\boldsymbol{\gamma}}] = \hat{\boldsymbol{\Gamma}}\mathbf{A}^H(\sigma^2\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Gamma}}\mathbf{A}^H)^{-1}\mathbf{A}\hat{\boldsymbol{\Gamma}}$.

To further update the $\boldsymbol{\gamma}$, the expectation maximization (EM) algorithm is applied, and the update rule for $\boldsymbol{\gamma}$ can be given as

$$\gamma_i = \boldsymbol{\mu}_x^2(i) + \boldsymbol{\Sigma}_x(i,i). \tag{2.32}$$

The updated hyperparameter with nonzero values leads to the sparse solution for $\mathbf{x}$. The further improvement in estimating the γ can be updated using the fixed point update rule given in [41–44]. The fixed point update rule is advantageous over EM algorithms in terms of faster convergence. The fixed point update rule can be found by differentiating the above cost function with respect to γ and equating to zero, computed as

$$\hat{\gamma}_m^{new} = \hat{\gamma}_m^{old} \frac{||\mathbf{y}^H \mathbf{\Sigma}^{-1} \mathbf{a}_m||_2^2}{\mathbf{a}_m^H \mathbf{\Sigma}^{-1} \mathbf{a}_m} \tag{2.33}$$

$$\tag{2.34}$$

$$= \hat{\gamma}_m^{old} \frac{Tr[\mathbf{S}_y \mathbf{\Sigma}^{-1} \mathbf{a}_m \mathbf{a}_m^H \mathbf{\Sigma}^{-1}]}{\mathbf{a}_m^H \mathbf{\Sigma}^{-1} \mathbf{a}_m} \tag{2.35}$$

where $\mathbf{S_y} = \mathbf{y}\mathbf{y}^H$ is the sample covariance matrix and $Tr[.]$ denotes the trace (sum of diagonal elements) of a matrix. The results of SBL and further extension as multi-snapshot SBL are given in [45–48].

$$\hat{\gamma}_m^{new} = \hat{\gamma}_m^{old} \frac{\sum_{l=1}^{L} |\mathbf{y}_l^H \mathbf{\Sigma}_{\mathbf{y}_l}^{-1} \mathbf{a}_m|^2}{\sum_{l=1}^{L} \mathbf{a}_m^H \mathbf{\Sigma}_{\mathbf{y}_l}^{-1} \mathbf{a}_m} \tag{2.36}$$

The results in the literature show that SBL performance is very robust to noise, coherence, and multipath effects. It has been shown that compared to other existing methods, SBL gives high resolution even for a single snapshot. Hence, it is a potential candidate to perform well in real-time scenarios for better localization of sources.

2.4 Comparison of traditional and CS-based algorithms

In this section, Table (2.1) presents a comparison of the different localization algorithms discussed so far. The table lists each method along with its advantages, disadvantages, and computational complexity per iteration. The following terms are used to describe the complexity.

- N - number of sensors,
- M - number of candidate points on predefined angular grid,
- L - number of snapshots,
- K - number of sources,
- N_k - number of DFT components and
- I - number of iterations
- N_τ - number of computed for each GCC function in the time-lag domain.

Algorithm	Description	Disadvantage	Complexity
CBF	spatial filtering, robust to noise	low resolution, difficult to localize multiple sources	$O(MNL)$
MUSIC	high resolution subspace based method	requires more number of snapshots	$O(N^3 + N^2ML)$
GCC-PHAT	time domain method, computationally efficient	difficult to localize multiple sources	$O(N_k N_\tau N^2) + O(M)$
SRP-PHAT	time domain method, averaging across GCC-PHAT pairs, robust to reverberation	computationally costly, need more sensor data	$O(MN_kN)$
M-BP	CS based method, l_1 minimization, uses convex optimization	computationally costly, intractable for large size data	$O(M^3L^3)$
OMP	CS based method, computationally efficient	greedy algorithm, need more measurements	$O(NMK)$
M-SBL	Bayesian probabilistic framework, coherent processing of frequencies	computationally costly	$I(N^3 + N^2ML)$

Table 2.1 Comparison of traditional and CS-based localization algorithms

2.5 Data-driven Methods

Recently, data-driven methods have shown promising results for sound source localization in reverberant and low SNR scenarios [19, 49–51]. As a result, various deep neural network (DNN) based architectures such as feedforward neural networks (FFNN), convolutional neural networks (CNN), recurrent neural networks (RNN), convolutional recurrent neural network (CRNN) and encoder-decoder architectures are proposed in recent years [19]. Most of the reported works have indicated the superiority of DNN-based sound source localization (SSL) methods over classical localization approaches in terms of high resolution and low erroneous DOAs [19, 52–57]. Typically low-level signal representations such as waveforms or spectrograms, power spectrums obtained from traditional methods such as CBF, MUSIC, GCC-PHAT are fed into the DNN architectures and the features are learned to improve the localization accuracy. In this section, we explore two specific approaches: regression and classification, using DNN for solving the DOA estimation problem. For an extensive literature survey on different DNN techniques used for DOA estimation, please refer to [19].

2.5.1 Localization as classification problem

In DNN methods, DOA estimation is often approached as a classification problem, where each class represents a specific zone in the search space. The search space is divided into several subregions of similar size, and a neural network is trained to provide a probability of active source presence for each subregion. This classification approach utilizes feedforward layers with softmax or sigmoid activation functions in the final layer to produce spatial (pseudo)-spectrums indicating high probabilities of source

activity in corresponding zones [19]. Peak-picking algorithms are then employed to extract DOA estimates, either by selecting the J highest peaks for known source counts or by choosing peaks above a threshold for joint estimation of source count and localization. Spherical coordinates are commonly used, with the azimuth angle quantized into N_θ regions [55, 56, 58–62]. Elevation estimation is less explored. Recent SSL neural networks often estimate both azimuth and elevation using separate output layers or a single layer representing zones on the unit sphere [63, 64]. Distance estimation is challenging and has received limited attention. Some studies also consider Cartesian coordinates, dividing the horizontal plane into regions for classification [65, 66]. Classification methods suffer from a decreasing angular difference between regions far from the microphone array, resulting in regression as the preferred method for estimating Cartesian coordinates.

2.5.2 Localization as regression problem

Regression-based DNN networks provide source location estimates as continuous values, which offers the advantage of potentially more accurate DOA estimation without quantization. However, it has two drawbacks: the need for a known or assumed number of sources, as regression alone, cannot determine source activity, and the source permutation problem inherent in multi-source localization, common in deep learning-based source separation methods [67, 68]. In regression-based sound source localization, the choice of coordinate type to be estimated is often driven by the context or application, as regression typically requires only a small number of output neurons [57, 69, 70]. In terms of coordinate estimation, spherical coordinates such as azimuth and elevation have been used by various methods with different network architectures and output representations. Cartesian coordinates have also been explored, with systems estimating (x, y) or all three coordinates (x, y, z) using regression techniques [60, 71, 72]. In the context of the DCASE 2019 Challenge, several SELD systems have used regression for Cartesian coordinates [7, 51, 73, 74].

2.5.3 SELD tasks

In the recent past, polyphonic sound event localization and detection (SELD) problems have garnered a lot of attention among researchers, which combine the detection and localization tasks and have many practical applications [7, 51, 54]. SELD is a crucial task in acoustic scene analysis, aiming to identify and locate multiple sound events within an audio signal while providing detailed spatial information like azimuth and elevation angles and temporal occurrence. It comprises two key components: sound event detection, for identifying sound event presence and temporal boundaries, and sound source localization, for estimating spatial coordinates. SELD has wide-ranging applications such as surveillance, robotics, and immersive audio, empowering machines to better perceive and interact with their acoustic environment. In [51], the authors presented the pioneering work of SELDnet, the first paper addressing the SELD task, introducing a convolutional recurrent neural network that achieves simultaneous sound event recognition, localization, and tracking while showcasing robust performance under diverse conditions and emphasizing the significance of larger real-life training datasets for enhanced real-world

applicability. Since its introduction, researchers have proposed various model architectures and features to improve SELD performance [70, 72–75].

2.6 Gridless algorithms

The grid-based methods discretize the search space into a predefined angular grid of possible DOAs, resulting in limited resolution, particularly when the actual DOAs don't lie on predefined grid points. In contrast, gridless localization algorithms enable a more detailed search in the continuum, potentially enhancing resolution and localization accuracy. In this section, we discuss gridless methods NOMP, SFW and briefly discuss ANM based methods.

2.6.1 Atomic Norm Minimization (ANM) Based Methods

To overcome the basis mismatch problem in compressive sensing problems, the DOA estimation problem is formulated in a continuous angular space and the atomic norm is used as a sparse promoting measure for general signals [76]. ANM is a mathematical optimization problem that has infinitely many unknowns and is solved efficiently over a few optimization variables in the dual domain with semidefinite programming [76]. The DOA estimation problem is formulated as a minimization problem where the objective is to minimize the atomic norm of the sparse signal subject to the observed measurements and the array geometry. The atoms correspond to possible DOAs in continuous angular space. The atomic norm of a signal is a measure of its sparsity in the atom domain. ANM based optimization problem is solved using semi-definite programming in 1D and 2D scenarios [20, 76–84]. In [77], the authors present ANM for the exact recovery of the unknown frequencies even if the continuous dictionary is not incoherent at all and does not satisfy any sort of restricted isometry conditions. In [78], the authors present positive semidefinite programs for ANM in recovering high-dimensional frequencies by transforming dual problems to equivalent positive semidefinite program by using positive trigonometric polynomials. In [76] ANM is applied for grid-free compressive beamforming. The work in [85] unifies the two techniques; one is based on covariance fitting from a statistical perspective and termed as the gridless SPICE, and the other uses the deterministic atomic norm by interpreting gridless SPICE (GLS) as atomic norm methods in various scenarios for MMV model. Additionally, gridless methods have been applied for non-uniform arrays and wideband processing [86–91].

2.6.2 Newtonized Orthogonal Matching Pursuit (NOMP)

The Newtonized OMP (NOMP) algorithm is a generalization of OMP to the continuous angular space ($\boldsymbol{\theta} \in [0^\circ, 360^\circ]$) that employs Newton's steps to refine source parameters in each iteration [21]. The iterative process comprises two main phases: identifying a new source and applying Newton refinements to enhance the parameters of previously identified sources. The Newton refinement process plays a critical role for two reasons: 1) avoiding potential basis mismatch resulting from discretizing

a continuous parameter space, and 2) providing feedback to locally improve parameters estimated in earlier iterations [21]. The primary steps of NOMP are briefly outlined below:

- **Grid-based initial estimation:** The DOA and corresponding amplitude are estimated for each source sequentially within the predefined discretized grid $\boldsymbol{\theta}_d$. The coarse estimates and amplitudes are computed by solving

$$\hat{\theta}_k = \arg\max_{\boldsymbol{\theta}_d} |\langle \mathbf{r}, \mathbf{a} \rangle|, \tag{2.37}$$

  where $\mathbf{r}$ is the residual initialized as measurement vector $\mathbf{y}$. The corresponding source amplitude is obtained using the following:

$$\hat{x}_k = \mathbf{a}^H(\hat{\theta}_k)\, \mathbf{y}. \tag{2.38}$$

- **Local Newton optimization:** Ideally, the DOA should be estimated on the continuum, i.e., solving (2.37) over $\boldsymbol{\theta} \in [0^\circ, 360^\circ]$. Newton's step helps to search over the continuum by locally refining the estimate ($\hat{\theta}_k$) obtained by picking the maximum over the discrete set $\boldsymbol{\theta}_d$. Once the locally optimized DOA is obtained, the source amplitude and residual measurement get updated.

- **Global cyclic feedback optimization:** After completing the local optimization, we circularly optimize the DOA and amplitude of all the identified K acoustic sources. This is the additional step in NOMP, which provides feedback for local refinements of previously estimated DOAs. This helps to better explain the measurements in light of new information regarding the presence of another source. This feedback is presented in the form of an updated residue. This step is crucial for fast convergence and high estimation accuracy.

- **Update by least squares:** Here the residual is updated by projecting measurements $\mathbf{y}$ onto the subspace spanned by the estimated sources. This ensures that the residual energy is the minimum possible for the current set of estimated DOAs.

However, the method encounters difficulties with coarse grids, where local Newton iterations fail to converge towards the objective function's local minimum [21, 92].

2.6.3 Sliding Frank-Wolfe (SFW)

An alternative gridless approach is the Sliding Frank-Wolfe (SFW) algorithm, which solves the Beurling LASSO problem, i.e., a traditional LASSO in the continuum [22, 93]. Let us consider that the sources are situated within a specified region denoted as $\mathcal{B}$. The arrangement of these sources is characterized by a measure denoted as ν. This measure is a function that accepts a subset of $\mathcal{B}$ as input and produces a positive, real, or complex value. The purpose of this measure ν is to represent the distribution of sources within the domain of interest $\mathcal{B}$, eliminating the necessity for a discrete grid. Dirac mass function is one example of measure ν that can model a point source of unit amplitude at angle

$\theta \in \mathcal{B}$ [91]. The Beurling LASSO problem is formulated as

$$\nu^* = \underset{\nu \in \mathcal{G}}{\arg\min}\ \frac{1}{2}\left|\left|\int_{\mathcal{B}} \mathbf{a}(\theta)\, d\nu - \mathbf{y}\right|\right|_2^2 + \lambda|\nu|(\mathcal{B}) \tag{2.39}$$

where $\mathcal{G}$ is the set of complex measures defined on $\mathcal{B}$. The main difference between LASSO and Beurling LASSO is that the DOA of the estimated sources is not limited to an angular finite grid. The SFW algorithm solves the Beurling LASSO problem by iteratively adding the Dirac masses to the measure, alternating with local updates of the DOA of the Dirac masses [91]. In each iteration of SFW, a new source is added, and the DOA and source amplitude of all sources are optimized locally and jointly. Under specific conditions, particularly when the solution is a finite sum of Dirac masses and is unique, it has been demonstrated that this algorithm converges within a finite number of iterations [93]. The key steps in SFW are summarized in Algorithm 2.

Algorithm 2 Sliding Frank-Wolfe Algorithm

1. $\Lambda^{[0]} \leftarrow \emptyset, \mathbf{r}^{[0]} \leftarrow \mathbf{y}, tol = 1e^{-10}$
2. **for** $k = 1, \ldots, K$
3. Find the next source:
   $\hat{\theta}_k = \underset{\theta_i \in \mathcal{B}}{\arg\max}\ \frac{1}{L}\sum_{l=1}^{L} \left|\mathbf{a}^H(\theta_i)\, \mathbf{r}^{[k-1]}\right|^2 \qquad (a)$
4. $\Lambda^{[\frac{k-1}{2}]} = \{\Lambda^{[k-1]},\ \hat{\theta}_k\}$
5. Optimize the amplitude:
   $\mathbf{x}^{[\frac{k-1}{2}]} = \underset{\mathbf{x} \in \mathbf{X}}{\arg\min}\ \frac{1}{2}\ \left|\left|\mathbf{A}(\Lambda^{[\frac{k-1}{2}]})\, \mathbf{x} - \mathbf{y}\right|\right|_{\mathcal{F}}^2 + \lambda||\mathbf{x}||_1 \qquad (b)$
6. Optimize the amplitudes and parameters:
   $\mathbf{x}^{[k]}, \Lambda^{[k]} = \underset{\Lambda \subset \mathcal{B}, \mathbf{x} \in \mathbf{X}}{\arg\min}\ \frac{1}{2}\ ||\mathbf{A}(\Lambda)\, \mathbf{x} - \mathbf{y}||_{\mathcal{F}}^2 + \lambda||\mathbf{x}||_1 \qquad (c)$
7. $\mathbf{r}^{[k]} \leftarrow \mathbf{y} - \mathbf{A}(\Lambda^{[k]})\, \mathbf{x}^{[k]}$
8. **end for**

Here Λ represents the set of estimated K DOAs. The residual $\mathbf{r}$ is initialized as a measurement vector ($\mathbf{y}$) and updated at each iteration. An iteration consists of the following steps:

- **Adding a source** Similar to the NOMP algorithm, a coarse DOA is estimated and used as initialization for solving optimization problem (a) in algorithm 2.
- **Local amplitude estimation** The estimated DOA is used to compute the amplitude, which is used as initialization to solve the optimization problem (b) in algorithm 2.
- **Joint optimization:** In the last step, both DOAs and amplitudes are jointly optimized as (c) in algorithm 2. This problem is non-convex. At the end of each iteration, Λ and residual get updated.

The same steps are repeated until the K DOAs and their corresponding amplitudes are optimized. The SFW has been extended to 3D acoustic source localization in a grid-free setting [91].

Chapter 3

Analysis of DOA Estimation Algorithms on Real-world Data

This chapter focuses on the performance analysis of DOA estimation algorithms using real-world recordings from two datasets: LOCATA and DCASE. The discussion begins by introducing the LOCATA dataset and explaining its processing steps. It describes how directional information is obtained from the recordings and then applies DOA estimation algorithms such as CBF, MUSIC, and SBL. Furthermore, the chapter analyzes three wideband signal models and compares different variants of wideband SBL (referred to as SBL1, SBL2, and SBL3) that make various assumptions about the source signal power spectrum. The DOA estimation performance of the SBL algorithms is compared with the wideband processing of CBF and MUSIC.

The DCASE dataset is a collection of audio recordings used for developing and evaluating algorithms for SELD task. Deep learning models are typically designed to handle multichannel audio data captured by microphone arrays for the SELD task. We also explore the DNN based architecture and conduct experimental validation using DCASE dataset. We address the issue of unrealistic DOA estimates that arise in many methods due to the absence of temporal continuity models. To improve the accuracy of DOA estimation, we propose an update rule that incorporates predicted DOAs and their derivatives. In summary, this chapter provides a comprehensive and detailed experimental validation of DOA estimation algorithms using real datasets.

3.1 LOCATA dataset and processing

The IEEE-AASP Challenge on sound source LOCATA offers an open-access data corpus of indoor multi-channel audio recordings with multiple mobile sources and ground truth information for performance evaluation. Since its release, various DOA estimation methods have been applied to this dataset, leading to an active area of research and evaluation [12, 94, 95]. The LOCATA development dataset performs the estimation of performing DOA. The LOCATA dataset comprises recordings collected indoors and features various microphone arrays as shown in Fig. 3.1, including: i) a 12-microphone pseudo-spherical array named robothead, ii) a spherical array called eigenmike with 32 sensors, iii) a 15-microphone non-uniform planar array dicit, and iv) a hearing aid with 4 mics. The data is collected indoors, and the dataset offers a range of localization and tracking tasks listed as

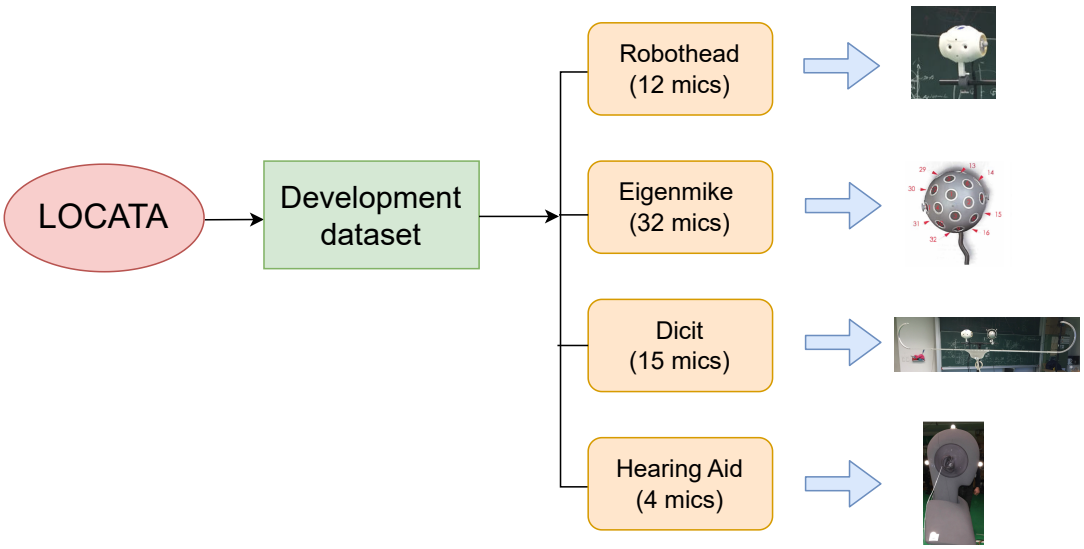

Figure 3.1 Description about LOCATA dataset

- **Task 1:** single static source
- **Task 2:** multiple static sources
- **Task 3:** single moving source
- **Task 4:** multiple moving sources
- **Task 5:** single moving source when the array is also in motion
- **Task 6:** multiple moving sources when the array is also in motion

The LOCATA recordings have challenging scenarios such as near-field sources, reverberation, and ambient noise (from a road in front of the building, measurement noise, and traffic sounds outside the recording space). The recordings took place indoors in a 7.1 × 9.8 × 3 m^3 room, featuring a reverberation time of approximately 0.55 seconds [96]. Further details and assumptions about the LOCATA dataset can be found in [12, 94, 95, 97]. The data processing pipeline is shown in Fig. 3.2. We first perform a short-time Fourier transform (STFT) on audio signals. Each block comprises 100 STFT frames/snapshots and has a 90% overlap in block-level processing. The DOA estimation algorithms are then applied to each data block, and the angular spectrum is computed. The DOA is obtained by identifying the peak locations in the angular power spectrum. After estimating the DOAs, we perform error analysis using the available ground truth, considering errors only when the voice activity detector (VAD) [98] indicates the presence of voice. During data processing, the following parameter values are utilized: an FFT size of 1024, a frequency range of [800, 2800] Hz, and a snapshot/frame duration of 0.03 seconds. For multi-frequency processing using CBF and MUSIC, the power spectrum is computed at each frequency, and averaging is performed across narrowband spectrums. On the other hand, the wideband SBL1 and SBL3 process all frequencies coherently (to be discussed later).

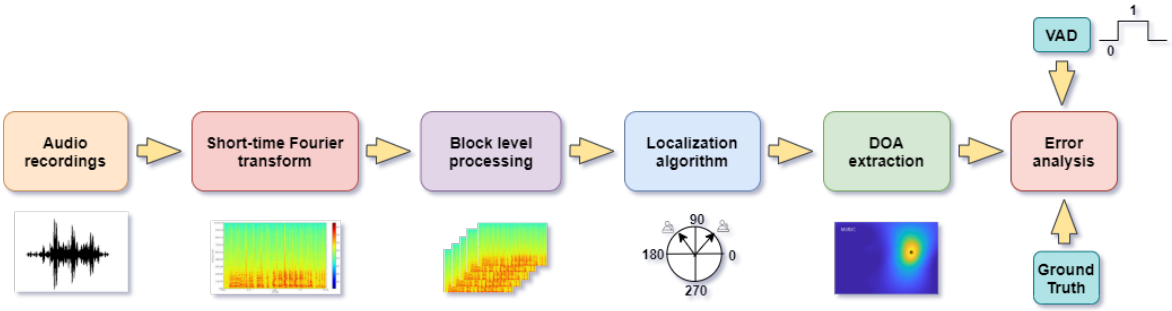

Figure 3.2 Various steps involved for LOCATA processing

3.2 Performance analysis of CBF, MUSIC and SBL

As discussed earlier, CBF exhibits robustness to noise but lacks resolution, making it challenging to localize closely spaced sources. On the other hand, while the MUSIC method offers high resolution, it typically requires a large number of snapshots. In challenging environments with random noise and low SNR, there is a need for high-resolution methods that can operate with fewer snapshots. For localizing audio sources with rich frequency spectra, utilizing multi-frequency SBL methods is advantageous [46, 99]. Here, we evaluate SBL along with traditional DOA estimation methods of CBF and MUSIC on various source localization tasks from the LOCATA dataset. While processing multi-frequency data, the narrowband spectra obtained from CBF and MUSIC are averaged across the frequencies. The recordings from **robot-head** and **eigenmike** for Task 1, 3, and 5 are considered for single source localization. The comparative study shows that SBL significantly outperforms CBF and MUSIC on all considered tasks.

3.2.1 Performance metric

The DOA is estimated as the peak location of the power spectrum computed by DOA estimation algorithms for each block. The spectrum is computed with 1° resolution both in azimuth and elevation. The error between estimated and true DOAs are computed at block level during voice activity periods [98]. In instances where the signal energy surpasses the predefined threshold, the VAD is active/on; otherwise, it remains deactivated. The mean absolute error (MAE), root mean squared error (RMSE), and standard deviation (Std Dev) of the estimates are computed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\phi_i - \hat{\phi}_i)^2}{N}}, \tag{3.1}$$

$$\text{MAE} = \frac{\sum_{i=1}^{N}|(\phi_i - \hat{\phi}_i)|}{N}, \tag{3.2}$$

$$\text{Std Dev} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|(\phi_i - \hat{\phi}_i) - \frac{\sum_{i=1}^{N}(\phi_i - \hat{\phi}_i)}{N}|^2}, \tag{3.3}$$

where ϕ, $\hat{\phi}$, and N are the true DOA, estimated DOA, and the total number of DOAs, respectively. For LOCATA, we compute the RMSE for each recording and report the average RMSE over all recordings for each of the tasks. A source is said to be misdetected if the estimated DOA is more than ζ° away from the true DOA. The probability of detection (P_d) for DOA estimation algorithms is also computed as $P_d(\zeta) = 1 - \frac{N_{\text{miss}}(\zeta)}{N_{\text{total}}}$, where $N_{\text{miss}}(\zeta)$ is the number of misdetections (over all the recordings in each task) and N_{total} is total number of blocks.

3.2.2 Results for robot-head array

For robot-head, DOA estimation and tracking errors have been computed for all the tasks as shown in Table 3.1. Localization of single stationary talker (Task 1) and all algorithms give relatively low error as seen from Table 3.1. It can be seen from Fig. 3.3 that in the case of Task 1 using SBL, 97% of sources are detected within 10° of true DOA for the robothead. For Task 3, the moving talker causes its distance from the stationary microphone array to change and has poorer DOA estimation performance compared to Task 1. In Task 5, the talker is moving as well as the microphone arrays installed on the platform are moving. In terms of probability of detection, SBL performs significantly better than CBF and MUSIC for low values of ζ (Fig. 3.3) and the computed errors are least for SBL, followed by that of MUSIC and CBF for all tasks.

Task	Method	MAE		RMSE		Std Dev	
		az	el	az	el	az	el
T1	CBF	4.48	10.0	5.27	12.2	0.04	0.12
	MUSIC	1.96	3.94	2.20	5.91	0.01	0.06
	SBL	**1.10**	**3.57**	**1.25**	**3.72**	**0.01**	**0.01**
T3	CBF	4.37	6.09	8.37	12.3	0.12	0.18
	MUSIC	8.70	10.6	15.8	16.5	0.23	0.21
	SBL	**3.82**	**3.16**	**6.37**	**5.45**	**0.08**	**0.07**
T5	CBF	15.7	11.7	36.7	18.6	0.58	0.25
	MUSIC	8.36	9.85	23.4	17.2	0.58	0.24
	SBL	**2.98**	**3.93**	**10.5**	**7.57**	**0.17**	**0.11**

Table 3.1 Error performance of robot-head array for Tasks 1,3 and 5 (averaged over all recordings)

3.2.3 Results for eigenmike array

The MAE for Task 1, 3, and 5 using eigenmike is shown in Fig. 3.4. Due to rotations of the eigenmike within the shock mount, it is susceptible to scattering effects [94], which results in high azimuth error (Fig. 3.4, note that the two plots have different vertical range). The estimates of azimuth and elevation for an eigenmike recording from Task 3 are shown in Fig. 3.5 and Fig. 3.6. The estimates obtained from SBL are much closer to ground truth compared to CBF and MUSIC when voice activity is present (denoted by 1 in VAD plot). The VAD plot in Fig. 3.5 and Fig. 3.6 shows the output of the voice activity detector. In instances where the signal energy surpasses the predefined threshold, the VAD is on (shown as 1); otherwise, it remains deactivated, denoted by 0.

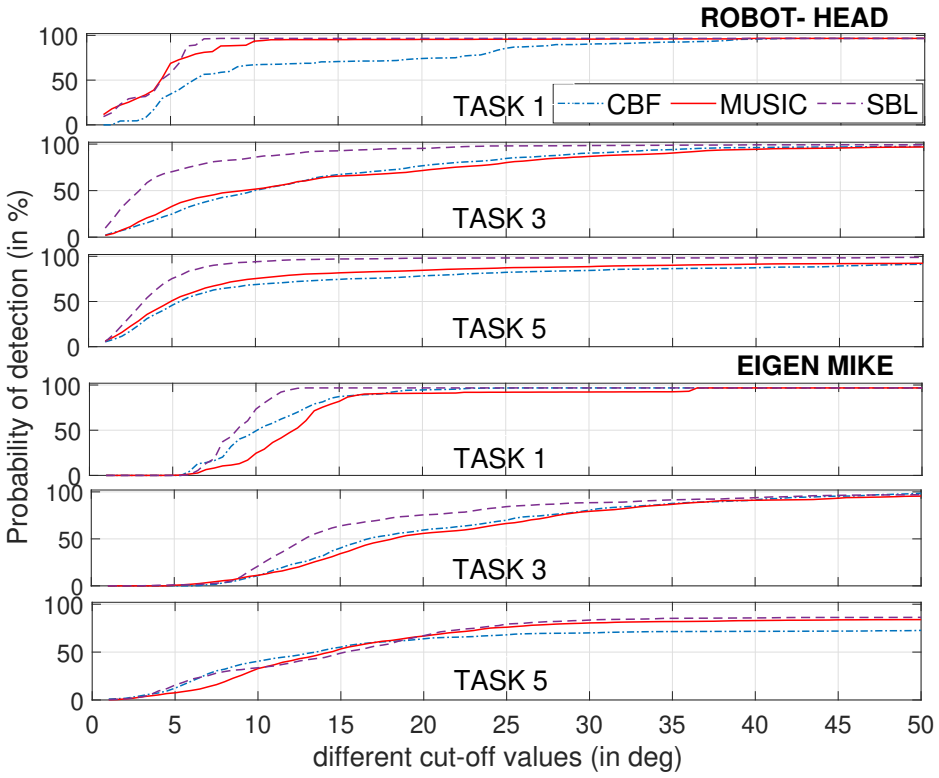

Figure 3.3 Probability of detection (P_d) vs cutoff ζ

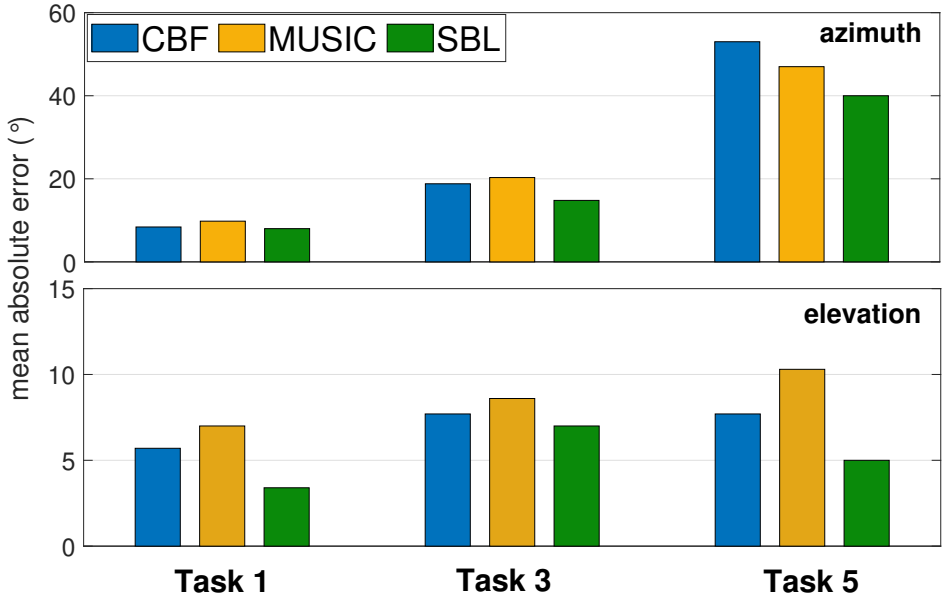

Figure 3.4 Azimuth and elevation error using eigenmike for Task1, 3 and 5 averaged over all recordings

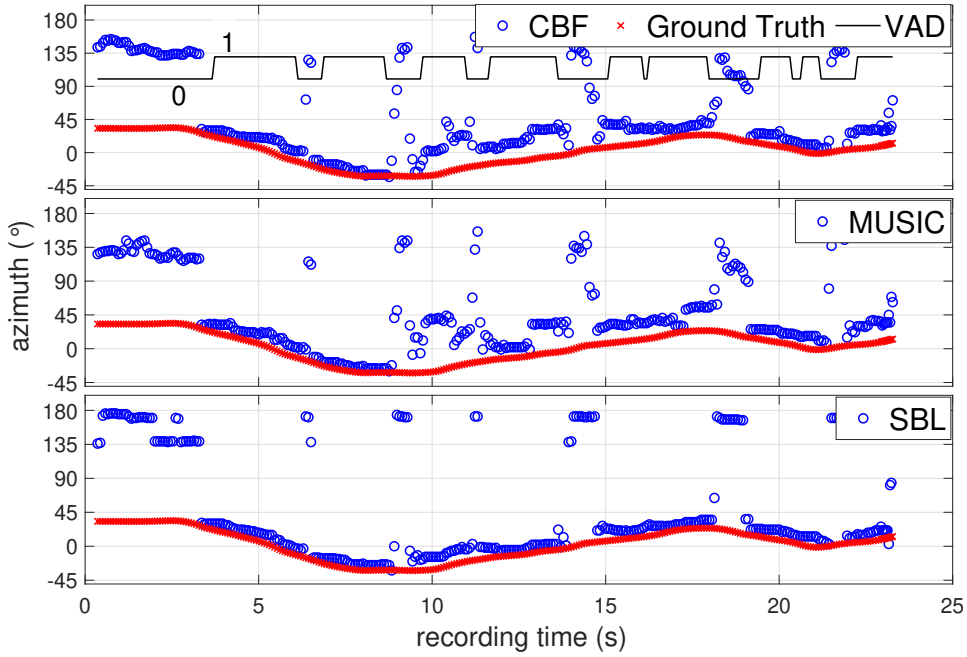

Figure 3.5 Azimuth DOA estimates (in °) of a single target using eigenmike array, Task 3, recording 2

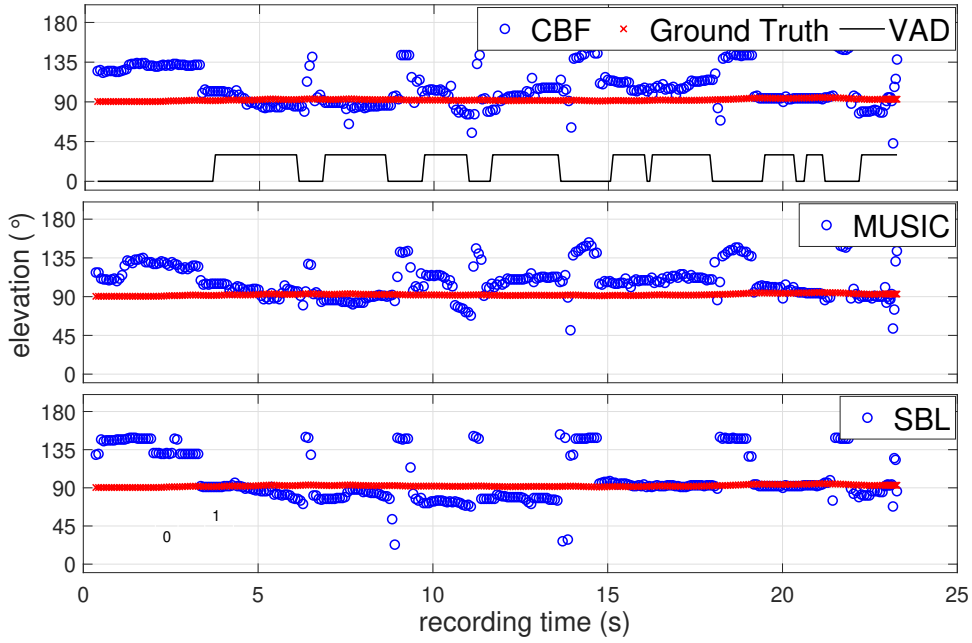

Figure 3.6 Elevation DOA estimates (in °) of a single target using eigenmike array, Task 3, recording 2

3.3 Performance analysis of wideband signal models and SBL algorithms

As demonstrated in the earlier section, SBL has been successful in DOA estimation due to its robustness and high resolution using a few snapshots. Most wideband SBL algorithms make the simplifying assumption that distinct sources have the same power spectrum across frequency bands [42, 45, 46, 99, 100]. However, this assumption may not be true in practice (for example, speech signals). Recently, a realistic model where the source variance is allowed to change by the same factor for each source across frequencies is proposed in [101]. We consider three signal models: **Model1** has a flat signal spectrum [42, 45, 46, 99, 100], **Model2** has independent spectrum for each source, and **Model3** [101] has a shared colored spectrum with independent magnitude scaling for each source. We investigate the performance of three wideband SBL variants: SBL1, SBL2, and SBL3 (each derived under the corresponding model assumption). We discuss three variants of wideband SBL (SBL1, SBL2, and SBL3) with different source signal power spectrum assumptions. The localization performance of SBL algorithms is compared with wideband processing of CBF and MUSIC. The experimental validation is presented using simulated data and experimental LOCATA data. This comparative study shows that SBL3, which simultaneously enforces sparsity and models frequency-dependent signal spectrum, performs superior in most scenarios.

3.3.1 Wideband signal models

We discuss three wideband signal models, each with distinct assumptions on the signal spectrum. Consider a multi-frequency, multi-snapshot signal model as

$$\mathbf{Y}_f = \mathbf{A}_f \mathbf{X}_f + \mathbf{N}_f, \tag{3.4}$$

where $f = 1, \ldots, F$ is frequency index, $\mathbf{Y}_f \in \mathbb{C}^{N \times L}$ is the L snapshot observation matrix received from N sensors, $\mathbf{A}_f \in \mathbb{C}^{N \times M}$ is the sensing matrix with mth column consisting of the steering vector $\mathbf{a}_{fm} = [1, e^{j\pi d \frac{c}{f_m} \sin\theta_m}, \ldots, e^{j\pi(N-1)d \frac{c}{f_m} \sin\theta_m}]^T$ for a source located at angle θ_m, $\mathbf{X}_f \in \mathbb{C}^{M \times L}$ is source amplitude, and $\mathbf{N}_f \in \mathbb{C}^{N \times L}$ is additive gaussian noise. The angles θ_m range over the 1-D angular search grid $[-\frac{\pi}{2}, \frac{\pi}{2}]$ containing M discrete locations in azimuth. The columns of $\mathbf{X}_f$ are assumed to be sparse with K non-zero entries corresponding to K active sources with $K < N$ and $K \ll M$. We want to estimate the K active DOAs from the M candidate DOAs. The columns of $\mathbf{X}_f$ are assumed zero-mean complex Gaussian with diagonal covariance matrix $\mathbf{\Gamma}_f$. As discussed next, the assumptions made on $\mathbf{\Gamma}_f$ give rise to various signal spectrum models.

Model1: For Model1, the variance of source amplitudes is constant across frequency bands; hence the spectrum is flat, i.e.,

$$\mathbf{\Gamma}_1, ..., \mathbf{\Gamma}_F = \mathbf{\Gamma}. \tag{3.5}$$

Model2: For Model2, we assume the source variance is changing across frequency bands, hence a colored spectrum,

$$\mathbf{\Gamma}_1, ..., \mathbf{\Gamma}_F. \tag{3.6}$$

Model3: Model1 discussed above is too restrictive as practical sources rarely have a flat spectrum. Though Model2 is very general, it does not impose sparsity constraints across frequency bands. An intermediate model that allows for spectrum to vary with frequency band and also keep sparsity same across frequency bands can be given as [101],

$$\mathbf{\Gamma}_f = c_f \mathbf{\Gamma}. \tag{3.7}$$

The parameter c_f accounts for the non-flat frequency spectrum. Note that this model restricts all the sources to have the same spectrum up to a scaling constant.

3.3.2 Wideband SBL

In this section, we briefly discuss wideband SBL algorithms [45, 46, 100]. SBL uses a Bayesian framework with prior parameterized by $\mathbf{\Gamma}_f$. SBL estimates $\mathbf{\Gamma}_f$ whose diagonal is sparse [43] and hence columns of $\mathbf{X}_f$ are also sparse. The non-zero locations of this diagonal give us the required DOAs. Under assumptions of independence across snapshots and frequencies, the likelihood is

$$p(\mathbf{Y}_{1:F}|\mathbf{X}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{Y}_f|\mathbf{X}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} p(\mathbf{y}_{fl}|\mathbf{x}_{fl}). \tag{3.8}$$

In multi-snapshot, multi-frequency, SBL formulation [44, 100], the source amplitudes $\mathbf{x}_{fl}$ are independent, zero-mean, complex Gaussian vectors with possibly frequency-dependent diagonal covariance $\mathbf{\Gamma}_f = \text{diag}(\boldsymbol{\gamma}_f) = \text{diag}([\gamma_{f,1}, \ldots, \gamma_{f,M}])$ giving the prior

$$p(\mathbf{X}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{X}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} p(\mathbf{x}_{fl}). \tag{3.9}$$

The sparsity of source amplitude vectors is related to the sparsity of the parameter vector $\boldsymbol{\gamma}_f$. As prior (3.9) and likelihood (3.8) are Gaussian, the evidence $p(\mathbf{Y}_{1:F})$ is also Gaussian

$$p(\mathbf{Y}_{1:F}) = \prod_{f=1}^{F} p(\mathbf{Y}_f) = \prod_{f=1}^{F}\prod_{l=1}^{L} \mathcal{CN}(\mathbf{y}_{fl}; \mathbf{0}, \mathbf{\Sigma}_f), \tag{3.10}$$

where $\mathbf{\Sigma}_f = \sigma_f^2\mathbf{I} + \mathbf{A}_f\mathbf{\Gamma}_f\mathbf{A}_f^H$ is covariance matrix, which changes with assumption on $\mathbf{\Gamma}_f$ and $\mathcal{CN}()$ denotes complex Gaussian density function. The SBL method estimates the unknown param-

eters $\boldsymbol{\gamma}_f, f = 1, \ldots, F$ by maximizing the evidence $p(\mathbf{Y}_{1:F})$

$$\{\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_F\} = \underset{\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_F\}}{\arg\max}\ p(\mathbf{Y}_{1:F})\,, \tag{3.11}$$

$$= \underset{\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_F\}}{\arg\min} \sum_{f=1}^{F} \sum_{l=1}^{L} \Big(\mathbf{y}_{fl}^H \boldsymbol{\Sigma}_f^{-1} \mathbf{y}_{fl} - \log|\boldsymbol{\Sigma}_f| \Big). \tag{3.12}$$

SBL1: The first variant of multi-frequency SBL follows Model1 and assumes $\boldsymbol{\Gamma}_f$ constant across all frequencies i.e. flat spectrum. For SBL1, the covariance matrix becomes $\boldsymbol{\Sigma}_f = \sigma_f^2 \mathbf{I} + \mathbf{A}_f \boldsymbol{\Gamma} \mathbf{A}_f^H$ and only a single vector $\boldsymbol{\gamma}$ has to be estimated. To find the minimum of this non-convex objective function, we differentiate with respect to $\boldsymbol{\gamma}$ and equate to zero. For details about this procedure, see [44, 100]. The resulting fixed-point update is,

$$\gamma_m^{\text{new}} = \gamma_m^{\text{old}} \left(\frac{\sum_{f=1}^{F} \sum_{l=1}^{L} |\mathbf{y}_{fl}^H \boldsymbol{\Sigma}_f^{-1} \mathbf{a}_{fm}|^2}{\sum_{f=1}^{F} \mathbf{a}_{fm}^H \boldsymbol{\Sigma}_f^{-1} \mathbf{a}_{fm}} \right), \tag{3.13}$$

where γ_m is the m^{th} element of $\boldsymbol{\gamma}$ and $\mathbf{a}_{fm}$ is the m^{th} column of the dictionary matrix $\mathbf{A}_f$. At convergence, the estimate $\hat{\boldsymbol{\gamma}}$ is sparse [44, 100], which in turn enforces source amplitudes $\mathbf{x}$ to be sparse. As γ_m is the source power corresponding to m^{th} DOA, $\hat{\boldsymbol{\gamma}}$ is also called the SBL flat power spectrum.

SBL2: The flat signal spectrum assumption of SBL1 is restrictive and often violated for real-world signals. Following Model2, in SBL2 we assume the variance of source amplitudes to be frequency dependent. Solving the optimization in (3.12) is then equivalent to computing $\boldsymbol{\gamma}_f$ using the update rule (3.13) at each frequency. A single $\hat{\boldsymbol{\gamma}}$ is obtained by averaging,

$$\hat{\boldsymbol{\gamma}} = \frac{1}{F} \sum_{f=1}^{F} \hat{\boldsymbol{\gamma}}_f. \tag{3.14}$$

SBL3: It follows Model3 and frequency dependence of source variance is modelled using the parameter c_f as $\boldsymbol{\Gamma}_f = c_f \boldsymbol{\Gamma}$. Assuming c_f to be known, solving (3.12) gives the same update rule as (3.13) with $\boldsymbol{\Sigma}_f = \sigma_f^2 \mathbf{I} + c_f \mathbf{A}_f \boldsymbol{\Gamma} \mathbf{A}_f^H$. An update rule for c_f is derived by differentiating evidence with respect to c_f [101]. Here we use an empirical estimate $c_f \propto \text{Trace}(\mathbf{S}_{\mathbf{y}_f})$, where $\mathbf{S}_{\mathbf{y}_f} = \frac{1}{L} \mathbf{Y}_f \mathbf{Y}_f^H$ and normalize so that $\sum_{f=1}^{F} c_f = 1$. Note that $\text{Trace}(\mathbf{S}_{\mathbf{y}_f})$ is a measure of signal energy at f-th frequency. This estimate is intuitive as it assigns a higher c_f value for dominant frequencies present in the signal. An alternate estimate of cf could be the largest singular value of the matrix $\mathbf{S}_{\mathbf{y}}$. However, from our simulations, we observed that both give similar performance. The computational complexity of SBL1, SBL2 and SBL3 are same and given by $\mathcal{O}(F \times (N^3 + MN^2 + MNL))$ [100].

Noise estimate: Along with estimating $\boldsymbol{\Gamma}_f$, it is advantageous to estimate the variance of unknown noise for faster convergence of SBL as it controls the sharpness of peaks [44]. Stochastic maximum

likelihood [44, 100, 102] is used for efficient estimation of σ_f^2 given by

$$\hat{\sigma}_f^2 = \frac{\text{Trace}[(\mathbf{I}_n - \mathbf{A}_{\mathcal{M},f}\mathbf{A}_{\mathcal{M},f}^{+})\mathbf{S}_{\mathbf{y}_f}]}{N-K}, \tag{3.15}$$

where $\mathcal{M}$ is the set of indices where $\hat{\boldsymbol{\gamma}}$ is non-zero ($|\mathcal{M}| = K =$ number of sources), $\mathbf{A}_{\mathcal{M},f}$ is the matrix $\mathbf{A}_f$ restricted to columns corresponding to indices $\mathcal{M}$ and $\mathbf{A}_{\mathcal{M},f}^{+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}_{\mathcal{M},f}$. For noise estimation, K is assumed to be known.

3.3.3 Performance analysis

We apply DOA estimation algorithms on both simulated data and audio signals from the LOCATA dataset to compare the performance of different models and algorithms. A source is misdetected if the estimated DOA is more than 30° away from the true DOA. We report the probability of detection (P_d) as (1 - # missed sources/ # total active sources). For the correctly detected sources, RMSE is computed as $\sqrt{\frac{\sum_{i=1}^{N_s}\sum_{j=1}^{N}(\phi_{ij}-\hat{\phi}_{ij})^2}{N_{sim}\ N_s}}$, where ϕ, $\hat{\phi}$, and Ns are the true DOA, estimated DOA, and the number of sources respectively. N_{sim} represents the number of Monte Carlo simulations for simulated data and the number of blocks for LOCATA. For LOCATA, we compute the RMSE and PD, averaged over all recordings.

3.3.4 Simulated models

Signals are generated using all three models discussed in Section 3.3.1 following various assumptions on $\boldsymbol{\Gamma}$. For Model1, $\gamma_m = 1$ for each source. For Model2, γ_m is drawn from uniform distribution $\text{U}[0, 10]$ at each frequency. For Model3 we sample the parameters $c_f \sim \text{U}[0, 10]$. We define signal-to-noise-ratio $\text{SNR} = 10\log_{10}(\frac{\sigma_x^2}{\sigma_n^2})$, where σ_x^2 and σ_n^2 are variance of signal and noise respectively. For Model1 $\sigma_x^2 = \gamma_m = 1$, and for Model2 and Model3, it is taken to be the variance of the uniform distribution $\text{U}[0, 10]$. For all models, the noise variance is computed using signal variance and the desired SNR value. For simulations, we use 5 sensors with 8 cm spacing in a uniform linear array and equispaced frequencies from 800–2100 Hz. For simulations, DOAs are computed over a predefined angular grid from -90° to 90° with spacing of 1°.

3.3.5 Results for simulated data

We compare the results of SBL with CBF and MUSIC, and we obtain their wideband power spectrum as the average of their narrowband power spectrum. To compare the methods, the RMSE is computed over 500 Monte Carlo simulations at different SNRs and the number of snapshots. Fig. 3.7 shows the RMSE vs SNR (ranging from -10 to 10 dB) and snapshots (ranging from 1 to 100). For both plots, we consider two sources located at -10° and -25°. For RMSE vs SNR, 100 snapshots and for RMSE vs snapshots, SNR = 10 dB are fixed. For RMSE vs SNR, SBL1 and SBL3 performances are very similar across all the models and SNR values. At SNR $\geqslant$ 2 dB, all SBL algorithms have similar performance

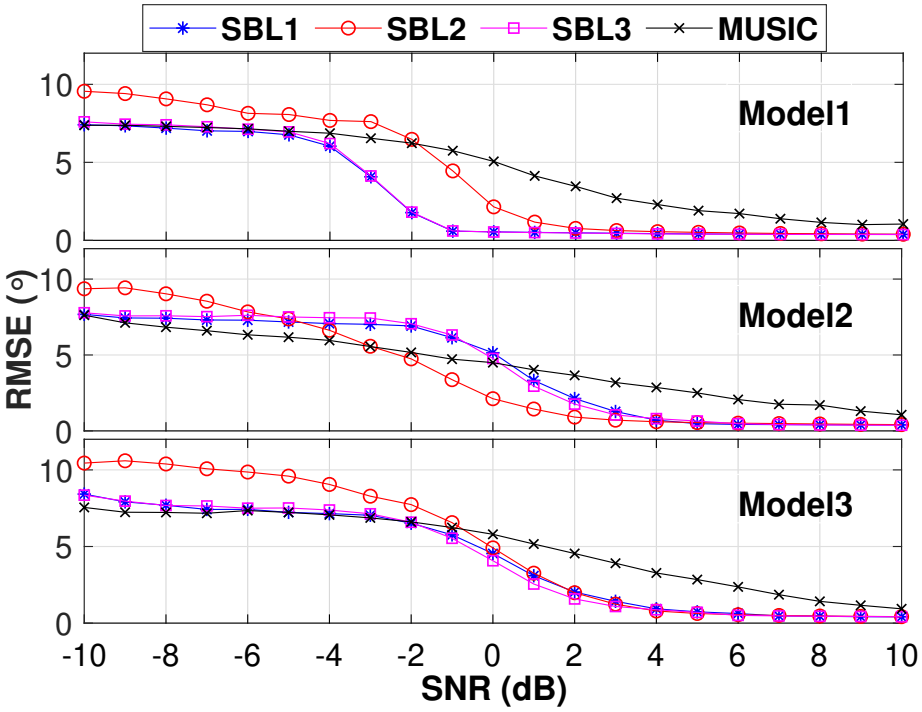

Figure 3.7 Performance analysis of wideband signal models and algorithms for different SNR values

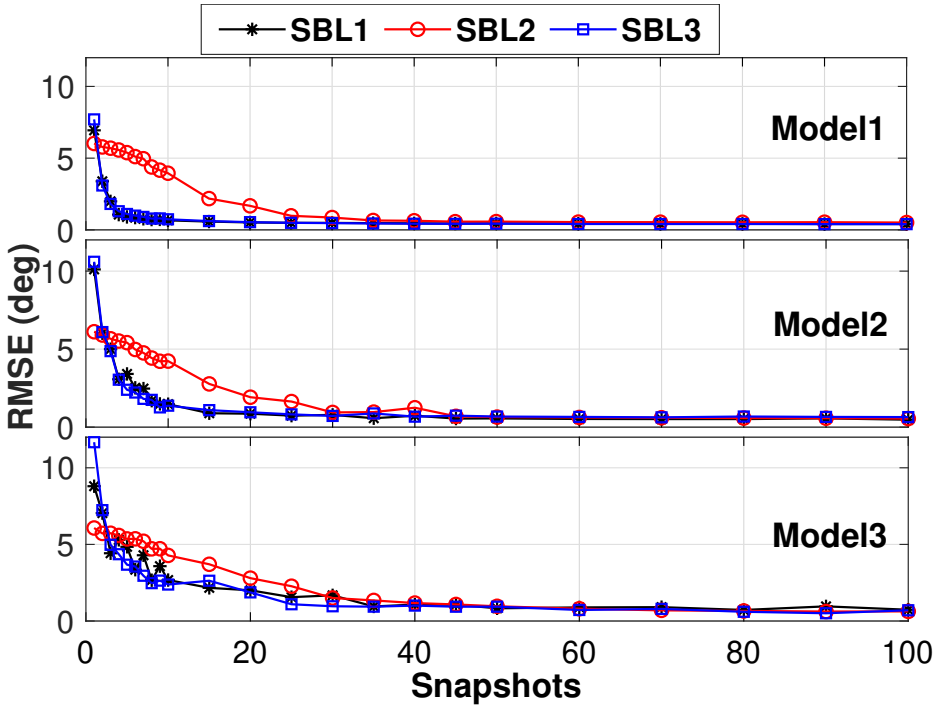

Figure 3.8 Performance analysis of wideband signal models and algorithms for different snapshots

and error approaches 0 as SNR increases. In this range, MUSIC has a higher error than SBL. At lower SNR, the SBL1, SBL3, and MUSIC show similar errors, while SBL2 has higher errors. For intermediate SNR in Model2, SBL2 works best.

We can see from Fig. 3.8 that for all three models, the RMSE for SBL2 is always higher than for SBL1 and SBL3. The performance of SBL1 and SBL3 is similar for Model1, whereas SBL3 performs better than SBL1 for Model2 and Model3 for lower values of snapshots. Fig. 3.9 shows the spectrums using simulated data with 100 snapshots for two sources located at 1° and 12.5° at 15 dB SNR. CBF has a poor resolution as only one peak is obtained for all the models. SBL1 and SBL3 localize both sources, and the spectrum is sparse. MUSIC and SBL2 also localize both sources but have relatively less sharp peaks.

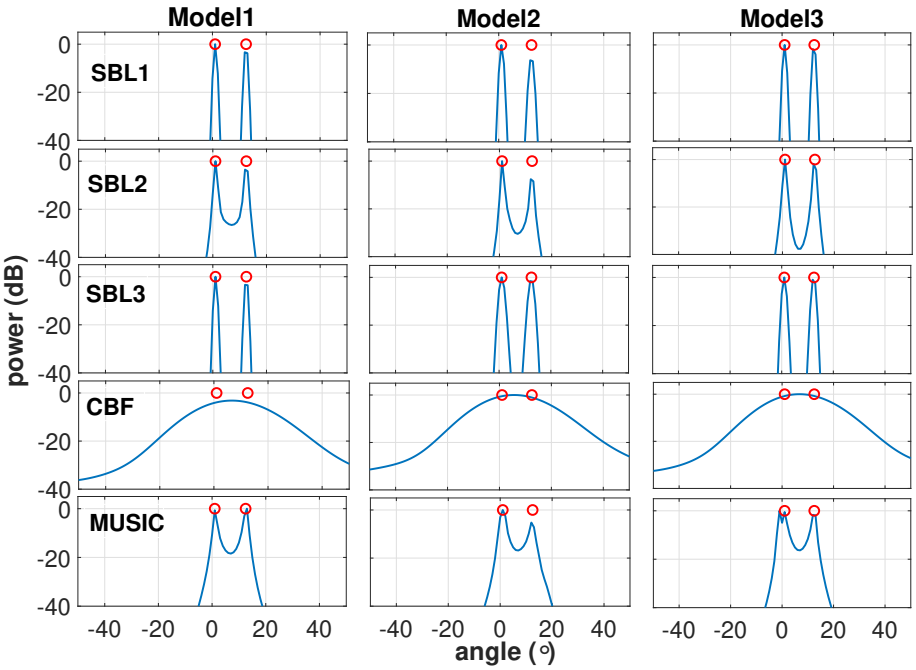

Figure 3.9 Spectrum of two sources at $[1,\ 12.5]^\circ$ from various algorithms for wideband signal models

3.3.6 Results for LOCATA dataset

The LOCATA [12] development dataset is used for the performance comparison of algorithms. We consider the recordings from **robothead**, **dicit**, and **dummy** arrays. We perform azimuth localization using a predefined DOA grid from -180° to 180° with a spacing of 1°. We note that Model2 and/or Model3 from Section 3.3.1 better characterize the speech signals in this dataset. We follow the LOCATA processing as discussed earlier, and the error analysis is done using the computed estimates and the available ground truth. The error between estimated and true DOAs is computed at block level during voice activity periods [98]. We use frequencies between 300–2500 Hz ($F = 47$ equispaced frequencies are chosen). Table 3.2 and 3.3 report RMSE errors and detection probability for various tasks and arrays.

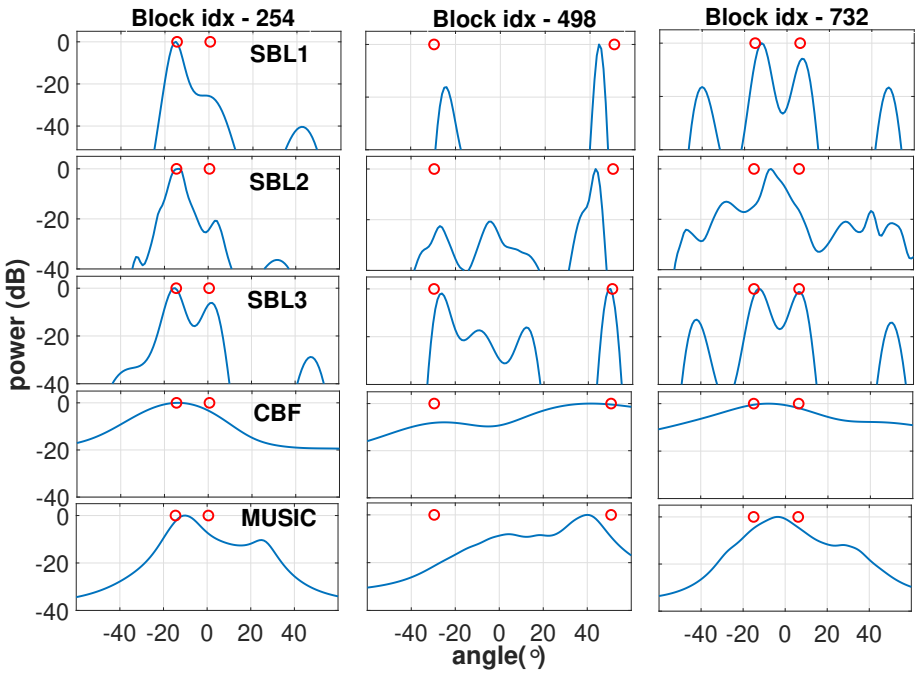

Figure 3.10 Spatial spectrum for few selected blocks of recording-1 in Task 4 using dicit array (LOCATA)

Dicit: A uniform linear subarray of 7 sensors with 32 cm spacing is considered from a 15-sensor dicit array. For dicit, the angular grid is considered from $-90°$ to $90°$. We consider the recordings from Tasks 2, 4, and 6 with multiple sources. The robustness of SBL3 can be seen from Fig. 3.9 and Fig. 3.10 shows the spectrums obtained from all algorithms for both simulated data and LOCATA data. Fig. 3.10 shows the spectrum obtained from different blocks in Task 4, recording-1 using a dicit array. These selected blocks are representative of observations throughout the recording. CBF and MUSIC show poor performance compared to SBL1 and SBL2, which localize both sources in 2 out of 3 blocks. Multiple spurious peaks produced by SBL2 are visible. SBL3 localizes both sources in all the blocks and shows superior performance compared to other algorithms. This is supported by RME, and detection probability (PD) averaged across recordings of Task 2, 4, and 6 as shown in Table 3.2 and 3.3. This shows that for real-world signals, SBL3 is more effective.

Dummy: The algorithms are compared on all tasks (1–6) foa r dummy with 4 sensors. Fig. 3.11 shows the DOA estimates for Task 5, recording 2 of the hearing aid. Task 5 consists of recording from a moving source when both the source and array are in motion. The shaded region represents the active voice period. It can be seen that for SBL3, the estimates are closer to the ground truth. The RMSE and PD are reported in Table 3.2 and 3.3. SBL3 consistently outperforms all other algorithms. For Task 2, SBL1 has low RMSE, but it also has low detection probability (see 3.3). **Robothead:** We report the azimuth error for Task 2, 4, and 6 for the localization of multiple static and moving sources in Table 3.2. Fig. 3.12 shows the DOA estimates of Task 6 recording-2 from block index 45 to 135 when both array

Array	Task	Localization algorithm				
		CBF	**MUSIC**	**SBL1**	**SBL2**	**SBL3**
Dummy	1	8.6	6.8	2.9	5.1	**1**
	2	11.4	8.4	**3.2**	8.8	7.4
	3	9.1	12.2	6.9	10.5	**5.8**
	4	14.6	14.4	13.2	**13.1**	17.1
	5	11.1	13.7	9.7	13.1	**6**
	6	**15.3**	15.8	16.1	15.8	18
Dicit	2	17.2	15	13.1	17.9	**12**
	4	15.2	12.8	12.8	12.5	**11.2**
	6	16	12.7	**9**	10.2	**9**
Robothead	2	10.3	**10**	12.2	11.5	12.8
	4	19.2	18.8	17	16.3	**16**
	6	16.5	16.1	**15.3**	16.7	**15.3**

Table 3.2 RMSE (in °) using dummy, dicit and robothead (averaged over all recordings).

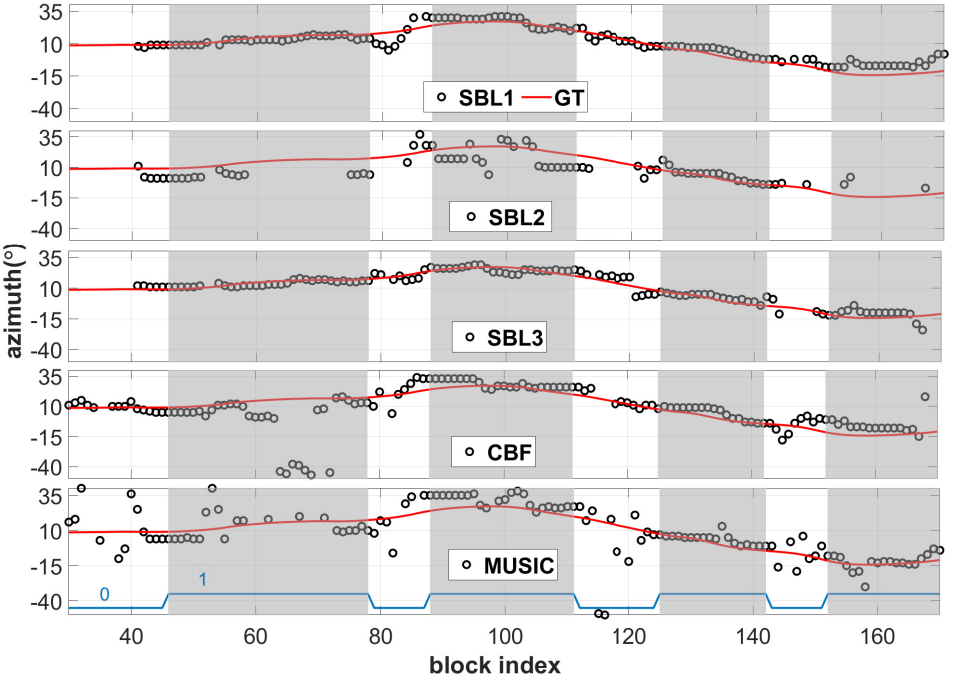

Figure 3.11 The estimated DOAs and ground truth (GT), Task 5, recording 2, dummy array (LOCATA)

and sources are in motion. It can be observed that both SBL1 and SBL3 give comparable estimates for the shown part of the recording. For blocks 90 to 110, SBL3 estimates both sources accurately; however, it estimates only one source accurately from blocks 120 to 130, as only one peak is obtained in the spectrum. Similarly, SBL1 estimates both sources accurately from blocks 120 to 130 but estimates one source only from blocks 90 to 110. The RMSE and PD reported in Table 3.2 and 3.3 show that SBL3 performs best when averaged over all recordings.

Array	Task	Localization algorithm				
		CBF	**MUSIC**	**SBL1**	**SBL2**	**SBL3**
Dummy	1	99	99	**100**	99	**100**
	2	**90**	40	29	67	85
	3	81	74	71	58	**85**
	4	41	31	46	50	**55**
	5	**100**	99	**100**	99	**100**
	6	16	18	38	33	**45**
Dicit	2	52	62	60	70	**72**
	4	32	36	76	68	**78**
	6	88	**92**	88	**92**	**92**
Robothead	2	80	85	86	**94**	91
	4	46	63	81	81	**87**
	6	45	44	44	38	**47**

Table 3.3 Probability of detection (in %) using dummy, dic,it and robothead (averaged over all recordings).

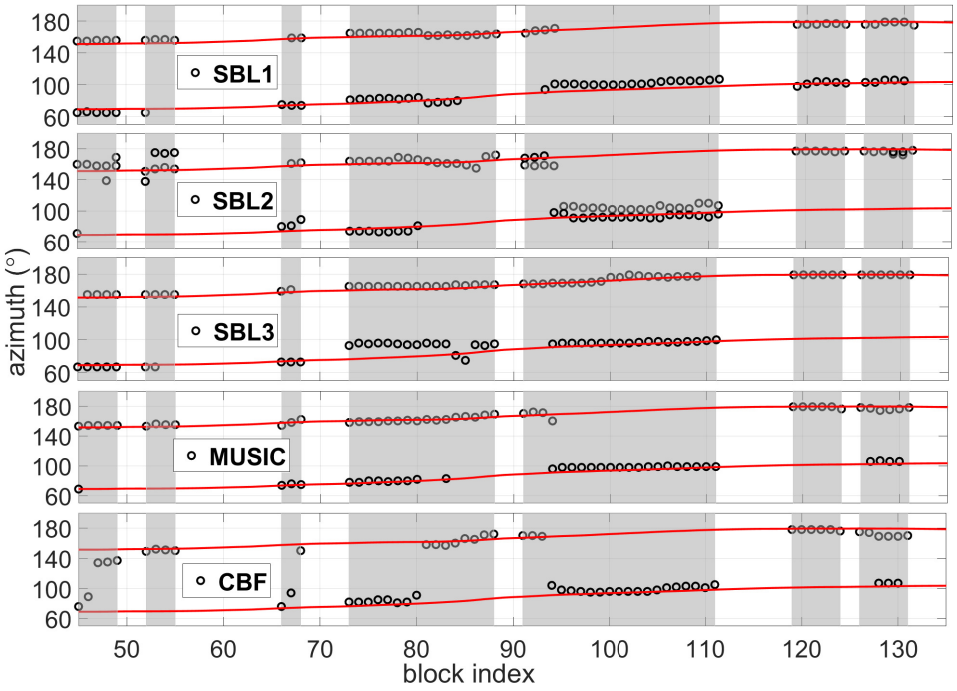

Figure 3.12 The estimated DOAs and ground truth (GT) on LOCATA dataset, Task 6, recording 2 using robothead array

3.4 Improving DOA estimation accuracy via derivative prediction

Traditional DOA estimation methods rely on the analytical properties of array signals and do not generalize well to non-ideabehaviorsrs such as multi-path signal propagation and uncertain noise characteristics. Data-driven methods provide a new way to learn non-ideal behaviors by using general func-

tion approximators such as neural networks and show promising results for source localization tasks. In real-world scenarios, sudden large changes in the DOA are unexpected. To capture this, we train our network to learn and predict DOA derivatives to maintain temporal continuity. DOA derivatives provide information about the rate of change in the x, y, z positions. Thus, the combined DOA and DOA derivative prediction provides a mechanism to suppress sudden DOA changes (if predicted by the network) and estimate realistic, smooth motion trajectories. In this study, we focus on improving the localization accuracy of an existing model [74] by predicting both the DOAs and their derivatives, i.e., changes in the x, y, z positions over time. We compare the existing localization models (considering the detection ground truth to be known), which predicts only DOAs, with the model, which predicts both DOAs and their derivatives. Our experiments reveal that DOA estimation can be a challenging task, even for immobile sound sources. This research aims to understand better deep learning-based models for DOA estimation tasks, focusing on combining DOAs with their derivatives to improve the source trajectories. This work focuses on the effectiveness of incorporating the derivative information to obtain smoother trajectories.

3.4.1 Model architecture

3.4.1.1 Features

The SALSA-Lite was introduced as an efficient computational version of the Spatial Cue-Augmented Log-Spectrogram (SALSA) feature for MIC (audio format) data [73, 74]. For M-channel audio recording, SALSA-Lite is a $(2M-1)$ channel feature consisting of M log-power spectrogram with $(M-1)$ frequency-normalized interchannel phase differences (NIPDs). The NIPD (Λ) approximating the relative distance of arrival (RDOA) can be written as

$$\Lambda(t,f) \approx -\frac{c}{2\pi f} \arg |\boldsymbol{H}_1^*(t,f)\boldsymbol{H}_{2:M}(t,f)| \approx [d_{12}(t) \dots d_{1M}(t)], \tag{3.16}$$

where $\boldsymbol{H}_m(t,f) = e^{\frac{j2\pi f d_{1m}(t)}{c}}$ is the array response for any arbitrary array structure under the far-field assumption and $d_{1m}(t)$ is the RDOA between the first (reference) and m^{th} mic. The SALSA-Lite provides the exact time-frequency positioning between the spectrogram and the NIPD, resulting in the model being able to localize multiple overlapping sources.

3.4.1.2 Architecture

Figure 3.13 shows the neural network architecture designed to simultaneously predict the DOAs and their derivatives. Here, the derivative represents the change in DOAs between two time frames. The SALSA-lite is fed to the CRNN network, which consists of one convolutional layer, one average pooling layer followed by four ResNet22 blocks [103] in the network body [73, 74]. The output othe f ResNet block is fed into a two-layer bidirectional Gated Recurrent Unit (GRU) followed by two distinct regression heads for predicting DOAs and their derivatives in Cartesian coordinates ($x, y,$ and z), respectively. Unlike the SELD network in [74], we focus only on the DOA estimation task and replace the

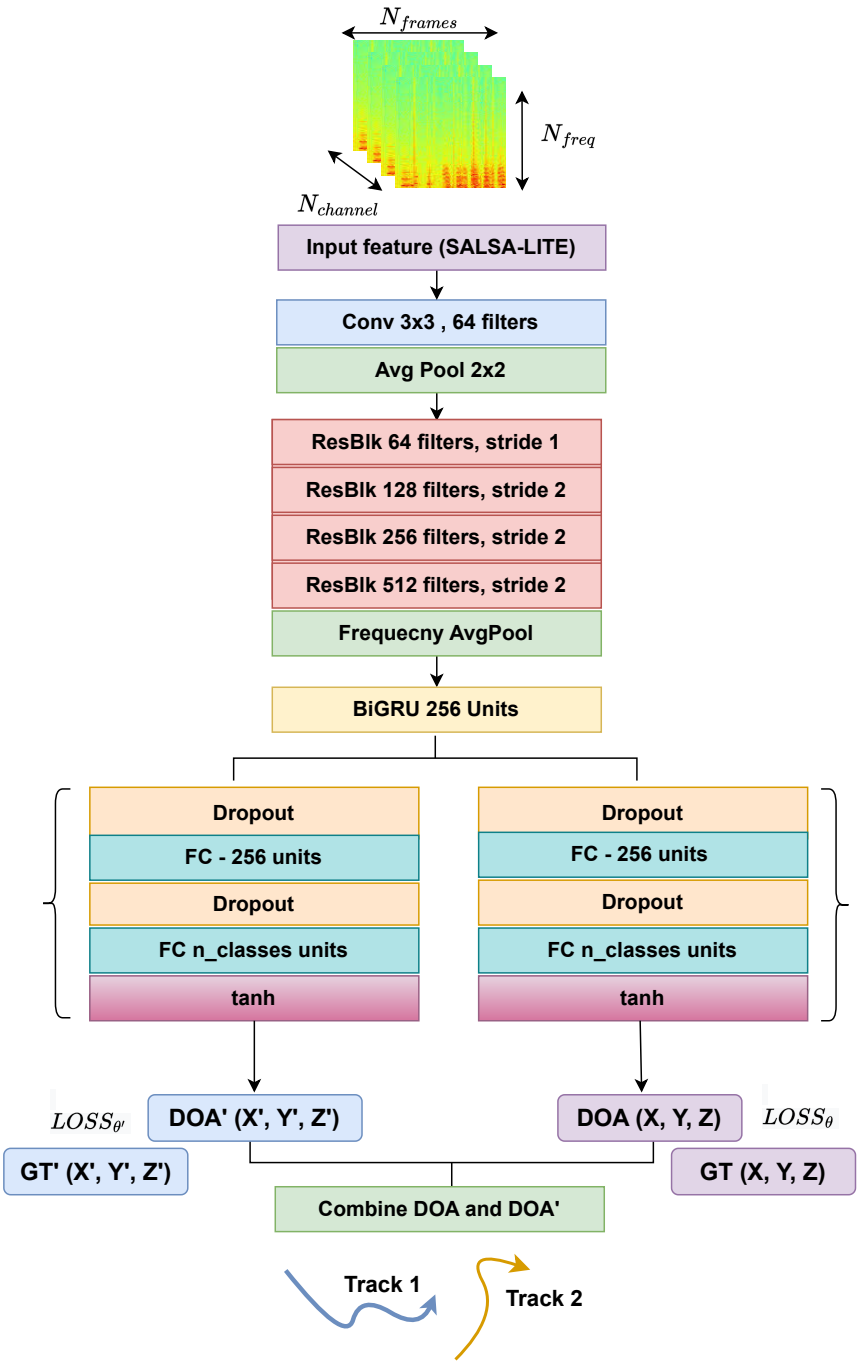

Figure 3.13 Model architecture predicting both DOAs and DOA derivatives

detection head with the regression head, which predicts the derivatives of DOAs (x', y', z') at different timestamps as shown in Fig. 3.13. Along with predicting the intermediate DOAs, the additional derivative information helps to obtain better overall DOA estimates. Once the network predicts the DOAs and

their derivatives, the final DOAs are obtained using the following update equation

$$\hat{x}_n^{\text{final}} = \frac{\hat{x}_n + (\hat{x}_{n-1} + \hat{x}'_n)}{2}, \quad n = 0, 1, \ldots, N-1 \tag{3.17}$$

where $\hat{x}'_n$ is prediction of the DOA derivative in x position (i.e., $x_n - x_{n-1}$) at n^{th} time and $\hat{x}'_0 = 0$ is the first derivative assumed to be zero. Similarly, the update rule for $\hat{y}_n^{\text{final}}$ and $\hat{z}_n^{\text{final}}$ can be obtained. By additionally incorporating derivative predictions, we expect the DOA of moving targets to be estimated more accurately. The number of active sources is assumed to be known for both the regression heads (predicting DOAs and derivatives). The ground truth is used to compute the losses for both prediction heads. The mean squared error (MSE) loss is minimized while training both the network heads and can be written as

$$\text{LOSS}_x = w_1 \sum_{n=1}^{N} (x_n - \hat{x_n})^2 + w_2 \sum_{n=1}^{N} (x'_n - \hat{x_n}')^2, \tag{3.18}$$

$$\text{LOSS}_{\text{total}} = \text{LOSS}_x + \text{LOSS}_y + \text{LOSS}_z, \tag{3.19}$$

where x_n and x'_n are ground truth DOAs and their derivatives at n^{th} timestamps, and $\hat{x_n}$ and $\hat{x_n}'$ represent the predicted DOAs and the predicted derivatives at n^{th} timestamps respectively. In (3.18), LOSS_x represents loss computed for x positions and similarly LOSS_y and LOSS_z can also be computed. The total loss minimized by the network is given in (3.19). Note that using appropriate weights for DOA loss and its derivatives loss is important; we use equal weights for both DOA and derivatives loss ($w_1 = w_2 = 0.5$). However the loss weights (w_1 and w_2) can be automatically optimized as described in [104] and can be incorporated in (3.18).

3.4.2 Simulation results

3.4.2.1 Dataset and training

The TAU-NIGENS Spatial Sound Events 2021 dataset has been employed in this work to analyze the performance of the proposed models. The dataset comprises 600 recordings, each one minute long and with four channels, and is in the MIC data format with a sampling frequency of 24 kHz. The dataset includes a diverse range of sound events featuring both stationary and mobile sources from 12 distinct classes. For this study, 400 recordings are used for training, while 100 recordings are allocated to both validation and testing. The angular range for azimuth and elevation angles are $[-180,\ 180)^\circ$ and $[-45,\ 45]^\circ$, respectively. For feature extraction, we followed the same setup as in [74], and frequencies from 50 Hz to 2 kHz are processed to avoid the aliasing. While training, Adam optimizer [105] is used, with the initial learning rate as 3×10^{-4} which linearly decreases to 10^{-4} over the last 15 epochs. A total of 70 epochs with 32 batch size are processed. The validation set is used for model selection, whereas the test data is used for the performance analysis.

The DCASE data provides ground truth DOA labels every 0.1 seconds. While generating the ground truth for DOA derivatives, the derivatives are considered as the difference between the previous and current DOA. As a significant gap in DOA update time would result in an unreliable derivative estimate, the derivative is considered zero if the source appears for the first time or a missing source reappears after more than 20 frames ($\approx$ 2 sec). We believe this is a reasonable choice since human motion patterns do not change abruptly within short time spans.

3.4.2.2 Accuracy metric

Following the framework of our baseline method [74], we use the detection ground truth to compute the error only for the frames where the sources are present/active. In this study, DOA/spatial error $\Delta\sigma$ is computed as the angular distance between the predicted and true positions, [54, 106].

$$\Delta\sigma = \arccos\left(\mathbf{n}_{\text{true}}.\mathbf{n}_{\text{pred}}\right).\frac{180}{\pi}, \tag{3.20}$$

where $\mathbf{n}_{\text{true}}$ and $\mathbf{n}_{\text{pred}}$ are unit norm vectors corresponding to the true and predicted positions, respectively. In applications requiring indoor audio source localization, knowing the angular distance of a source is more important than its Euclidean distance [12, 107].

Following the DCASE challenge convention [108], a source is considered as localized only when the DOA error (averaged across the active frames within a block) is less than 20°; we call these cases as true positive (TP). The false negative (FN) counts the number of incidences when the averaged spatial distance is more than 20°. The evaluation criteria and threshold of 20° is used from the original DCASE challenge TP criteria [51,54,106,108]. We report the probability of detection (P_{d} in %) as the fraction of frames where the distance between the predicted and true positions is less than 20°. Since the network is designed to provide only one prediction per class, the error was not calculated for the cases where multiple sources were present from the same class in the frame.

3.4.2.3 Effect of combining derivative

This subsection demonstrates the effect of combining the predicted derivatives with predicted DOAs. Fig. 3.14 shows the predicted source trajectories from both Model-1 (with derivative estimation) and Model2 (without derivative estimation) along with class-wise mean absolute error (MAE) computed for the correctly detected sources for one of the recordings. The cross ($\times$) in MAE plot denotes the cases when the models have not detected the source. It can be seen that Model-1 detects more sources, hence resulting in higher MAE for some classes compared to Model-2. The final DOA exhibits a smoother trajectory since the outliers are eliminated by combining the derivatives via the proposed update rule (3.17). The update rule is helpful even for static sources. We observed that Model-2 gives more erroneous DOAs for static sources than Model-1. Overall, Model-1's estimates are closer to true trajectories, resulting in higher P_{d}. For this recording, the average P_{d} for static and moving sources for

Model-1 and Model-2 are, $P_{ds} = 64\%$, $P_{dm} = 78\%$, $P_{ds} = 53\%$, and $P_{dm} = 52.4\%$, respectively. The total P_d averaged over 100 recordings from the test data is reported in Table 3.4.

From Table 3.4, it is evident that both Model-1 and Model-2 show similar performance for the clean dataset. We observed that Model-1's performance degrades when the network predicts the erroneous DOAs; hence combining them with equal weights leads to incorrect estimates. As a correction step, the current DOA prediction with derivative and the past prediction can be weighted depending on the threshold. A choice must be made depending on confidence in the present and past predictions.

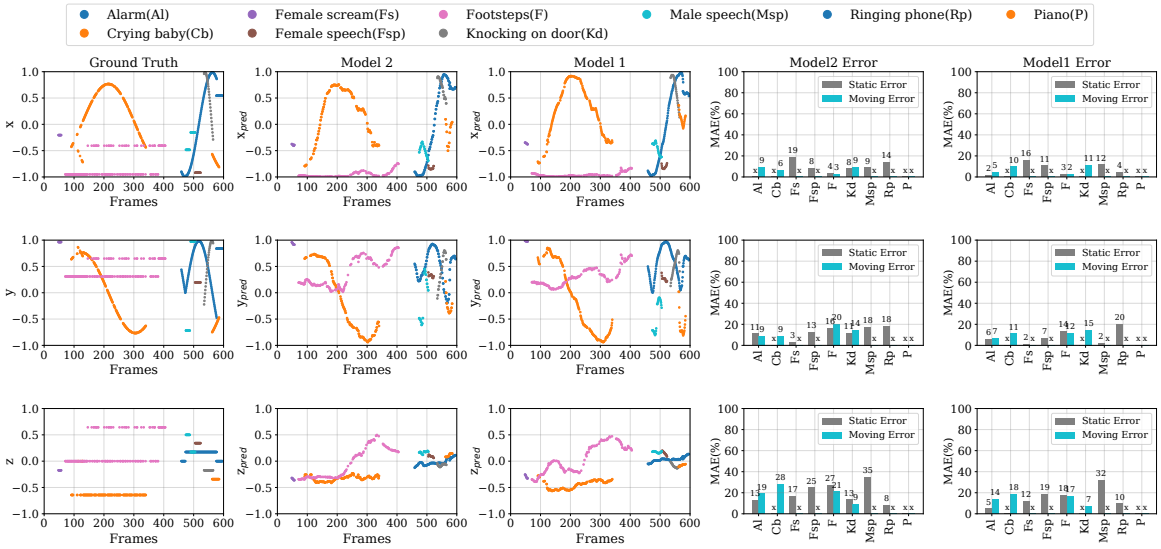

Figure 3.14 Effect of derivatives: true and predicted trajectories from Model-1 and Model-2 with class-wise MAE.

SNR	Model	TP_s	TP_m	FN_s	FN_m	P_{ds}	P_{dm}
Clean	Model-1	28843	21523	13056	9892	68.8	68.5
	Model2	27637	21389	14262	10026	65.9	68
-2dB	Model-1	17169	13097	24730	18318	40.9	41.7
	Model2	17199	12487	24700	18928	41	39.7
-5dB	Model-1	15812	10919	26087	20496	37.7	34.7
	Model2	14136	10522	27763	20893	33.7	33.4

Table 3.4 Performance of Model-1 and Model-2 at different SNR (averaged over test data).

3.4.2.4 Effect of transfer learning

To speed up the training process and reduce the risk of overfitting, we repeat the experiments using the pre-trained CRNN weights from an existing SELD model using SALSA-Lite, where the best model is obtained at the 47^{th} epoch [74]. The dataset, architecture, and framework for the pre-trained model

detailed in [74] is the same as the CRNN body used in this work. Keeping the CRNN body's weights fixed using the pre-trained SELD model increases the overall P_{d} is increased by 10 % for both Model-1 and Model-2, as shown in Table 3.5. From Fig. 3.15, it can be seen that Model-1 outperforms Model2 with higher P_{d} and lower DOA error.

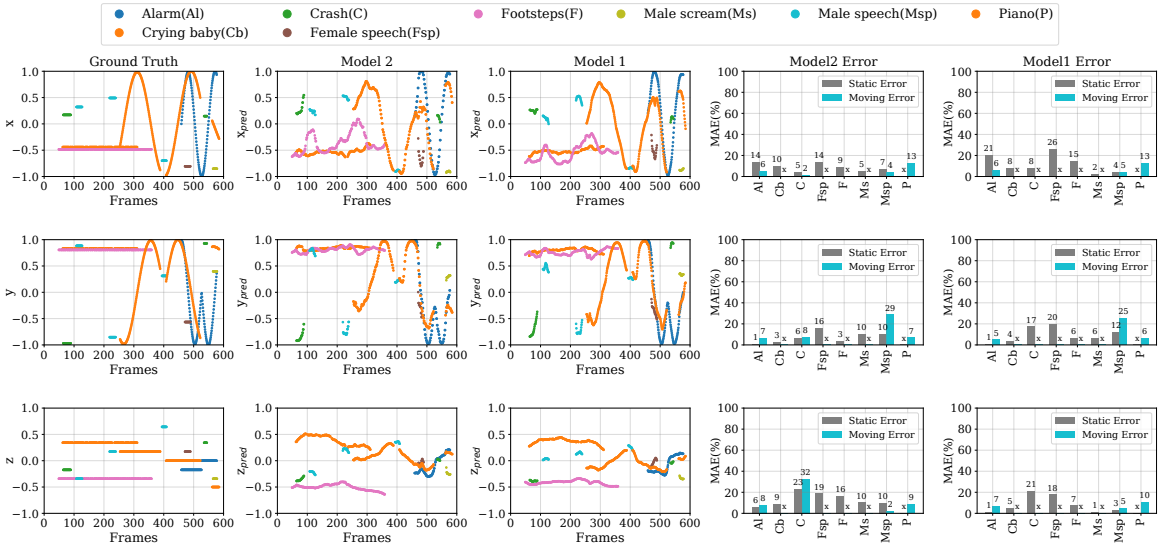

Figure 3.15 Effect of transfer learning: true and predicted trajectories from Model-1 and Model-2 with classwise MAE.

SNR	Model	$\mathbf{TP}_{\mathrm{s}}$	$\mathbf{TP}_{\mathrm{m}}$	$\mathbf{FN}_{\mathrm{s}}$	$\mathbf{FN}_{\mathrm{m}}$	$\mathbf{P}_{\mathrm{ds}}$	$\mathbf{P}_{\mathrm{dm}}$
Clean	Model-1	33033	24687	8866	6728	78.8	78.5
	Model-2	33431	24531	8468	6884	79.7	78
-2dB	Model-1	18349	12249	23550	19166	43.7	39
	Model-2	17777	11906	24122	19509	42.4	37.8
-5dB	Model-1	15365	10060	26534	21355	36.7	32
	Model-2	15404	9816	26495	21599	36.7	31.2

Table 3.5 Performance of Model-1 and Model-2 using the pretrained CRNN SELD model at different SNR (averaged over test data).

3.4.3 Effect of low SNR levels

In order to assess the robustness of the models, we introduced synthetic additive white Gaussian noise to the recordings from TAU-NIGENS Spatial Sound Events 2021 dataset despite the presence of unknown interference and noise in the dataset. The results from Table 3.4 indicate a significant degradation in the P_{d} of both models as the SNR level decreases. Nevertheless, Model-1 exhibits superior performance in noisy scenarios due to the improved final DOAs resulting from estimated derivatives.

The impact of SNR level on the source trajectories obtained from both models is presented in Fig. 3.16. Our analysis suggests that Model-1 provides more reliable estimates than Model-2.

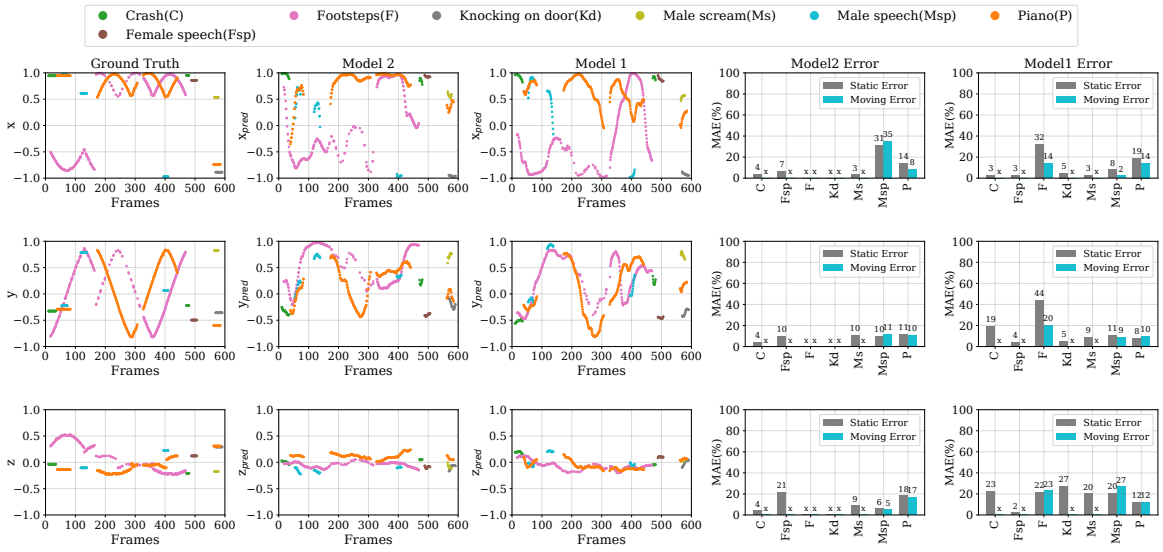

Figure 3.16 Effect of low SNR levels: true and predicted trajectories from Model-1 and Model-2 with classwise MAE.

3.5 Summary

In this chapter, we have considered DOA estimation as a compressive sensing problem and solved it using sparse Bayesian learning algorithm. We use three array structures to show DOA estimation results for different tasks of LOCATA dataset. The results show that CBF and MUSIC work well for a stationary or slow-moving source but are error-prone in challenging tasks where the source and/or array is moving. Multi-frequency SBL was observed to be robust to these challenges and performed well in all the tasks. The study shows that SBL significantly outperforms CBF and MUSIC on all tasks.

Later, we analyzed three wideband signal models and derived wideband SBL update rules for each of these models. We compared SBL methods with classical DOA estimation methods of CBF and MUSIC on recordings of LOCATA and simulated data. Simulations show that SBL3 performs best across all the signal models. SBL2 gives higher errors due to multiple false peaks. For hearing aid, SBL3 work best for LOCATA data processing. For dicit and robothead SBL1 and SBL3 show similar performance (Task 4 and 6). This study shows that Model3 is an effective signal model that accurately balances sparsity and power spectrum for real-world signals.

At last, we show the significance of predicting DOA derivatives in conjunction with DOAs (Model-1) for enhancing the overall DOA estimation performance compared to solely predicting DOAs (Model-2). Furthermore, we demonstrate that Model-1 is resilient to noise and performs better than Model-2

under low SNR conditions. Given the broad range of applications of SELD tasks, our analysis reveals that estimating DOAs and their derivatives cumulatively improves the source trajectories and overall performance. In the future, it would be intriguing to investigate the potential impact of incorporating higher-order derivatives in SELD tasks where detection and DOA estimation are simultaneously performed.

Chapter 4

Parametric Models and Algorithms for DOA Trajectory Localization

4.1 Introduction to Trajectory Localization

Traditional localization methods rely on block-level processing to extract the directional information from multiple measurements processed together [14, 23–25, 28, 49, 109, 110]. They estimate fixed DOA within a block followed by tracking filters on these block estimates. This works well for slow-moving targets but may not be ideal for fast motion. However, in real-world scenarios, the DOA is not constant across the snapshots, which can lead to limitations in the performance of localization algorithms. In [111], a sequential SBL algorithm was proposed to estimate time-varying DOAs, while in [49, 50], neural network-based methods were used to obtain trajectories directly. Despite these advancements, algorithms are still needed to accurately estimate DOA trajectories while being computationally efficient.

This chapter introduces a signal model incorporating source motion using parametric trajectories and accounts for DOA motion (linear or nonlinear) within the block duration. This provides better DOA estimates compared to models assuming fixed DOA. It can also be extended to other parametric models, albeit with a further increase in processing complexity. Parametric trajectories can potentially eliminate the need for tracking filters by implicitly performing both localization and tracking. We refer to this as trajectory localization (TL). In trajectory localization, source trajectories are estimated instead of point estimates. In this chapter,

- We developed two trajectory models to account for dynamic source DOA: (a) harmonic trajectory model and (b) polynomial trajectory model.
- We developed an extension of CBF called TL-CBF to perform parametric trajectory estimation.
- Reformulated the trajectory model in a sparse signal framework and developed TL-SBL and TL-OMP algorithms for trajectory localization.
- We developed two gridless algorithms to estimate the trajectory parameters: (a) SFW for trajectory localization (TL-SFW) and (b) NOMP for trajectory localization (TL-NOMP).
- We formulated multi-frequency signal models and developed extensions of TL-SFW and TL-NOMP to perform trajectory localization using multi-frequency signals.

- We performed a comprehensive performance analysis of the proposed signal models and algorithms to study the impact of signal-to-noise ratio (SNR), number of snapshots, resolution limits, grid step size, and computational complexity.
- We experimentally validate proposed algorithms on real-world recording from the LOCATA dataset.

4.2 Signal model

This section briefly overviews the static DOA model for ease of reading and discusses the proposed parametric trajectory model. Polynomial and harmonic trajectory models are used to model complex trajectories. The linear trajectory model, presented in our earlier work [13, 112], can be recovered as a particular case of the polynomial model.

4.2.1 Static DOA Model

In this subsection, the DOA is assumed to be constant within a block. Let $\mathbf{y} \in \mathbb{C}^N$ be the measurement vector received from an $N-$sensor uniform linear array (ULA), when K sources are present:

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{a}(\theta_k) x_k + \mathbf{n} = \mathbf{A}_{\text{sv}}\mathbf{x} + \mathbf{n} \tag{4.1}$$

where $\mathbf{A}_{\text{sv}} = [\mathbf{a}(\theta_1) \dots \mathbf{a}(\theta_K)]$ is a matrix whose columns are steering vectors where $\mathbf{a}(\theta_k)$ is steering vector corresponding to the source direction θ_k and $k = 1, \dots, K$. $\mathbf{x} = [x_1, \dots, x_K]$ is the source amplitude vector and $\mathbf{n} \in \mathbb{C}^N$ is the additive noise.

When a sequence of L observations is available, the above narrowband model can be extended to multiple measurement vector (MMV) model [41, 44] as:

$$\mathbf{Y} = \mathbf{A}_{\text{sv}}\mathbf{X} + \mathbf{N} = [\mathbf{A}_{\text{sv}}\mathbf{x}_1 \dots \mathbf{A}_{\text{sv}}\mathbf{x}_L] + \mathbf{N} \tag{4.2}$$

where $\mathbf{Y} = [\mathbf{y}_1 \dots \ \mathbf{y}_L] \in \mathbb{C}^{N \times L}$ is the L snapshot observation matrix, $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_L] \in \mathbb{C}^{K \times L}$ represents the source amplitudes of K sources over $L > 1$ snapshots, and $\mathbf{N} = [\mathbf{n}_1 \dots \mathbf{n}_L] \in \mathbb{C}^{N \times L}$ accounts for the additive noise across L snapshots. Under static DOA assumption, the source directions (θ_k) do not change with time and are determined by analyzing the block of L snapshots.

4.2.2 Parametric models for DOA trajectory

In practical situations, sources are often in motion, making the assumption of constant DOA impractical. This presents a challenge in accurately estimating the DOA for moving sources. To overcome this issue, we modelled and estimated linear DOA trajectories within block duration. However, the linear assumption does not always hold true, as sources can exhibit complex, nonlinear motion. To address

this limitation, we introduce two general trajectory models that capture linear and nonlinear motion – polynomial and harmonic trajectories. In this work, we consider polynomial and harmonic trajectory models, however any parametric model can be used to capture the linear/nonlinear motion of DOAs within the blocks.

4.2.3 Polynomial model

Signals with constant/changing amplitude and polynomial phase are useful in many applications [113–116]. For example, chirp signals representing second-order phase polynomials, quadratic FM signals corresponding to third-order phase polynomials, and radar returns from targets with constant acceleration featuring second-order phase terms [113, 115]. The work in [117] derived the Cramer-Rao lower bound on the variance of estimated parameters of signals with constant amplitude and polynomial phase. In [118], a parameter estimation study is performed for non-stationary signals with time-varying amplitudes and polynomial phase. The work in [119] explores the estimation of multicomponent polynomial-phase signals impinging on a multi-sensor array, leveraging state-space modeling. In [120], Cramer–Rao bounds and maximum likelihood estimation for random amplitude phase-modulated signal are studied. Addressing the estimation challenge for signals comprising one or more components, [121] employs a linear parametric model representing amplitude and phase functions. The derived Cramer-Rao bound emphasizes the independent estimation of amplitude and phase parameters. The spatial time-frequency distribution concept has been developed and employed in [122] to localize spatial sources, where the PPS-like sources have been given great importance. In [123], the Extended Kalman filter is applied to estimate the multi-component polynomial phase system. Exploiting spatial information given by a multi-sensor array is shown to provide a high convergence rate. In the literature, RF signal impinging on a single sensor whose phase is modulated (not linear) typically as a polynomial function, whereas the proposed polynomial trajectory model characterizes the DOA at each snapshot, and when the DOA is changing, the phase difference along the sensor array also changes.

We define a p^{th} order polynomial trajectory as a function of snapshot number as

$$\theta^l = \phi + \sum_{p=1}^{P} \alpha_p \left(\frac{l}{L-1} \right)^p, \quad l = 0, 1, \ldots, L-1 \tag{4.3}$$

where θ^l represents the DOA at l^{th} snapshot and $\boldsymbol{\omega} = (\phi, \alpha_1, \ldots, \alpha_p)$ denotes the vector of trajectory parameters for a source. The first order polynomial $(p = 1)$ corresponds to the linear trajectory model,

$$\theta^l = \phi + \alpha_1 \left(\frac{l}{L-1} \right), \quad l = 0, 1, \ldots, L-1, \tag{4.4}$$

whereas the zeroth order polynomial $(p = 0)$ corresponds to the static DOA case. Note that increasing the number of parameters in the model allows for complex trajectories but, at the same time, leads to higher computations in the trajectory estimation algorithms.

4.2.4 Harmonic trajectory model

Alternate to the polynomial model, we can use the harmonic trajectory model as discussed in [124] to generate trajectories,

$$\theta^l = \phi + \sum_{q=1}^{Q} \{\alpha_q \sin q\nu l + \beta_q \cos q\nu l\} \tag{4.5}$$

where ν denotes the fundamental frequency of sinusoidal signals to be added and $\boldsymbol{\omega} = (\phi, \alpha_1, \ldots, \alpha_Q, \beta_1, \ldots, \beta_Q)$ denotes the vector of trajectory parameters for a source. These trajectories are guaranteed to be bandlimited, with the maximum frequency being $Q\nu$. We choose Q based on expected DOA changes within a block. As in the case of polynomial trajectories, increasing Q increases the computational cost of trajectory estimation algorithms.

4.2.5 Observation model

Let $\boldsymbol{\omega}_k \in \boldsymbol{\Psi}$ be the vector of parameters defining the k^{th} source DOA trajectory, where $\boldsymbol{\Psi}$ is the continuous trajectory space. Define $\tilde{\mathbf{A}}(\boldsymbol{\omega}_k) \in \mathbb{C}^{N\times L}$ to be the trajectory steering matrix containing all the steering vectors as the DOA varies for the k^{th} trajectory, i.e., $\tilde{\mathbf{A}}(\boldsymbol{\omega}_k) = \left[\mathbf{a}(\theta_k^1) \ldots \mathbf{a}(\theta_k^L))\right] = \left[\mathbf{a}_1^k \ldots \mathbf{a}_L^k\right]$, where θ^l represents the l^{th} snapshot DOA in an L-length block. Let $\tilde{\mathbf{X}}_k = \text{diag}(\mathbf{x}^k), \mathbf{x}^k = [x_k^1 \ldots x_k^L]^T$ be the diagonal matrix of L complex amplitudes for the k^{th} source. Thus, the MMV observation matrix when K sources are present can be expressed as,

$$\mathbf{Y} = \sum_{k=1}^{K} \tilde{\mathbf{A}}(\boldsymbol{\omega}_k)\tilde{\mathbf{X}}_k + \mathbf{N} = \sum_{k=1}^{K} \tilde{\mathbf{A}}_k\tilde{\mathbf{X}}_k + \mathbf{N}\,, \tag{4.6}$$

$$= \bar{\mathbf{A}}(\mathcal{W})\bar{\mathbf{X}} + \mathbf{N} = \bar{\mathbf{A}}\bar{\mathbf{X}} + \mathbf{N}\,, \tag{4.7}$$

where $\bar{\mathbf{X}} = [\tilde{\mathbf{X}}_1 \ldots \tilde{\mathbf{X}}_K]^T$, $\bar{\mathbf{A}}(\mathcal{W}) = [\tilde{\mathbf{A}}_1 \ldots \tilde{\mathbf{A}}_K]$, and $\mathcal{W} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K\} \subset \boldsymbol{\Psi}$. Here $\bar{\mathbf{X}}$ consists of K diagonal matrices (of size $L \times L$) stacked vertically. Let $\mathcal{X}_K^L$ be the set of all such vertically stacked diagonal matrices, thus $\bar{\mathbf{X}} \in \mathcal{X}_K^L$.

In contrast to the static DOA MMV model (4.2), (4.7) represents the dynamic DOA MMV model, which accounts for source motion through $\bar{\mathbf{A}}(\mathcal{W}$. In trajectory localization, our aim is to estimate parameters $\boldsymbol{\omega}_k$ defining the trajectory for all $\boldsymbol{\omega} \in \mathcal{W}$ the sources from the given observation matrix.

4.2.6 Sparse model

The above model (4.7) can be reformulated as a sparse signal model, allowing us to apply sparse signal processing algorithms for DOA trajectory estimation. To demonstrate the sparse formulation, we consider the linear trajectory estimation where $\boldsymbol{\omega} = (\phi, \alpha)$. Let us consider a finely sampled grid $\boldsymbol{\Psi}_d \subset \boldsymbol{\Psi}$ in (ϕ, α) space. Let the uniformly sampled points in this trajectory space be denoted by $\boldsymbol{\Psi}_d = \{(\phi_1, \alpha_1), \ldots, (\phi_1, \alpha_{M_2}), \ldots, (\phi_{M_1}, \alpha_{M_2})\}$. A sparse model for (4.7) can be written as

$$\mathbf{Y} = \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \tilde{\mathbf{A}}(\phi_{m_1}, \alpha_{m_2}) \tilde{\mathbf{X}}_{m_1 m_2} + \mathbf{N} \tag{4.8}$$

$$= \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \tilde{\mathbf{A}}_{m_1,m_2} \tilde{\mathbf{X}}_{m_1,m_2} + \mathbf{N} \tag{4.9}$$

$$= \tilde{\mathbf{A}}_s \, \tilde{\mathbf{X}}_s + \mathbf{N} \,, \tag{4.10}$$

where $\tilde{\mathbf{A}}_{m_1,m_2} \triangleq \tilde{\mathbf{A}}(\phi_{m_1}, \alpha_{m_2})$ and $\tilde{\mathbf{X}}_{m_1,m_2}$ are the changing DOA steering vector matrix and source amplitude matrix for the source at $(\phi_{m_1}, \alpha_{m_2})$. Among all the potential $M_1 M_2$ source trajectories, only K trajectories are present in a given block and $K \ll M_1 M_2$. This sparsity is modelled by matrices $\tilde{\mathbf{X}}_{m_1,m_2}$, of which only K are non-zero. Here, we assume that the true sources lie on the grid. For compact expression we define, $\tilde{\mathbf{A}}_s = [\tilde{\mathbf{A}}_{1,1} \dots \tilde{\mathbf{A}}_{M_1,M_2}] \in \mathbb{C}^{N \times M_1 M_2 L}$, and $\tilde{\mathbf{X}}_s = [\tilde{\mathbf{X}}_{1,1} \dots \tilde{\mathbf{X}}_{M_1,M_2}]^T \in \mathbb{C}^{M_1 M_2 L \times L}$. The above MMV model can be equivalently written as a single measurement model (SMV) [125, 126] by vectorizing the observation matrix $\mathbf{Y}$ and appropriately changing the terms on the right-hand side. Performing a column-wise vectorization operation on $\mathbf{Y}$, we get

$$\text{vec}(\mathbf{Y}) = \mathbf{y}_v = \tilde{\mathbf{A}}_v \tilde{\mathbf{x}}_v + \mathbf{n}_v \tag{4.11}$$

$$\tilde{\mathbf{A}}_v = [\mathbf{I}_L \otimes \tilde{\mathbf{A}}_{1,1}, \dots, \mathbf{I}_L \otimes \tilde{\mathbf{A}}_{M_1,M_2}] \tag{4.12}$$

$$\tilde{\mathbf{x}}_v = [\text{diag}(\tilde{\mathbf{X}}_{1,1})^T, \dots, \text{diag}(\tilde{\mathbf{X}}_{M_1,M_2})^T]^T \tag{4.13}$$

$$\mathbf{n}_v = \text{vec}(\mathbf{N}) \,, \tag{4.14}$$

where $\mathbf{I}_L \otimes \tilde{\mathbf{A}}_{m_1,m_2} \in \mathbb{C}^{NL \times L}$ is the column-wise Kronecker product (Khatri–Rao product) of $\mathbf{I}_L$ and $\tilde{\mathbf{A}}_{m_1,m_2}$, and $\mathbf{I}_L$ is the $L \times L$ identity matrix. Here, the diag$(\cdot)$ operation on a square matrix returns the diagonal of the matrix as a column vector. The sparsity structure of matrix $\tilde{\mathbf{X}}_s$ is translated into a block sparse structure of the vector $\tilde{\mathbf{x}}_v$. In the next section, we adapt SBL [43] and OMP algorithms to signal model (4.11) giving trajectory localization SBL and OMP, i.e. TL-SBL and TL-OMP.

4.3 Grid-based algorithms for trajectory localization

Grid-based algorithms use a predefined grid where each grid point represents a possible trajectory parameter to be estimated. The algorithm then analyzes the array measurements to determine the most likely parameters by comparing the signal characteristics at different grid points. In this section, we discuss grid-based methods for trajectory localization. We briefly describe existing methods [112] of TL-CBF and TL-SBL and introduce an extension of orthogonal matching pursuit for the trajectory model called TL-OMP. We conclude this section by showcasing grid-based TL algorithms for linear trajectory estimation with $\boldsymbol{\omega} = (\phi, \alpha)$ as described in (4.4).

4.3.1 TL-CBF

A modification of the conventional beamforming (CBF) [24] algorithm for the linear trajectory model is presented in [112]. We refer to it as CBF for trajectory localization, i.e., TL-CBF. The original CBF algorithm computes the angular power spectrum at a predefined DOA grid by analyzing the correlation between the observations and the steering vectors [17]. The DOA estimates are determined from the peaks of this angular power spectrum. The TL-CBF extends this by computing the power spectrum using the following expression,

$$P_{\text{TL-CBF}}(\boldsymbol{\omega}) = \frac{1}{L}\sum_{l=1}^{L} |\mathbf{a}_l^H(\boldsymbol{\omega})\,\mathbf{y}_l|^2\,, \tag{4.15}$$

where the power spectrum $P_{\text{TL-CBF}}(\boldsymbol{\omega})$ is a scalar-valued function of the vector variable ω. The power is computed over a discrete trajectory space ($\boldsymbol{\omega} \in \boldsymbol{\Psi}_d$) with M potential grid points for $\boldsymbol{\omega}$. The locations of peaks in the spectrum are the estimated DOA trajectories. Figures 4.3 and 4.4 show the 2D and 3D view of the same TL-CBF spectrum (4.15), and the locations of the peaks provide trajectory parameters.

4.3.2 TL-SBL

A derivative of SBL, TL-SBL, has been developed and applied to estimate DOA trajectory parameters. Here, we derive the TL-SBL update rule following the approach in [45, 99–101]. The block sparse structure of $\tilde{\mathbf{x}}_v$ has similarities with the static DOA MMV model [125, 126].

Prior: We assume that source amplitudes are i.i.d. across snapshots having zero-mean complex Gaussian distribution

$$p(\text{diag}(\tilde{\mathbf{X}}_m)) \sim \mathcal{CN}(\mathbf{0}, \gamma_m \mathbf{I}_L)\,, \tag{4.16}$$

where γ_m is the variance, in this section, we use a simplified notation where the double index (m_1, m_2) is replaced by a single index m and correspondingly the indices $\{(1,1),\ldots,(M_1,M_2)\}$ are renumbered as $\{1,2,\ldots,M_1M_2\}$. We additionally assume the amplitudes are independent across sources. Thus unknown $\tilde{\mathbf{x}}_v$ is Gaussian distributed and parametrized by the vector $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_{M_1M_2}]$.

Likelihood: Assuming the noise to be zero-mean complex Gaussian distributed and i.i.d across sensors and snapshots, the data likelihood can be given as

$$p(\mathbf{y}_v|\tilde{\mathbf{x}}_v;\sigma^2) = \mathcal{CN}(\mathbf{y}_v;\tilde{\mathbf{A}}_v\tilde{\mathbf{x}}_v,\sigma^2\mathbf{I}_{NL})\,, \tag{4.17}$$

where σ^2 is the noise variance.

Evidence: In SBL, $\boldsymbol{\gamma}$ is estimated using evidence maximization (or type-II maximum likelihood) where evidence is

$$p(\mathbf{y}_v;\boldsymbol{\gamma}) = \int_{\tilde{\mathbf{x}}_v} p(\mathbf{y}_v|\tilde{\mathbf{x}}_v;\sigma^2)\, p(\tilde{\mathbf{x}}_v;\boldsymbol{\gamma})\, d\tilde{\mathbf{x}}_v\,. \tag{4.18}$$

Since both prior and likelihood are Gaussian, from properties of Gaussian densities, we get evidence $p(\mathbf{y}_v;\boldsymbol{\gamma})$ to be Gaussian with zero-mean and let the covariance matrix be $\boldsymbol{\Sigma}_{\mathbf{y}_v}$. The log evidence can thus be expressed as

$$\log p(\mathbf{y}_v;\boldsymbol{\gamma}) \propto \log|\boldsymbol{\Sigma}_{\mathbf{y}_v}| - \mathbf{y}_v^H \boldsymbol{\Sigma}_{\mathbf{y}_v}^{-1}\mathbf{y}_v\,, \tag{4.19}$$

$$\text{where} \quad \boldsymbol{\Sigma}_{\mathbf{y}_v} = \sigma^2\mathbf{I}_{NL} + \tilde{\mathbf{A}}_v\boldsymbol{\Sigma}_0\tilde{\mathbf{A}}_v^T\,, \tag{4.20}$$

$$\boldsymbol{\Sigma}_0 = \mathbf{E}(\tilde{\mathbf{x}}_v\tilde{\mathbf{x}}_v^H)\,. \tag{4.21}$$

Evidence maximization can be performed by expectation maximization (EM) algorithm [125–127], but its convergence is known to be slow [41, 43]. The TL-SBL method is based on a sparse modeling framework, and the update rule for computing the TL-SBL spectrum is given as

$$\hat{\gamma}_m^{\text{new}} = \hat{\gamma}_m^{\text{old}}\, \frac{\mathbf{y}_v^H\boldsymbol{\Sigma}_{\mathbf{y}_v}\hat{\mathbf{A}}_m\hat{\mathbf{A}}_m^H\boldsymbol{\Sigma}_{\mathbf{y}_v}^{-1}\mathbf{y}_v}{\text{Tr}[\boldsymbol{\Sigma}_{\mathbf{y}_v}^{-1}\hat{\mathbf{A}}_m\hat{\mathbf{A}}_m^H]}\,, \tag{4.22}$$

where $\hat{\mathbf{A}}_m = \mathbf{I}_L \otimes \tilde{\mathbf{A}}_m$, and Tr$[\cdot]$ denotes trace of a matrix. The m^{th} grid point represents a potential source ($\boldsymbol{\omega}_m$) with corresponding to the trajectory steering matrix $\tilde{\mathbf{A}}_m$. The vector $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_m]$ denotes the variance of source amplitude and, due to the hierarchical property of SBL, turned out to be sparse. The locations of non-zero entries of $\boldsymbol{\gamma}$ signify the source DOA trajectory estimates. An illustration of the TL-SBL spectrum (4.22) is shown in Fig. 4.5 and features well-defined peaks.

4.3.3 TL-OMP

To estimate the trajectory parameters, we modify the OMP algorithm [128, 129]. This greedy method iteratively selects the atoms from the dictionary on which the projection of the residual measurement matrix is maximum,

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}\in\boldsymbol{\Psi}_d}{\arg\max}\, \frac{1}{L}\sum_{l=1}^{L}\left|\mathbf{a}_l^H(\boldsymbol{\omega})\, \mathbf{r}_l^{[k-1]}\right|^2, \tag{4.23}$$

where $\mathbf{r}_l^{[k-1]}$ represents residual at l^{th} snapshot for k^{th} iteration. The residual for the next iteration is,

$$\mathbf{r}_l^{[k]} = \mathbf{r}_l^{[k-1]} - \mathbf{P}_l\, \mathbf{r}_l^{[k-1]}, \tag{4.24}$$

where $\mathbf{P}_l = \mathbf{a}_l \left(\mathbf{a}_l^H \mathbf{a}_l\right)^{-1} \mathbf{a}_l^H$ is the projection matrix and $\mathbf{a}_l^H \mathbf{a}_l = N$. This ensures that the residual observation vectors (at each of the l^{th} snapshots) are orthogonal to the corresponding steering vectors of the estimated source trajectories. The residual is initialized to the observation vector $\mathbf{r}_l^{[0]} = \mathbf{y}_l$. TL-OMP is a greedy algorithm that makes locally optimal choices at each step without considering the global impact, leading to suboptimal solutions. The TL-OMP spectrum (4.23) at various iterations are shown in Figure 4.6. A source is found at each iteration, and the residual is computed for next iteration.

4.3.4 Example

We compare the grid-based algorithms for linear DOA trajectories. Firstly, We compare the localization ability of TL-CBF and TL-SBL with traditional CBF and SBL and later show a performance comparison among various TL algorithms.

Example 1: In this experiment, a ULA with 10 sensors and $d = \frac{\lambda}{2}$ spacing is considered. TL-CBF and TL-SBL algorithms require a grid over the parameters ϕ and α. We choose ϕ in the range $[-90°, 90°]$ with $1°$ separation and α in the range $[-15, 15]$ with 1 unit separation. For CBF and SBL algorithms we set θ grid in the range $[-90°, 90°]$ with $1°$ separation. The SNR is 10 dB. Here we consider $K = 4$ off-grid sources with (ϕ, α) parameters $(-15.5, 2.5)$, $(-25.5, -6.5)$, $(47.5, 4.5)$, and $(71.5, -12.5)$ in a 100-snapshot block. The power spectrum in (ϕ, α) domain is shown in Fig. 4.2 (a) for TL-CBF & TL-SBL (top row) and in θ domain for CBF & SBL (bottom row). SBL and TL-SBL can identify all the off-grid sources, whereas CBF and TL-CBF miss a source. The corresponding DOA trajectories are shown in Fig. 4.1. The TL-CBF and TL-SBL algorithms can find accurate on-grid approximations of the true off-grid trajectories.

Example 2: Figure 4.3, 4.5, and 4.6 show the 2D spectrum obtained from TL-CBF, TL-SBL and TL-OMP, respectively. Observations are generated using a 10-sensor ULA with $\frac{\lambda}{2}$ spacing. In each block, $L = 30$ snapshots are processed at 5 dB SNR. The grid over linear parameters are set as $\phi = \{-85 : 2 : 85\}$ and $\alpha \in \{-5 : 0.5 : 5\}$. Four sources are present with trajectories $\{(-11, 3.5), (20, 1.5), (61, -2.25), (-52, -4.75)\}$. These include both on-grid and off-grid sources. Figures indicate both true and estimated trajectory parameters.

It can be seen from Figure 4.3 that TL-CBF has broad peaks, which makes it incapable of discerning closely spaced trajectories, leading to poor resolution. In addition, there are numerous spurious peaks associated with each source (see Figure 4.4 inset), which can cause repeated detection of the same source. In contrast, the TL-SBL spectrum in Figure 4.5 offers higher resolution than TL-CBF but is computationally intensive as the size of the search grid increases, making it unsuitable for real-time applications. On the other hand, the TL-OMP spectrum shown in Figure 4.6 can estimate the trajectory parameters accurately, but it is a greedy algorithm. The grid-based algorithms are prone to bias errors when the parameters are off-grid. In this section, we only discussed the case of linear trajectories, but these algorithms can be extended to other trajectories with the corresponding results presented later.

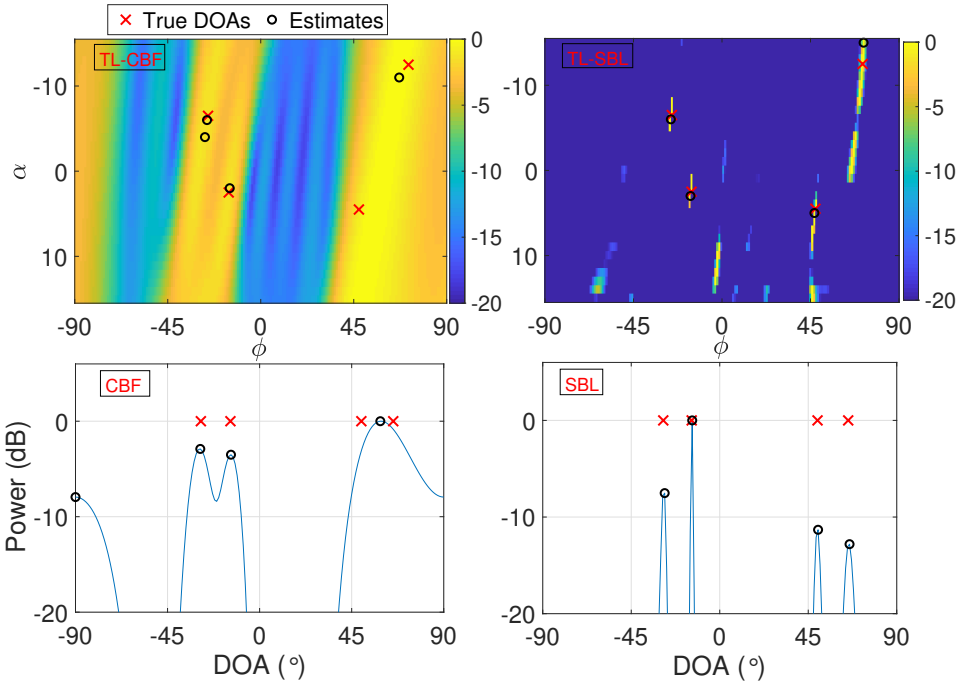

Figure 4.1 (**Example 1**) Corresponding DOA estimates of $K = 4$ off-grid sources obtained from TL-CBF, TL-SBL, CBF, and SBL algorithms at 10 dB SNR.

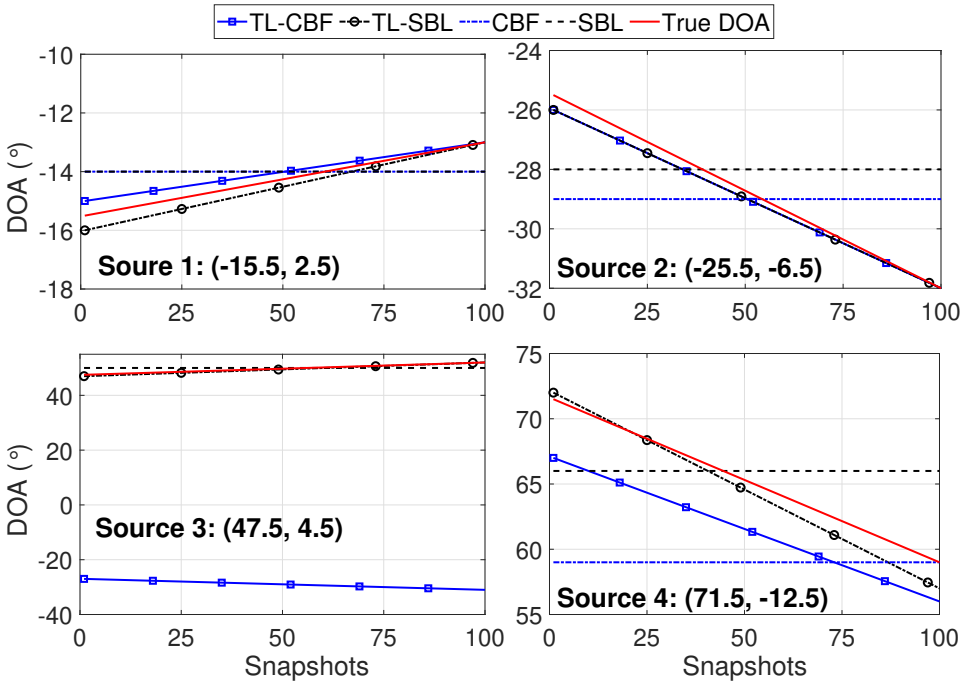

Figure 4.2 (**Example 1**) Power spectrum of $K = 4$ off-grid sources obtained from TL-CBF, TL-SBL, CBF, and SBL algorithms at 10 dB SNR.

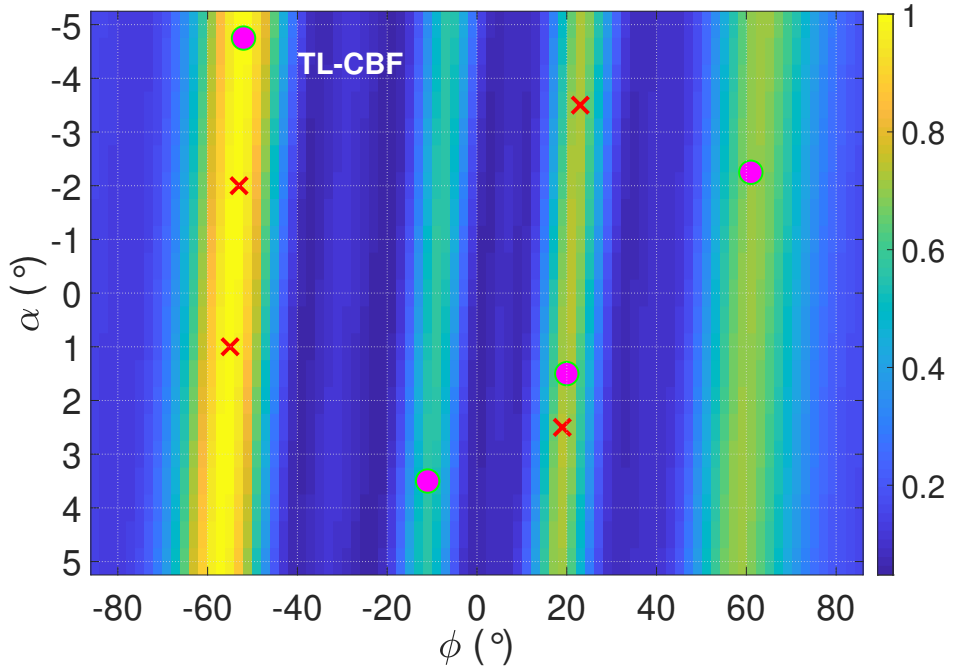

Figure 4.3 Example 2: TL-CBF spectrum for 4 source trajectories with true parameters $(-11, 3.5)$, $(20, 1.5)$, $(61, -2.25)$ and $(-52, -4.75)$ [circle]. Detected and assigned peaks are shown by red cross.

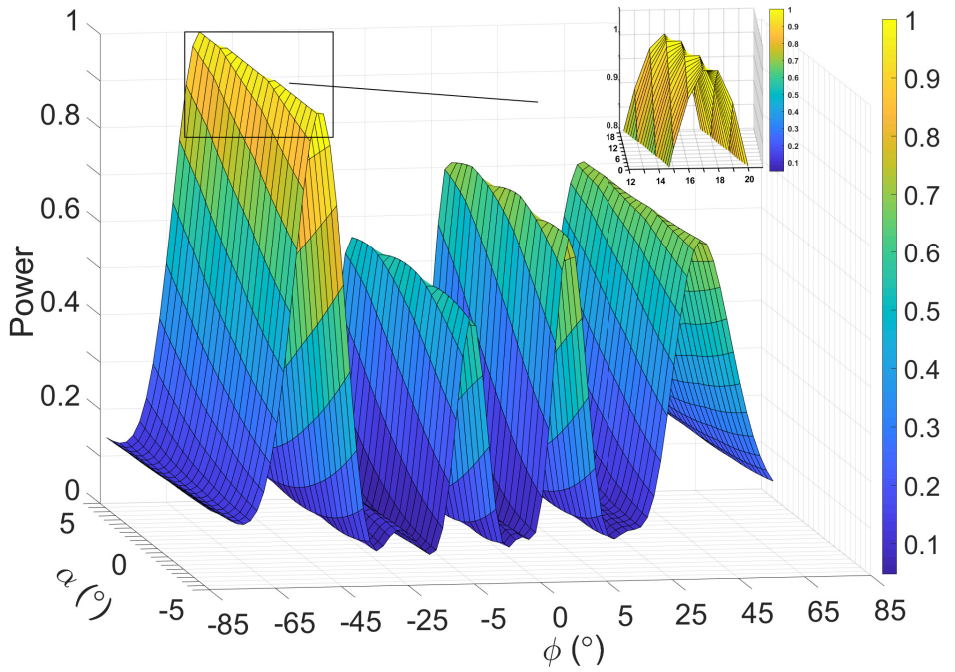

Figure 4.4 Example 2: 3D view of the TL-CBF spectrum with inset showing spurious peaks around the source $(-52, -4.75)$.

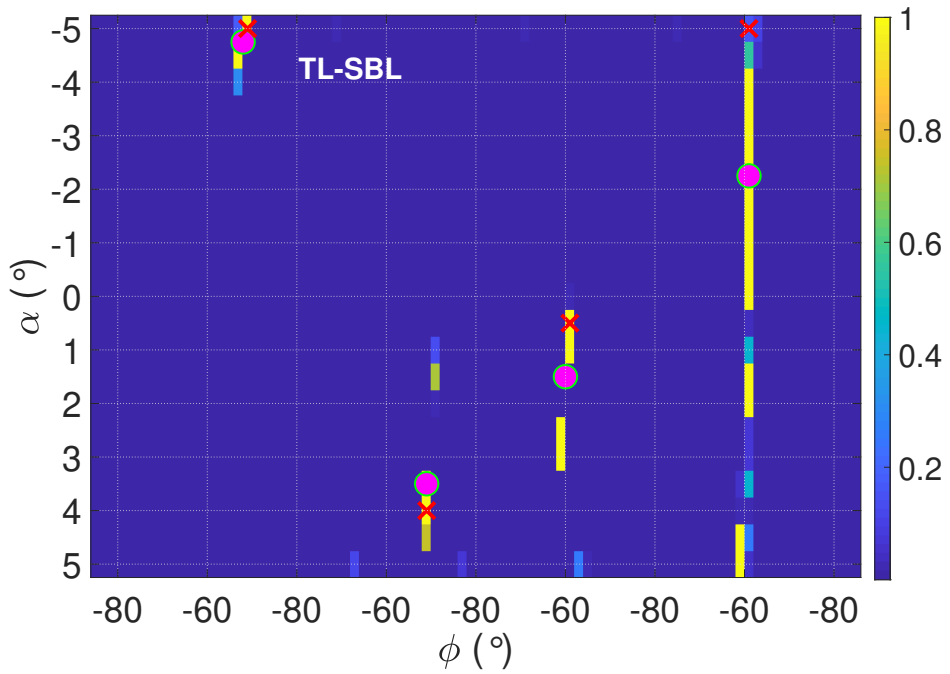

Figure 4.5 Example 2: TL-SBL spectrum for 4 source trajectories with true parameters $(-11, 3.5)$, $(20, 1.5)$, $(61, -2.25)$ and $(-52, -4.75)$ [circle]. Detected and assigned peaks are shown by red cross.

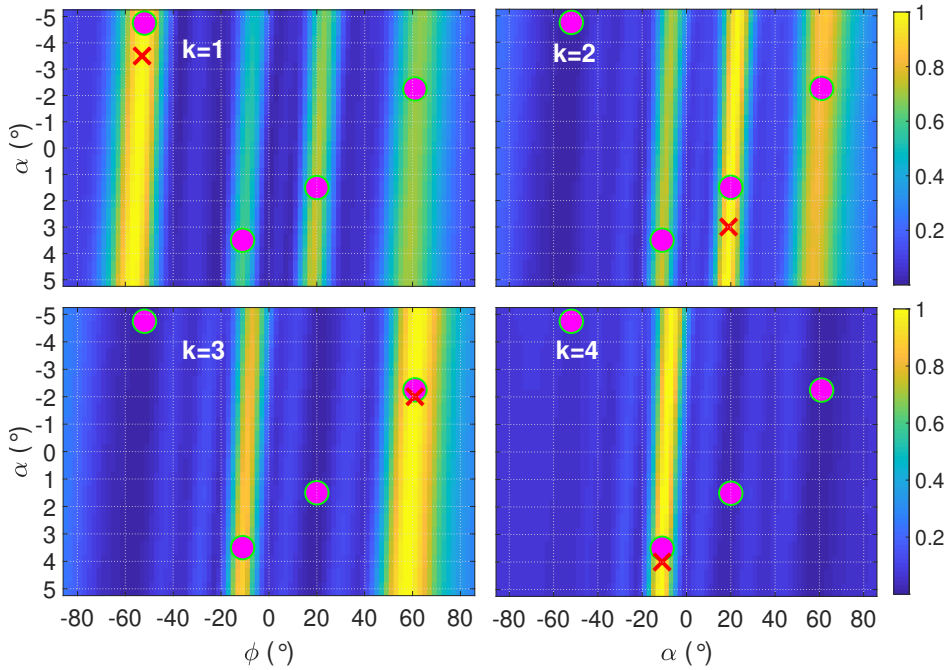

Figure 4.6 Example 2: TL-OMP spectrum at each iteration for 4 source trajectories with true parameters $(-11, 3.5)$, $(20, 1.5)$, $(61, -2.25)$ and $(-52, -4.75)$ [circle]. Detected and assigned peaks are shown by a red cross.

4.4 Gridless algorithms for trajectory localization

The performance of grid-based algorithms is limited when the true DOAs deviate from the grid or when the grid is too coarse, resulting in low resolution. Additionally, finer grid results in increased computational cost. In literature, various gridless methods have been proposed for DOA estimation to address the limitations of grid-based localization algorithms. Gridless localization has been formulated as an ANM problem and solved using semi-definite programming in 1D and 2D scenarios [20, 76–84]. Additionally, gridless methods have been applied for non-uniform arrays and multi-frequency processing [86–91]. The Newtonized OMP (NOMP) algorithm is a variation of OMP that employs Newton steps to refine source parameters in each iteration [21]. An alternative gridless approach is the Sliding Frank-Wolfe (SFW) algorithm [93], which solves the Beurling LASSO problem, i.e., a traditional LASSO in the continuum [22]. SFW has been extended to 3D acoustic source localization in a grid-free setting [91], and the choice of the regularization parameter is vital in obtaining accurate solutions. Here, We describe an alternate model for (4.7) and formulate the Beurling LASSO problem for gridless trajectory localization to address this limitation. We propose the TL-SFW and TL-NOMP algorithms to solve this and extend them for multi-frequency signals.

4.4.1 Beurling LASSO

Let there be K sources with the trajectory parameters $\mathcal{W} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K\} \subset \boldsymbol{\Psi}$. Similar to (4.1), the l^{th} snapshot can be expressed as

$$\mathbf{y}_l = \sum_{k=1}^{K} \mathbf{a}_l(\theta_k^l) x_k^l + \mathbf{n} = \sum_{k=1}^{K} \mathbf{a}_l(\boldsymbol{\omega_k}) x_k^l + \mathbf{n}\,. \tag{4.25}$$

Using Dirac mass $\delta_{\boldsymbol{\omega}}$ to represent a source with trajectory parameter $\boldsymbol{\omega} \in \boldsymbol{\Psi}$, we can reformulate (4.25)

$$\mathbf{y}_l = \int_{\Psi} \mathbf{a}_l(\boldsymbol{\omega})\, d\mu_l + \mathbf{n}\,, \tag{4.26}$$

$$\mu_l = \sum_{k=1}^{K} x_k^l\, \delta_{\boldsymbol{\omega}_k}\,, \tag{4.27}$$

where μ_l is the measure representing all the sources at the l^{th} snapshot. Across snapshots, the amplitudes change, but source trajectory parameters do not change. A Beurling LASSO problem is framed as,

$$\mu_l^* = \underset{\mu_l \in \mathcal{M}}{\arg\min}\ \frac{1}{2} \left|\left| \int_{\Psi} \mathbf{a}_l(\boldsymbol{\omega})\, d\mu_l - \mathbf{y}_l \right|\right|_2^2 + \lambda |\mu_l|\,, \tag{4.28}$$

where μ_l^* is the solution of the optimization problem, $\mathcal{M}$ is the set of complex measures defined on $\boldsymbol{\Psi}$, λ is the regularization parameter and $|\mu_l|$ represents any sparsity inducing norm of the measure μ_l. The regularization parameter λ can be tuned to find the number of sources. The single snapshot formulation

is not useful for estimating trajectories. The multiple snapshot extension could be written as,

$$\boldsymbol{\mu}^* = \underset{\mu_l \in \mathcal{M}, \forall l}{\arg\min} \frac{1}{2} \sum_{l=0}^{L-1} \left|\left| \int_{\Psi} \mathbf{a}_l(\boldsymbol{\omega})\, d\mu_l - \mathbf{y}_l \right|\right|_2^2 + \lambda \sum_{l=0}^{L-1} |\mu_l|\,, \tag{4.29}$$

where $\boldsymbol{\mu}^*$ is the collection of solution to all the measures $\{\mu_l^*, l = 0, 1, \ldots, L-1\}$. In this work, we assume the number of trajectories to be known; thus, we set $\lambda = 0$ and develop greedy iterative algorithms [91]. From the solution $\boldsymbol{\mu}^*$, we obtain estimates for the trajectory parameters $\mathcal{W}$ and their corresponding amplitudes using (4.27). In the presence of multi-frequency observations $\mathbf{Y}_f, f = 1, 2, \ldots, F$, a multi-frequency Beurling LASSO can be constructed by adding across frequencies.

4.4.2 Sliding Frank-Wolfe algorithm (TL-SFW)

Algorithm 3 TL-SFW pseudo-code to solve (4.29)

1. $\mathcal{W}^{[0]} \leftarrow \emptyset, \mathbf{R}^{[0]} \leftarrow \mathbf{Y}, tol = 1e^{-10}$
2. **for** $k = 1, \ldots, K$
3. Find the next source:
   $\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \boldsymbol{\Psi}}{\arg\max} \frac{1}{L} \sum_{l=1}^{L} \left| \mathbf{a}_l^H(\boldsymbol{\omega})\, \mathbf{r}_l^{[k-1]} \right|^2 \qquad (a)$
4. $\mathcal{W}^{[\frac{k-1}{2}]} = \{\mathcal{W}^{[k-1]},\ \boldsymbol{\omega}^*\}$
5. Optimize the amplitude:
   $\bar{\mathbf{X}}^{[\frac{k-1}{2}]} = \underset{\bar{\mathbf{X}} \in \mathcal{X}_k^L}{\arg\min} \frac{1}{2} \left|\left| \bar{\mathbf{A}}(\mathcal{W}^{[\frac{k-1}{2}]})\, \bar{\mathbf{X}} - \mathbf{Y} \right|\right|_{\mathcal{F}}^2 \qquad (b)$
6. Optimize the amplitudes and parameters:
   $\bar{\mathbf{X}}^{[k]}, \mathcal{W}^{[k]} = \underset{\mathcal{W} \subset \boldsymbol{\Psi}, \bar{\mathbf{X}} \in \mathcal{X}_k^L}{\arg\min} \frac{1}{2} \left|\left| \bar{\mathbf{A}}(\mathcal{W})\, \bar{\mathbf{X}} - \mathbf{Y} \right|\right|_{\mathcal{F}}^2 \quad (c)$
7. $\mathbf{R}^{[k]} \leftarrow \mathbf{Y} - \bar{\mathbf{A}}(\mathcal{W}^{[k]})\, \bar{\mathbf{X}}^{[k]}$
8. **end for**

MATLAB `fmincon` is used to solve equations (a), (b), (c)

We solve the Beurling LASSO problem (4.29) using greedy ($\lambda = 0$) Sliding Frank-Wolfe (SFW) algorithm [22,91,93]. The SFW algorithm for trajectory localization (TL-SFW) is detailed in Algorithm 3. We iteratively solve (4.29) by adding one source at a time. An empty set is denoted as $\emptyset$. $\mathbf{R}^{[k]}$ denotes the $N \times L$ residual matrix at the end of iteration k and is initialized as $\mathbf{R}^{[0]} = \mathbf{Y}$. Each iteration over K trajectories consists of the following steps:

(i) **Add a source:** Solve (4.23) to find a coarse trajectory estimate on the predefined grid $\boldsymbol{\Psi}_d$. Use this estimate as initialization to solve the global optimization problem (a) in Algorithm 3 to obtain $\boldsymbol{\omega}^*$.

(ii) **Amplitude estimation:** Initialize all the k source amplitudes as $\text{diag}(\tilde{\mathbf{X}}_k) = \text{diag}(\tilde{\mathbf{A}}^H(\boldsymbol{\omega}_k)\mathbf{Y})$ using the estimated trajectory parameters $\mathcal{W}^{[\frac{k-1}{2}]}$. Solve (b) to obtain optimized amplitudes $\bar{\mathbf{X}}^{[\frac{k-1}{2}]}$.

(iii) **Joint estimation:** Jointly optimize the trajectory parameters and amplitudes by solving (c). Initialization is done using $\mathcal{W}^{[\frac{k-1}{2}]}$ and $\bar{\mathbf{X}}^{[\frac{k-1}{2}]}$ for this non-convex optimization problem.

The algorithm is proven to converge in a finite number of iterations under certain constraints [93]. Optimizations (a), (b), and (c) are performed using the sequential quadratic programming algorithm [130] in the MATLAB 2018b function `fmincon`. For multi-frequency observations, problems (a) and (c) are respectively modified as,

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega}\in\boldsymbol{\Psi}}{\text{argmax}} \frac{1}{L}\sum_{f=1}^{F}\sum_{l=1}^{L}\left|\mathbf{a}_{lf}^H(\boldsymbol{\omega})\,\mathbf{r}_{lf}^{[k-1]}\right|^2 \tag{4.30}$$

$$\{\bar{\mathbf{X}}_f^{[k]}\}_{f=1}^{F}, \mathcal{W}^{[k]} = \underset{\mathcal{W}\subset\boldsymbol{\Psi},\bar{\mathbf{X}}_f\in\mathcal{X}_k^L}{\text{argmin}} \frac{1}{2}\sum_{f=1}^{F}\left|\left|\bar{\mathbf{A}}_f(\mathcal{W})\,\bar{\mathbf{X}}_f - \mathbf{Y}_f\right|\right|_{\mathcal{F}}^2 .$$

For multi-frequency processing, the trajectory parameters are estimated using the averaged spectrum over F frequencies. The optimization (b) is solved F times to obtain the amplitudes $\bar{\mathbf{X}}_f$ at each frequency. As the number of frequencies increases, the number of unknown parameters also increases, leading to a higher computational cost.

4.4.3 Newtonized OMP (TL-NOMP)

Newtonized orthogonal matching pursuit (NOMP) is a variant of OMP that incorporates Newton refinements to obtain precise off-grid estimates [21,92]. The NOMP algorithm for trajectory localization (TL-NOMP) is given in Algorithm 4.

NOMP has three main steps when adding a new source:

(i) **Find a source:** Obtain an initial coarse estimate $\boldsymbol{\omega}^*$ of source trajectory parameter by searching over the grid $\boldsymbol{\Psi}_d$ using (4.23) and estimate the corresponding amplitudes $\tilde{\mathbf{X}}^*$.

(ii) **Local Newton refinement:** Compute the Hessian matrix $(\mathbf{H})$ and gradient vector $(\boldsymbol{g})$ for the objective in (4.29) (assuming $\lambda = 0$). Refine the on-grid trajectory parameter estimate using single-step Newton's method over the continuum $\boldsymbol{\Psi}$.

(iii) **Global cyclic refinement:** Starting with the current residual $\mathbf{R}^*$ as the observation, add back each of the identified sources (one at a time) and optimize parameters using Local Newton refinement. Repeat until the convergence criteria are met.

The local Newton refinement provides an improvement on the initial on-grid parameter estimate. In contrast, the global cyclic refinement provides a feedback mechanism to improve the estimates accumulated from previous iterations. At the end of the k-th iteration, the residual $\mathbf{R}^{[k]}$ is updated using

Algorithm 4 TL-NOMP pseudo-code to solve (4.29)

1. $\mathcal{W}^{[0]} \leftarrow \emptyset,\ \mathbf{R}^{[0]} \leftarrow \mathbf{Y},\ tol = 1e^{-6}$
2. **for** $k = 1, \ldots, K$
3. Find the next source:
   $\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \boldsymbol{\Psi}_d}{\arg\max}\ \frac{1}{L} \sum_{l=1}^{L} \left| \mathbf{a}_l^H(\boldsymbol{\omega})\ \mathbf{r}_l^{[k-1]} \right|^2$
   $\text{diag}(\tilde{\mathbf{X}}^*) = \text{diag}\big(\tilde{\mathbf{A}}^H(\boldsymbol{\omega}^*)\mathbf{R}^{[k-1]}\big)$
4. Local Newton refinement:
   $\boldsymbol{\omega}^* = \boldsymbol{\omega}^* - \mathbf{H}^{-1}\boldsymbol{g}$
   $\text{diag}(\tilde{\mathbf{X}}^*) = \text{diag}\big(\tilde{\mathbf{A}}^H(\boldsymbol{\omega}^*)\mathbf{R}^{[k-1]}\big)$
5. $\mathcal{W}^{[\frac{k-1}{2}]} = \{\mathcal{W}^{[k-1]},\ \boldsymbol{\omega}^*\}$
6. Global cyclic refinement:
   $\mathbf{R}^* \leftarrow \mathbf{Y} - \bar{\mathbf{A}}(\mathcal{W}^{[\frac{k-1}{2}]})\ \bar{\mathbf{X}}^{[\frac{k-1}{2}]}$
   while $\left| ||\mathbf{R}^{[k-1]}||_f^2 - ||\mathbf{R}^*||_f^2 \right| < tol$
   for $i = 1, \ldots, k$
   $\hat{\mathbf{R}} = \mathbf{R}^* + \tilde{\mathbf{A}}(\boldsymbol{\omega}_i)\tilde{\mathbf{X}}_i$
   $\text{diag}(\tilde{\mathbf{X}}_i) = \text{diag}\big(\tilde{\mathbf{A}}^H(\boldsymbol{\omega}_i)\hat{\mathbf{R}}\big)$
   Local Newton refinement of $\boldsymbol{\omega}_i$ and $\tilde{\mathbf{X}}_i$
   $\mathbf{R}^{[k-1]} \leftarrow \mathbf{R}^*,\ \mathbf{R}^* \leftarrow \hat{\mathbf{R}} - \tilde{\mathbf{A}}(\boldsymbol{\omega}_i)\tilde{\mathbf{X}}_i$
   end for
   end while
7. Use (4.24) to find the orthogonal residual $\mathbf{R}^{[k]}$
8. **end for**

(4.24) where data is orthogonally projected onto steering vectors corresponding to identified source trajectories. For the multi-frequency implementation of NOMP, the objective in 4.29 is used instead.

Both TL-SFW and TL-NOMP solve (4.29) with the primary distinction lying in their refinement processes. Once the coarse trajectory parameter is found, the TL-SFW solves three optimization problems: i) optimizing the coarse trajectory parameters using (a), ii) optimization of amplitudes for each trajectory using (b), and iii) the joint estimation of estimated amplitude and trajectory parameters using (c). The residual is updated in each iteration, and trajectory and amplitudes are optimized accordingly. In contrast, the TL-NOMP uses a local single refinement process over each coarse on-grid estimate (instead of solving (a) in TL-SFW) and then a global cyclic refinement method to converge and attain the optimal solution, continuing until the stopping criteria are met. In the global cyclic refinement, the contribution of each source is removed, and the rest are optimized as described earlier. The implementation of SFW and NOMP is referred from [131], a helpful resource for developing TL-SFW and TL-NOMP.

4.5 Simulation Results

4.5.1 Simulation setup

We demonstrate various algorithms using simulations with linear and nonlinear trajectories. The performance of TL-SFW and TL-NOMP are compared with TL-CBF, TL-SBL and TL-OMP. A $10-$ sensor uniform linear array (ULA) with inter-sensor spacing $d = \frac{\lambda}{2}$ is used. Unless stated otherwise, simulations are for linear trajectories and narrowband signals. For grid-based methods TL-CBF, TL-SBL, and TL-OMP, we construct the following grid over trajectory parameters: $\phi \in \{-85 : 2 : 85\}$ and $\alpha \in \{-5 : 0.5 : 5\}$ resulting in a dictionary with $M = 86 \times 21 = 1806$ trajectory steering matrices $\tilde{\mathbf{A}}$. Throughout the simulations, we consider $L = 30$ snapshots within a block at an SNR of 5 dB. The source amplitudes and noise are complex Gaussian of the form $a + jb$ where a and b are generated using zero-mean Gaussians. The signal and noise variance are σ_x^2 and σ_n^2, respectively. The signal-to-noise ratio is defined as SNR $= 10\log_{10}(\frac{\sigma_x^2}{\sigma_n^2})$. For TL-SBL, the noise variance is assumed to be known and directly used in the update rule. However, an update rule for estimating the noise variance can also be derived [44, 100, 102].

To compare the localization accuracy of TL methods, we report RMSE. Let θ_k^l and $\hat{\theta}_k^l$ be the ground truth and estimated DOA obtained from trajectory parameters corresponding to the k^{th} source. The RMSE for k^{th} source is given by,

$$\text{RMSE}_k = \sqrt{\frac{\sum_{l=0}^{L-1} (\theta_k^l - \hat{\theta}_k^l)^2}{L}}, \quad k = 1, \ldots, K. \tag{4.31}$$

We perform 100 Monte Carlo trials and report the RMSE averaged across all the trials and sources. For TL-CBF and TL-SBL, if K sources are present, we identify $\hat{K} = K + 2$ peaks in the power spectrum. By considering more peaks, we overcome the problem of spurious peaks and get the best possible estimates closer to true trajectories. The Optimal SubPattern Assignment (OSPA) [132, 133] is used to solve the assignment problem between the $\hat{K}$ estimated trajectories and K true trajectories. Let $\hat{\mathcal{T}} \triangleq \left\{\hat{T}_1, \ldots, \hat{T}_{\hat{K}}\right\}$ be the set of $\hat{K}$ estimated trajectories and $\mathcal{T} \triangleq \{T_1, \ldots, T_K\}$ be the set of K true trajectories. The OSPA metric for sets $\mathcal{T}$ and $\hat{\mathcal{T}}$ is defined as

$$\text{OSPA}(\mathcal{T}, \hat{\mathcal{T}}) \triangleq \left[\frac{1}{\hat{K}} \min_{\pi \in \Pi_{\hat{K}}} \sum_{k=1}^{K} d_c(T_k, \hat{T}_{\pi(k)})^p + (\hat{K} - K)c^p\right]^{\frac{1}{p}} \tag{4.32}$$

where $K \leq \hat{K}$, the order parameter is $1 \leq p \leq \infty$ and c is the cutoff parameter. $\Pi_{\hat{K}}$ denotes the set of all permutations of length K with elements $\{1, \ldots \hat{K}\}$. The $d_c(T_k, \hat{T}_{\pi(k)}) \triangleq \min(c, d_t(T_k, \hat{T}_{\pi(k)})$, where $d_t(T_k, \hat{T}_{\pi(k)})$ denotes the error between two trajectories computed using (4.31). We choose $p = 2$ and $c = 100$. Once assigned, a source is said to be detected if the RMSE between ground truth and the assigned track is less than the detection threshold of 5°. We report the probability of detection P_d, i.e. the percentage of detected sources. The average RMSE is reported only for detected sources.

4.5.2 SNR

We perform simulations with SNR ranging from -10dB to 30dB. Four source trajectories (linear) are processed in a block containing $L = 30$ snapshots. The true trajectory parameters are $\mathcal{W} = \{(-11, 3.5), (20, 1.5), (61, -2.25), (-52, -4.75)\}$, such that some parameters are on-grid while the rest are off-grid. The minimum error achievable by on-grid methods for each of these trajectories are $0, 0.51, 0.15,$ and, 0.53 respectively, giving an average of 0.30. The error vs SNR and P_d vs SNR plots are shown in Figure 4.7. At low SNR, TL-CBF has the lowest RMSE; however, it exhibits lower P_d compared to other approaches as it fails to detect all the sources. Both TL-NOMP and TL-SFW outperform all the grid-based methods as they can optimize the parameters beyond the grid. As SNR increases, most algorithms reach saturation except TL-NOMP, which consistently enhances its performance. TL-SFW has a slightly better detection rate at low SNR than TL-NOMP. TL-SBL error saturates to the value of 0.30 beyond which its performance cannot improve since it can only find sources on the grid. It performs better than TL-OMP, which is a greedy algorithm.

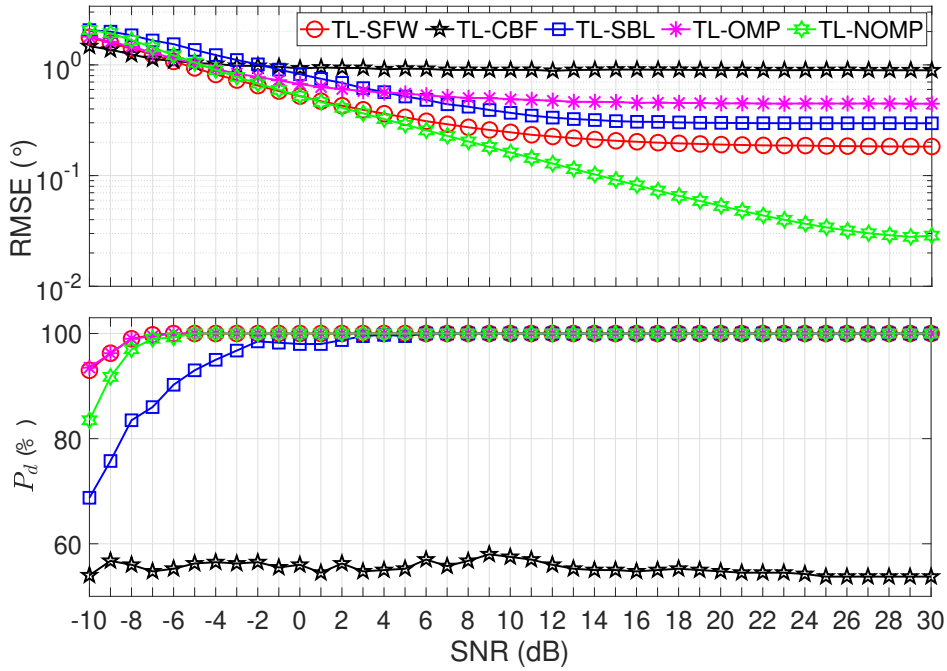

Figure 4.7 Evaluation of TL-methods for linear trajectory localization for various SNR values. RMSE vs SNR (top) and P_d vs SNR (bottom).

4.5.3 Snapshots

We evaluate algorithm performance with the number of snapshots ranging from 5 to 50 at 5 dB SNR. The true trajectory parameters are the same as above. Figure 4.8 shows that as the number of snapshots increases, the error decreases for all the algorithms. Both TL-SFW and TL-NOMP show superior performance compared to all the other methods. Grid-based methods exhibit higher error than

grid-free methods due to the bias present while estimating off-grid trajectory parameters, regardless of the number of snapshots. TL-CBF has higher P_d for fewer snapshots, which reduce with increasing snapshot number. This is likely due to the presence of spurious peaks (Figure 4.3), which become more prominent with increasing snapshots (4.15).

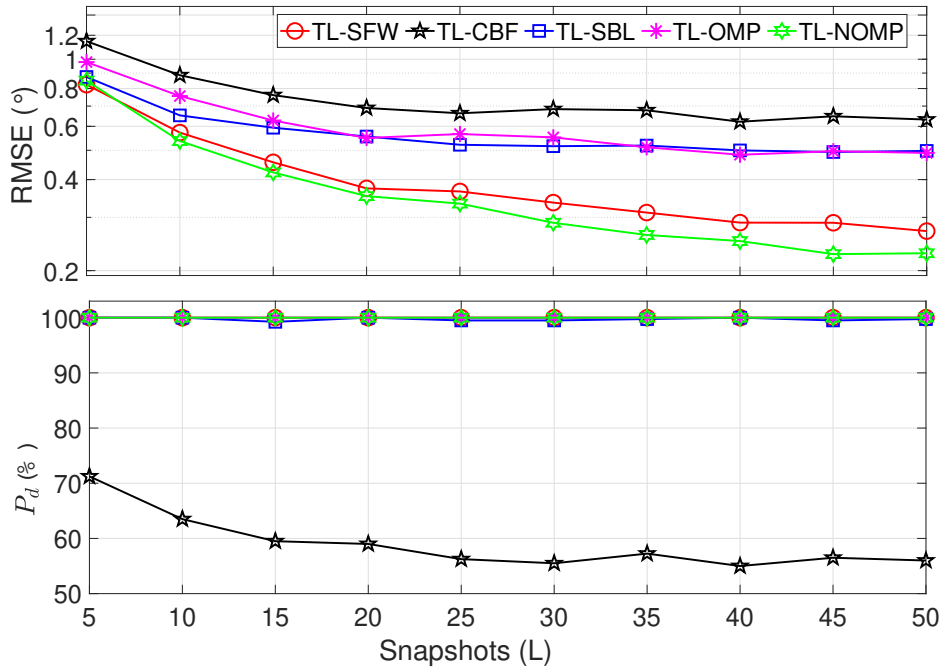

Figure 4.8 Evaluation of TL-methods for linear trajectory localization for various snapshots processed within a block. RMSE vs Snapshots (top) and P_d vs Snapshots (bottom).

4.5.4 Grid step-size

We analyze the impact of step-size (ϕ_{step}) used for creating ϕ grid in trajectory localization tasks. The grid over α is fixed with $\alpha \in \{-5 : 0.5 : 5\}$ while the grid over ϕ is made coarser by increasing ϕ_{step} from 1 to 10. Let $\boldsymbol{\phi}_g$ be the grid vector constructed using ϕ_{step} with N_ϕ grid points. For this ϕ_{step} experiment, the true parameters are $\big(\boldsymbol{\phi}_g(\lfloor N_\phi \times 0.2 \rfloor), 3.5\big)$, $\Big(\boldsymbol{\phi}_g(\lfloor N_\phi \times 0.45 \rfloor) + \frac{\phi_{\text{step}}}{2}, 1.5\Big)$, $\big(\boldsymbol{\phi}_g(\lfloor N_\phi \times 0.65 \rfloor), -2.5\big)$ and $\Big(\boldsymbol{\phi}_g(\lfloor N_\phi \times 0.9 \rfloor) + \frac{\phi_{\text{step}}}{2}, -4.75\Big)$ where $\lfloor . \rfloor$ denotes the floor of a real number. These source trajectories are chosen such that the true ϕ and α parameters have both on-grid and off-grid combinations. As the step-size increases, the grid becomes less refined, and the performance of grid-based methods is expected to degrade. Whereas TL-SFW and TL-NOMP are expected to perform better since they improve upon the initial on-grid estimates by performing optimization and refinement, respectively. This analysis is verified from simulation results shown in Figure 4.9. The impact of grid step-size on gridless methods is low with TL-NOMP being most robust to the coarseness of the ϕ grid.

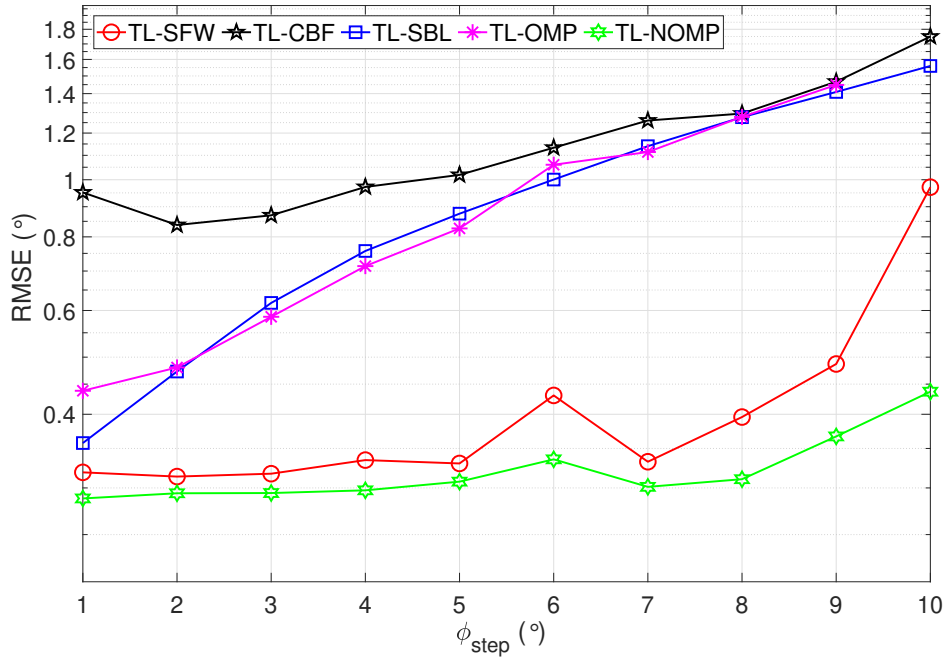

Figure 4.9 Error as function of parameter ϕ grid step-size with $L = 30$ snapshots at 5dB SNR.

4.5.5 Resolution

Resolution refers to the ability to distinguish between two nearby trajectories accurately. We consider 3 sources with linear trajectory parameters as follows $\mathcal{W} = \{(0, 3.5), (60, -4.5), (\zeta, 2.5)\}$. The 3^{rd} source trajectory varies as we increase ζ from -15 to 15. Specifically, its trajectory approaches that of the 1^{st} source and then diverges. We process 30 snapshots at 5 dB SNR. The results are shown in Figure 4.10. TL-CBF, TL-OMP, and TL-SFW have low resolution when dealing with closely spaced trajectories, as indicated by the peaks in the RMSE plot. Both TL-SBL and TL-NOMP outperform other methods, with TL-NOMP having the lowest error among all the methods. The detection performance of TL-SBL is influenced by our approach of selecting five peaks from the spectrum and subsequently identifying the three closest tracks after source association. Though there is a dip in error for all algorithms around $\zeta \in [-3, 3]$, it is likely due to repeated identification of the same source. It cannot be attributed to superior resolution ability.

4.5.6 Linear trajectory approximation for slowly moving sources

Here, we demonstrate how the linear trajectory model approximates and captures the DOA motion of slowly moving sources with nonlinear trajectories.

Example 1: We simulate a moving source with nonlinear DOA trajectory as shown in Fig. 4.11. The trajectory contains 31 non-overlapping blocks of $L = 50$ snapshots each. Within each block, the DOA trajectory is approximately linear. The maximum change in DOA within any block is $11.5°$.

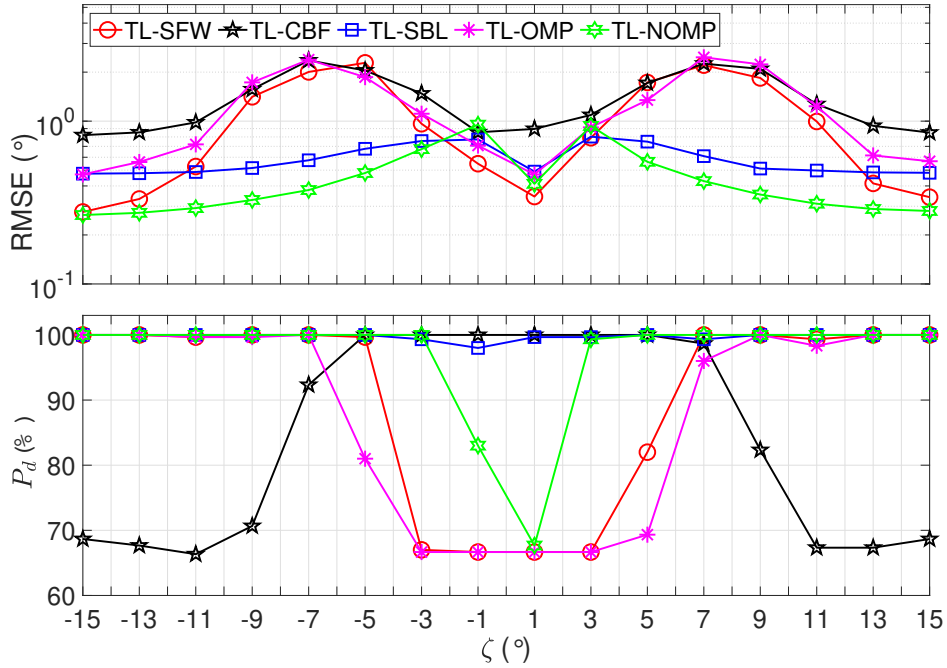

Figure 4.10 Error as a function of source proximity (ζ) with $L = 30$ snapshots at 5dB SNR. RMSE vs ζ (top) and P_d vs ζ (bottom).

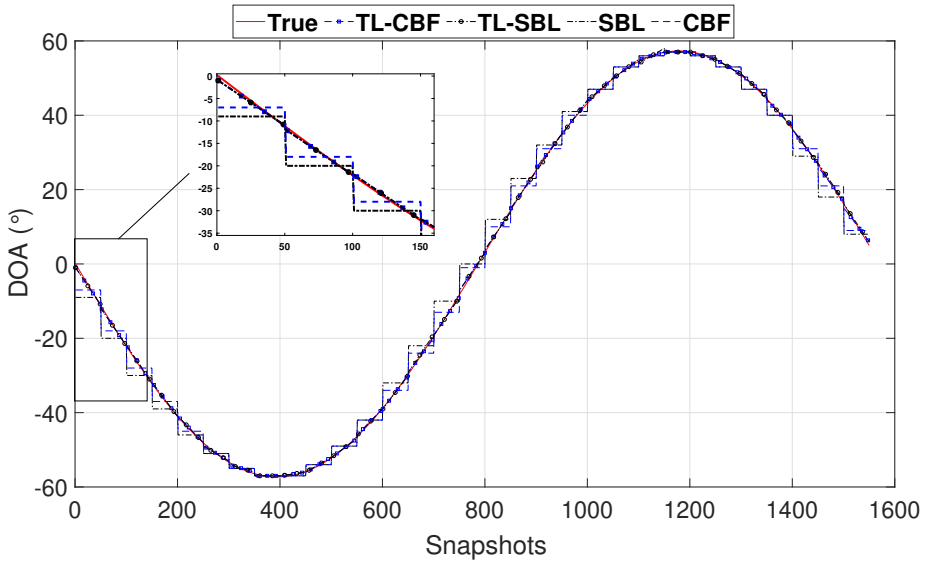

Figure 4.11 (**Example 1**) DOA estimates of CBF, TL-CBF, SBL, and TL-SBL for a moving source (10 dB SNR).

Trajectories estimated from TL-CBF and TL-SBL closely align with the true trajectory whereas CBF and SBL provide fixed DOA estimates in each block.

Example 2: Two moving sources with non-linear DOA trajectories are simulated in Fig. 4.12 (52 blocks with $L = 30$ snapshots each). The estimated DOAs by TL methods provide relatively smoother trajectories. The root-mean-square DOA error for non-crossing regions is $2.98°$, $3°$, $2°$, and $1.78°$ for CBF, TL-CBF, SBL, and TL-SBL respectively.

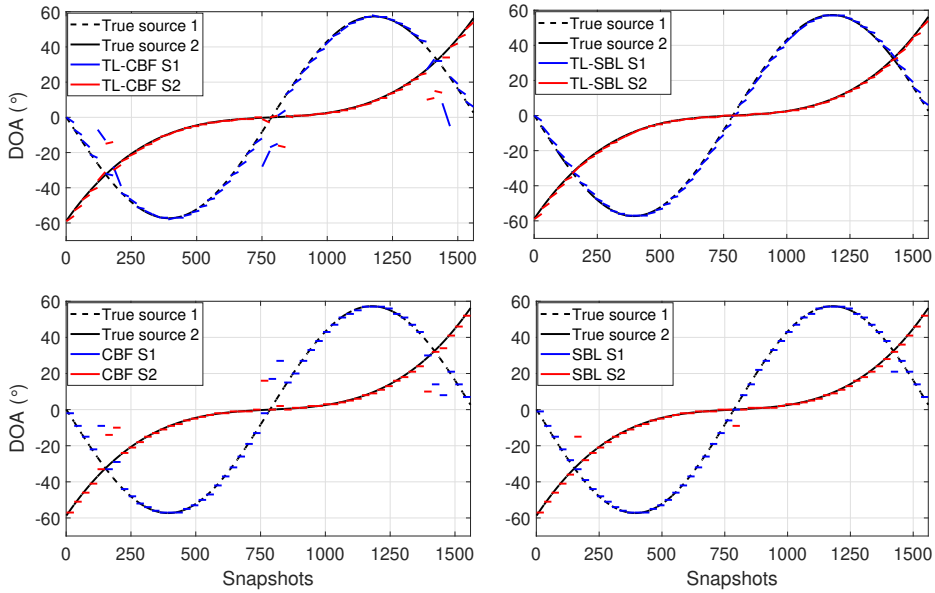

Figure 4.12 (**Example 2**) DOA estimates of CBF, TL-CBF, SBL, and TL-SBL for two moving sources (10 dB SNR).

4.5.7 Non-linear trajectories

Sample nonlinear trajectories, generated using 3 parameter quadratic and harmonic trajectory models, are shown in Figures 4.13 and 4.14, respectively. Each trajectory spans over $L = 40$ snapshots. Estimated trajectories, by processing observations at 20 dB SNR, using TL-SFW and TL-NOMP are shown as well. For both models, we construct the following grid over trajectory parameters: $\phi \in \{-85 : 2 : 85\}$ and $\alpha_1,\ \alpha_2,\ \beta_1 \in \{-5 : 0.5 : 5\}$, resulting in a dictionary with $M = 86 \times 21 \times 21 = 37926$ trajectory steering matrices $\tilde{\mathbf{A}}$. This is significantly larger than the number of grid points in the linear case. Figure 4.15 shows error vs SNR for nonlinear trajectory estimation. We set $L = 30$ and use sources with polynomial trajectories: $\mathcal{W} = \{(-60, 1, -3), (-31, 0.4, 3.6), (20, -3, 2), (51, 4, -0.2)\}$. TL-CBF frequently fails to detect trajectories, giving a poor detection rate of $P_d \approx 40\%$. TL-NOMP performs worse than TL-OMP at low SNR (both in RMSE and P_d) but recovers at higher SNR values, outperforming all other algorithms. TL-SFW shows marginal improvement over TL-OMP, with its error saturating at high SNR. All the results presented so far use a detection threshold of $5°$. Here, we investigate the effect of changing this detection threshold on detection probability P_d.

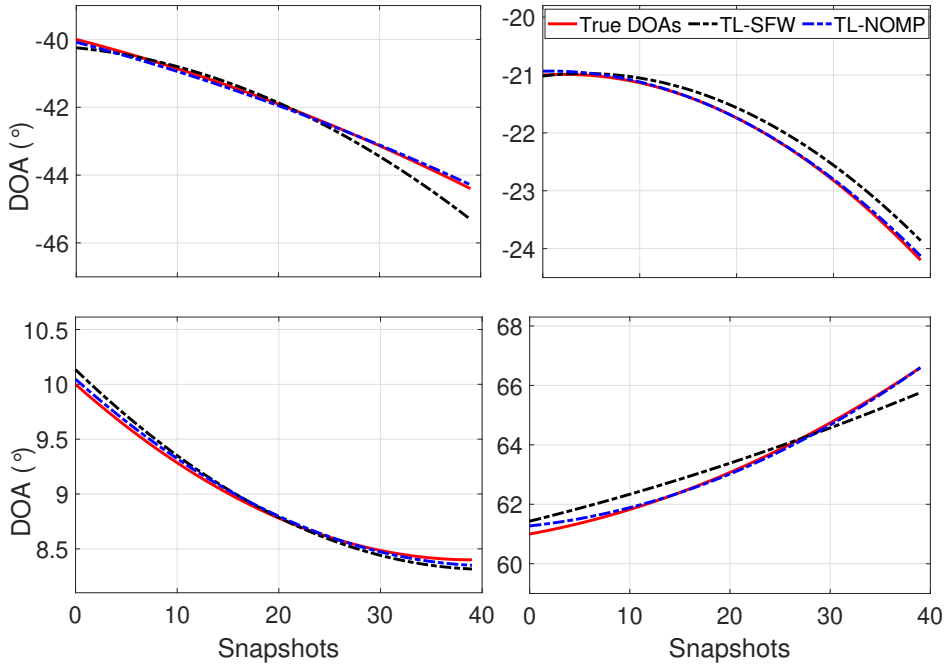

Figure 4.13 Quadratic model: True and estimated trajectories using TL-SFW and TL-NOMP for single block at 5 dB SNR. True trajectories: $(-40, -3, -1.4)$, $(-21, 0.4, -3.6)$, $(10, -3.2, 1.6)$, $(61, 2.4, 3.2)$.

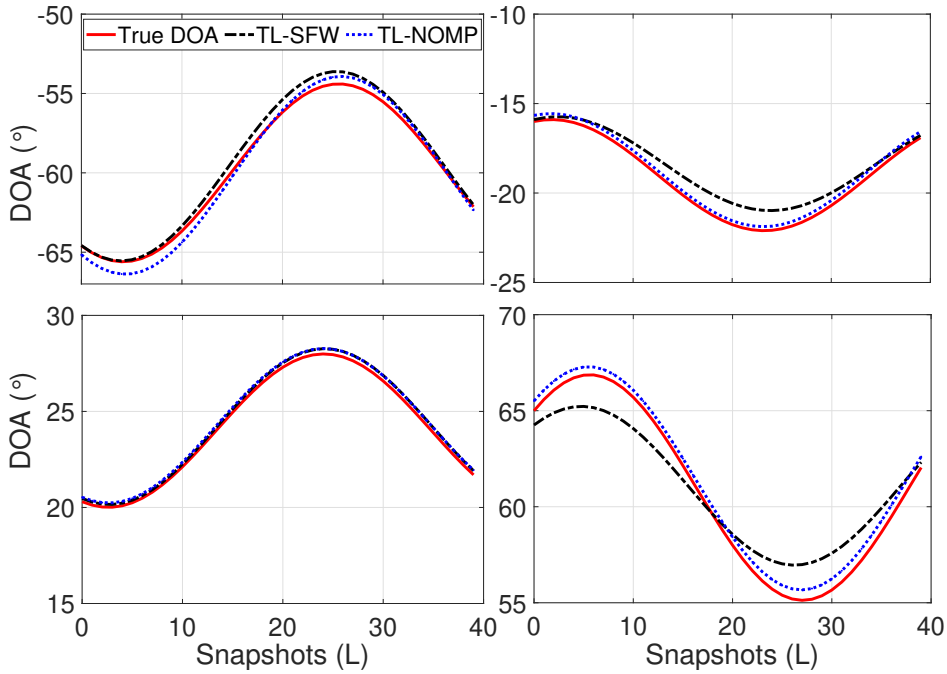

Figure 4.14 Harmonic trajectory model: True and estimated trajectories using TL-SFW and TL-NOMP for a single block at 5 dB SNR. The true trajectories are $(-60, -3.2, -4.6)$, $(-19, 0.8, 3)$, $(24, -1.5, -3.7)$, $(61, 4.3, 4)$.

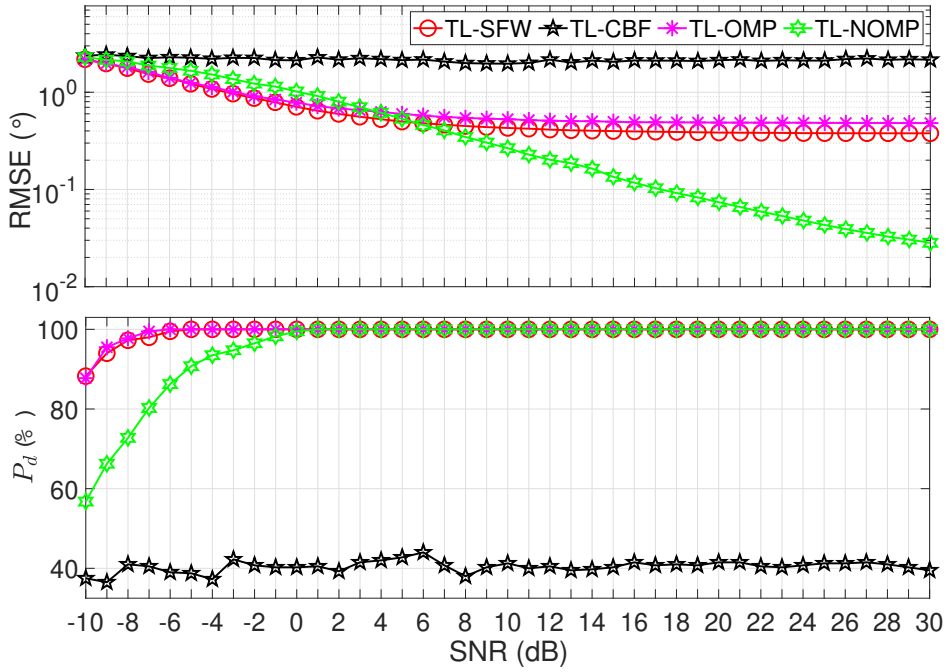

Figure 4.15 Performance of TL-methods for nonlinear trajectory localization at various SNR values. RMSE vs SNR (top) and P_d vs SNR (bottom).

Figure 4.16 depicts the P_d as the detection threshold is changed for select SNR values. As expected, an increase in the value detection threshold increases P_d. Similar to the inference from Figure 4.15, at lower SNR, the performance of TL-OMP is better than that of TL-NOMP, whereas TL-SFW shows superior detection performance at all SNR levels.

4.5.8 Computational effort

In this section, we present the computational time analysis of methods by varying snapshots from 5 to 50, at 5 dB SNR. We conduct experiments on a desktop equipped with an Intel(R) Core(TM) i7-8700 CPU operating at 3.19 GHz $\times$ 8 cores and 32 GB of memory. Figure 4.17 illustrates the computational time required by each method for estimating linear (top) and nonlinear (bottom) trajectories. TL-CBF and TL-OMP exhibit high computational efficiency leading to significantly shorter execution times when compared to other methods. For nonlinear trajectories, TL-SBL requires significantly longer execution times, even with a small number of snapshots. Hence, we omit TL-SBL results for the nonlinear case. The computational requirements of TL-NOMP are higher than that of TL-SFW.

4.5.9 Multi-frequency processing

We generate multi-frequency observations and apply TL algorithms. The TL-SFW processes the multi-frequency signals in a coherent manner (4.30), whereas other TL methods process them non-coherently. We extend the TL-CBF and TL-OMP to multi-frequency observations by summing the

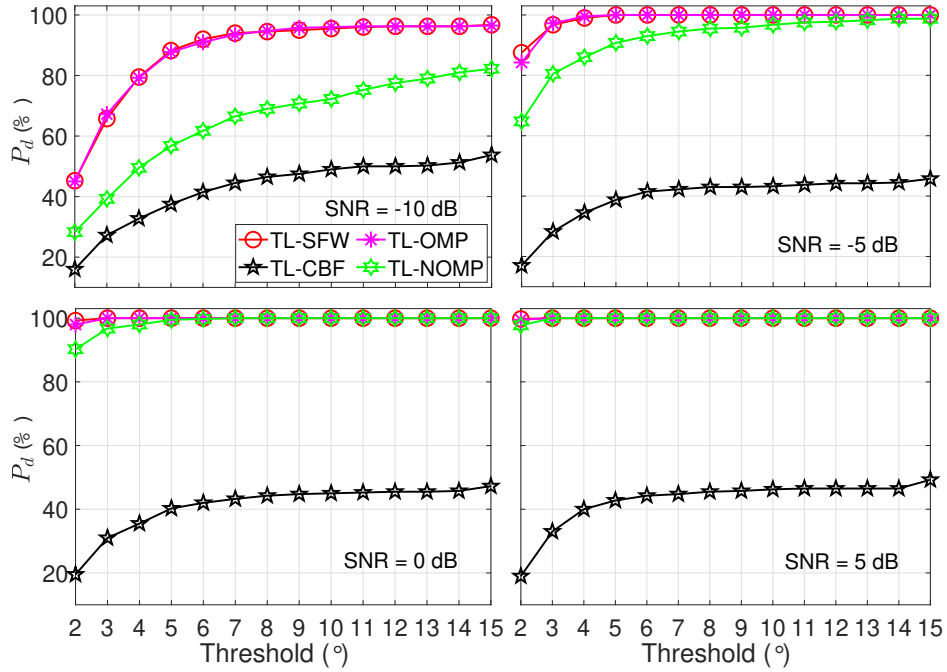

Figure 4.16 Performance of TL-methods for nonlinear trajectory localization with varying detection threshold at different SNR.

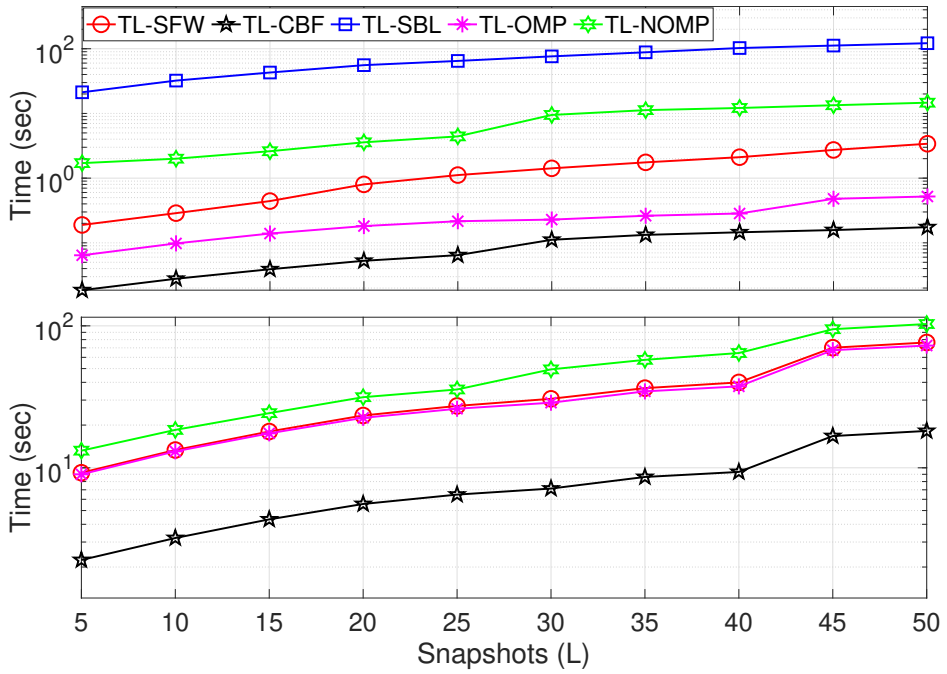

Figure 4.17 Complexity analysis of TL-methods for a varying number of snapshots. Linear trajectory model (top) and quadratic trajectory model (bottom).

spectrum across frequencies in (4.15) and (4.23). Due to its high computational complexity, we do not include multi-frequency [100, 134] TL-SBL. We examine the performance by increasing the number of frequencies processed as $F = 1, 3, 5,$ and 7 with corresponding frequencies 1600, $\{1400, 1600, 1800\}$, $\{1000, 1200, 1400, 1600, 1800\}$, and $\{1000, 1200, 1400, 1600, 1800, 2000, 2200\}$. Figure 4.18 shows that as the number of frequencies increases, the performance improves. The TL-NOMP offers the best performance among all and significantly improves over TL-OMP. TL-SFW shows degraded performance when more frequencies are used, which could be due to the additional amplitude parameters it has to estimate as the number of frequencies increases.

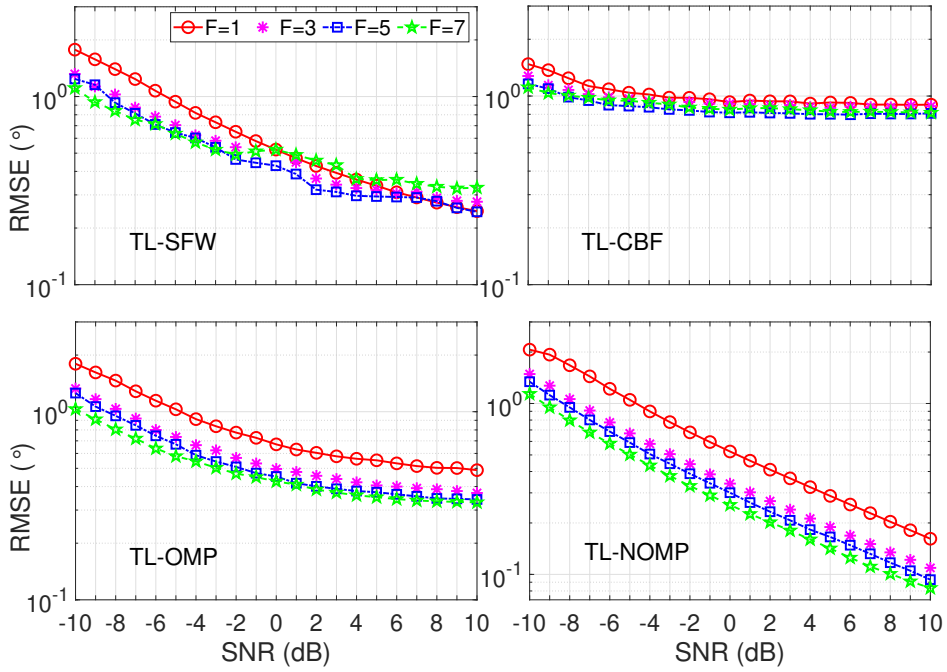

Figure 4.18 Performance of multi-frequency TL-methods for quadratic trajectories with different numbers of processed frequencies at various SNR.

4.5.10 Results on LOCATA

We apply the methods developed in this chapter to LOCATA experiment data [96]. A detailed description for the comprehensive understanding of the array configurations and recording conditions is provided in chapter 3. We create ground truth data for each snapshot to analyze the trajectory models and algorithms. During our dataset analysis, we noticed significant changes in the DOA within each block, enabling trajectory localization models to capture these variations. We consider audio signals from the dicit array in Task 4 (recording 2) with two moving sources. Data from a 7-sensor linear subarray of dicit with $d = 32$ cm inter-sensor spacing is processed (11 frequencies ranging from 200 to 450 Hz are utilized). Figure 4.19 shows the trajectory estimates for the two sources in a block consist-

ing of 30 snapshots (0.1950 s duration). The result shows that the proposed TL model can handle the real-world complexity and capture the DOAs changing within the block, hence improving the overall localization accuracy. The proposed algorithms do not explicitly model reverberations and structured noises that exist in real-world measurements. Thus, their performance may degrade when applied to realistic acoustic environments.

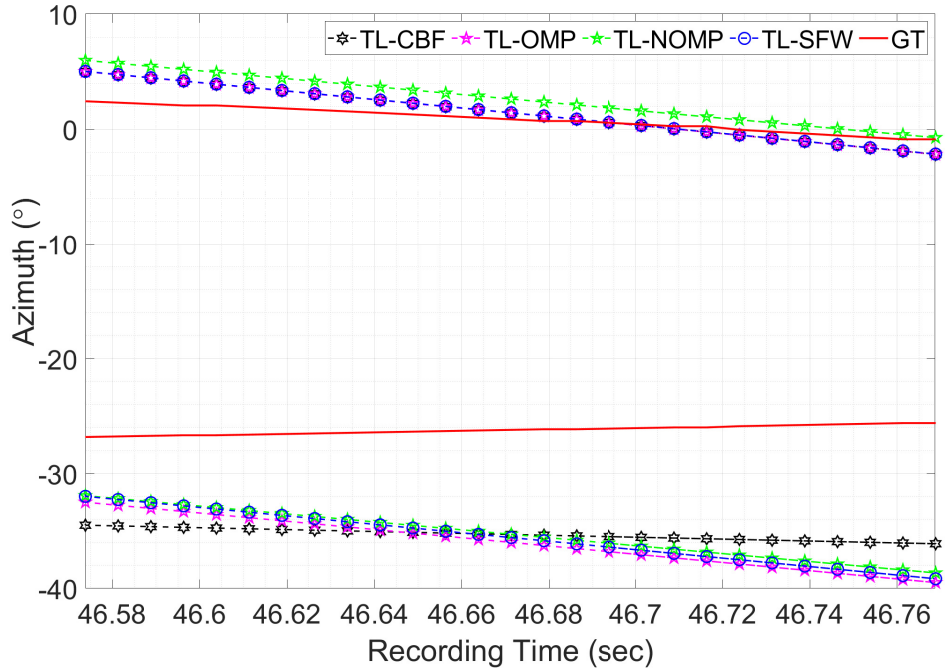

Figure 4.19 Trajectory estimates of two moving sources using dicit array, Task 4, recording 2 from LOCATA. Here, GT is ground truth

Algorithm	Noise resilience	Resolution	Effect of grid-step	Computation speed	Detection probability
TL-CBF	Low	Low	High	Fast	Low
TL-SBL	Medium	High	High	Slow	Medium
TL-OMP	High	Low	High	Fast	High
TL-NOMP	High	High	Low	Medium	Medium
TL-SFW	High	Medium	Medium	Fast	High

Table 4.1 Comparative analysis of various algorithms for trajectory localization.

4.6 Summary

In this chapter, we proposed two novel trajectory models: the harmonic and polynomial models. We developed TL-CBF, TL-OMP, and TL-SBL algorithms for estimating the trajectories. To overcome

the limitation of grid-based algorithms, we proposed two gridless algorithms for localizing the DOA trajectories – TL-SFW and TL-NOMP – and demonstrated their superior performance in extensive simulations. We also extended the algorithms for multi-frequency processing. The proposed models and algorithms are also validated using recordings from real-world LOCATA dataset. Table 4.1 summarizes the performance characteristics of various algorithms, highlighting their noise resilience, resolution, sensitivity to grid-step, speed, and detection probability. Among grid-based methods, TL-CBF and TL-OMP are fast but have low to moderate resolution, whereas TL-SBL is slow but has high resolution. Among the gridless methods, TL-SFW is preferable in scenarios where noise resilience, computational efficiency, and detection rate are prioritized. At the same time, TL-NOMP is more suitable for applications that require noise resilience, and high resolution and coarse parameter grids are tolerable. Overall, gridless algorithms outperform grid-based methods for trajectory localization.

Chapter 5

Conclusion

This thesis provides valuable insights into localization algorithms through comprehensive analysis of real-world datasets. The analysis of various localization algorithms (CBF, MUSIC, and SBL) and exploration of wideband models (Model1, Model2, and Model3) help in enhancing the understanding of localization algorithms and their practical applicability. The analysis on the LOCATA dataset shows that compressive sensing based SBL is a promising method and robust to the challenges posed in real-world scenarios. For task 1, 3 and 5, SBL consistently outperforms CBF and MUSIC across various tasks, establishing SBL as a robust approach in challenging scenarios.

Building on these findings, the thesis addresses wideband DOA estimation using SBL algorithms, evaluating a realistic signal model that accounts for changes in source variance across the frequency range. By applying and evaluating three wideband SBL variants (SBL1, SBL2, and SBL3), along with wideband versions of CBF and MUSIC, it is shown that SBL3, which incorporates a shared colored spectrum, performs best across different signal models and array configurations. This work enhances the understanding of wideband SBL algorithms and their applicability in real-world scenarios.

The thesis further explores the integration of Deep Neural Network based methods, highlighting the importance of predicting DOA derivatives alongside DOA for improving localization performance. A new model combining DOA and their derivatives is proposed, demonstrating improved performance under low signal-to-noise ratio (SNR) conditions using the TAU-NIGENS Spatial Sound Events 2021 dataset. This study underscores the significance of incorporating higher-order derivatives in sound event localization and detection tasks.

Additionally, the thesis introduces novel parametric signal models, such as polynomial and bandlimited models, to identify DOA trajectories that capture the dynamic motion of a source. Grid-based and gridless algorithms are developed to estimate these trajectories, with gridless methods like Sliding Frank-Wolfe and Newtonized Orthogonal Matching Pursuit overcoming the limitations of grid-based approaches by estimating parameters in continuous trajectory space. This improves localization accuracy and eliminates the need for tracking filters. The research also extends to wideband processing, providing detailed results on SNR, number of snapshots, resolution limits, grid step size, and computational complexity. The proposed algorithms are applied to challenging real-world recordings from the LOCATA dataset, which include challenging acoustic scenarios.

In future, the work in this thesis can further analyze the applicability of proposed algorithms in diverse acoustic scenarios accounting real-world implementation challenges such as near-field sources, ambient noise, and different reverberation effect. The TL-algorithms can be generalized to different data modalities to assess the efficacy of proposed model and algorithms. Additionally, a Cramér-Rao Bound (CRB) analysis can be conducted for the proposed TL models to evaluate their theoretical performance limits.

Moreover, the contributions of this thesis extend beyond sound waves and can be applied to other data types, including radio waves. This versatility significantly enhances the impact of the research across various applications. The outcomes of this research have the potential to contribute to a safer and more interconnected world, where smart devices can effectively perceive and track sources, thereby improving performance in numerous practical applications.

Bibliography

[1] L. Wan, G. Han, L. Shu, S. Chan, and T. Zhu, "The application of DOA estimation approach in patient tracking systems with high patient density," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 6, pp. 2353–2364, 2016. (Cited on page 1.)

[2] G. Han, L. Wan, L. Shu, and N. Feng, "Two novel DOA estimation approaches for real-time assistant calibration systems in future vehicle industrial," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1361–1372, 2015. (Cited on page 1.)

[3] B. Alenljung, J. Lindblom, R. Andreasson, and T. Ziemke, "User experience in social human-robot interaction," in *Rapid automation: Concepts, methodologies, tools, and applications*. IGI Global, 2019, pp. 1468–1490. (Cited on page 1.)

[4] S. Birchfield and D. Gillmor, "Acoustic source direction by hemisphere sampling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings*, vol. 5. IEEE, 2001, pp. 3053–3056. (Cited on page 1.)

[5] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor UAVs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8618–8628, 2019. (Cited on page 1.)

[6] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to "informed" sound source localization for hearing aid applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 16–20. (Cited on page 1.)

[7] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466. (Cited on pages 2 and 19.)

[8] E. Fishler, A. Haimovich, R. Blum, D. Chizhik, L. Cimini, and R. Valenzuela, "Mimo radar: an idea whose time has come," in *IEEE Radar Conference*, 2004, pp. 71–78. (Cited on page 2.)

[9] D. Grimes and T. O. Jones, "Automotive radar: A brief review," *Proceedings of the IEEE*, vol. 62, no. 6, pp. 804–822, 1974. (Cited on page 2.)

[10] X. Zeng, B. Wang, and K. J. R. Liu, "Driver arrival sensing for smart car using wifi fine time measurements," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 41–45. (Cited on page 2.)

[11] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 59–80, 2018. (Cited on page 2.)

[12] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, P. N. H. Barfuss and, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array Multichannel Signal Processing Workshop*, 2018, pp. 410–414. (Cited on pages 3, 23, 24, 34, and 41.)

[13] S. Jaiswal, R. Pandey, and S. Nannuru, "Deep architecture for DOA trajectory localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. (Cited on pages 4 and 47.)

[14] H. L. V. Trees, *Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)*. John Wiley & Sons, 2002. (Cited on pages 5, 8, and 46.)

[15] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1. (Cited on pages 5 and 6.)

[16] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*. Springer, 2001, pp. 157–180. (Cited on pages 5 and 12.)

[17] A. Xenaki, P. Gerstoft, and K. Mosegaard, "Compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 260–271, 2014. (Cited on pages 5, 9, 13, 14, and 51.)

[18] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Birkhäuser, New York, NY, 2013. (Cited on pages 5, 12, 13, and 14.)

[19] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022. (Cited on pages 5, 18, and 19.)

[20] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, 2013. (Cited on pages 5, 20, and 57.)

[21] B. Mamandipoor, D. Ramasamy, and U. Madhow, "Newtonized orthogonal matching pursuit: Frequency estimation over the continuum," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5066–5081, 2016. (Cited on pages 5, 20, 21, 57, and 59.)

[22] Y. D. Castro and F. Gamboa, "Exact reconstruction using beurling minimal extrapolation," *Journal of Mathematical Analysis and applications*, vol. 395, no. 1, pp. 336–354, 2012. (Cited on pages 5, 21, 57, and 58.)

[23] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996. (Cited on pages 5 and 46.)

[24] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, 1988. (Cited on pages 8, 9, 46, and 51.)

[25] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas Propagation*, vol. 34, no. 3, pp. 276–280, 1986. (Cited on pages 9 and 46.)

[26] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics Speech Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976. (Cited on page 10.)

[27] K. M. Varma, "Time delay estimate based direction of arrival estimation for speech in reverberant environments," Ph.D. dissertation, Virginia Tech, 2002. (Cited on page 11.)

[28] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000. (Cited on pages 11 and 46.)

[29] H. Do, F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2007, pp. I–121. (Cited on page 12.)

[30] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006. (Cited on page 12.)

[31] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. (Cited on page 13.)

[32] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010. (Cited on page 13.)

[33] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008. (Cited on page 13.)

[34] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008. (Cited on page 13.)

[35] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005. (Cited on pages 13 and 14.)

[36] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001. (Cited on pages 13 and 14.)

[37] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007. (Cited on pages 14 and 15.)

[38] A. Aich and P. Palanisamy, "On-grid DOA estimation method using orthogonal matching pursuit," in *2017 International Conference on Signal Processing and Communication (ICSPC)*. IEEE, 2017, pp. 483–487. (Cited on page 15.)

[39] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, 2004. (Cited on page 15.)

[40] D. Wipf, J. Palmer, and B. D. Rao, "Perspectives on sparse Bayesian learning," *Computer Engineering*, vol. 16, no. 1, p. 249, 2004. (Cited on page 15.)

[41] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Process*, vol. 55, no. 7, pp. 3704–3716, 2007. (Cited on pages 17, 47, and 52.)

[42] Z. Liu, Z. Huang, and Y. Zhou, "An efficient maximum likelihood method for direction-of-arrival estimation via sparse Bayesian learning," *IEEE Transactions on Wireless Commun.*, vol. 11, no. 10, pp. 1–11, 2012. (Cited on pages 17 and 29.)

[43] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Jun. 2001. (Cited on pages 17, 30, 50, and 52.)

[44] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Signal Processing Letter*, vol. 23, no. 10, pp. 1469–1473, Oct. 2016. (Cited on pages 17, 30, 31, 32, 47, and 61.)

[45] K. L. Gemba, S. Nannuru, P. Gerstoft, and W. S. Hodgkiss, "Multi-frequency sparse Bayesian learning for robust matched field processing," *Journal of Acoustical Society of America*, vol. 141, no. 5, pp. 3411–3420, 2017. (Cited on pages 17, 29, 30, and 51.)

[46] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Multi-frequency sparse Bayesian learning for matched field processing in non-stationary noise," *Journal of Acoustical Society of America*, vol. 144, no. 3, pp. 1943–1943, 2018. (Cited on pages 17, 25, 29, and 30.)

[47] P. Gerstoft, A. Xenaki, and C. F. Mecklenbräuker, "Multiple and single snapshot compressive beamforming," *Journal of Acoustical Society of America*, vol. 138, no. 4, pp. 2003–2014, 2015. (Cited on page 17.)

[48] D. P. Wipf and B. D. Rao, "Bayesian learning for sparse signal reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6. IEEE, 2003, pp. VI–601. (Cited on page 17.)

[49] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 29, pp. 300–311, 2020. (Cited on pages 18 and 46.)

[50] R. Opochinsky, G. Chechik, and S. Gannot, "Deep ranking-based DOA tracking algorithm," in *29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1020–1024. (Cited on pages 18 and 46.)

[51] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018. (Cited on pages 18, 19, and 41.)

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. (Cited on page 18.)

[53] S. Chakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 136–140. (Cited on page 18.)

[54] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019. (Cited on pages 18, 19, and 41.)

[55] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818. (Cited on pages 18 and 19.)

[56] D. Suvorov, G. Dong, and R. Zhukov, "Deep residual network for sound source localization in the time domain," *arXiv preprint arXiv:1808.06429*, 2018. (Cited on pages 18 and 19.)

[57] Q. Nguyen, L. Girin, G. Bailly, F. Elisei, and D. C. Nguyen, "Autonomous sensorimotor learning for sound source localization by a humanoid robot," in *Workshop on Crossmodal Learning for Intelligent Robot. in conj. with IEEE/RSJ IROS*, 2018. (Cited on pages 18 and 19.)

[58] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, *On sound source localization of speech signals using deep neural networks*. Technische Universität Berlin, 2019. (Cited on page 19.)

[59] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015. (Cited on page 19.)

[60] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 451–455. (Cited on page 19.)

[61] N. Ma, G. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Interspeech*, vol. 2015. International Speech Communication Association, 2015, pp. 160–164. (Cited on page 19.)

[62] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5. (Cited on page 19.)

[63] A. Fahim, P. Samarasinghe, and T. Abhayapala, "Multi-source DOA estimation through pattern recognition of the modal coherence of a reverberant soundfield," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 605–618, 2019. (Cited on page 19.)

[64] V. Varanasi, H. Gupta, and R. M. Hegde, "A deep learning framework for robust DOA estimation using spherical harmonic decomposition," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 1248–1259, 2020. (Cited on page 19.)

[65] G. L. Moing, P. Vinayavekhin, D. J. Agravante, T. Inoue, J. Vongkulbhisal, A. Munawar, and R. Tachibana, "Data-efficient framework for real-world multiple sound source 2d localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3425–3429. (Cited on page 19.)

[66] W. Ma and X. Liu, "Phased microphone array for sound source localization with deep learning," *Aerospace Systems*, vol. 2, no. 2, pp. 71–81, 2019. (Cited on page 19.)

[67] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Computer Speech & Language*, vol. 75, p. 101360, 2022. (Cited on page 19.)

[68] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245. (Cited on page 19.)

[69] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv:2106.06999*, 2021. (Cited on page 19.)

[70] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and D. M. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 885–889. (Cited on pages 19 and 20.)

[71] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6. (Cited on page 19.)

[72] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 915–919. (Cited on pages 19 and 20.)

[73] T. Nguyen, K. Watcharasupat, N. Nguyen, D. Jones, and W. Gan, "SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 30, pp. 1749–1762, 2022. (Cited on pages 19, 20, and 38.)

[74] T. Nguyen, D. Jones, K. Watcharasupat, H. Phan, and W. Gan, "SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 716–720. (Cited on pages 19, 20, 38, 40, 41, 42, and 43.)

[75] H. Phan, L. Pham, P. Koch, N. Q. Duong, I. McLoughlin, and A. Mertins, "On multitask loss function for audio event detection and localization," *arXiv preprint arXiv:2009.05527*, 2020. (Cited on page 20.)

[76] A. Xenaki and P. Gerstoft, "Grid-free compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 1923–1935, 2015. (Cited on pages 20 and 57.)

[77] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, 2013. (Cited on pages 20 and 57.)

[78] W. Xu, J. F. Cai, V. K. Mishra, M. Cho, and A. Kruger, "Precise semidefinite programming formulation of atomic norm minimization for recovering d-dimensional $(d \geq 2)$ off-the-grid frequencies," in *Information Theory and Applications Workshop (ITA)*. IEEE, 2014, pp. 1–4. (Cited on pages 20 and 57.)

[79] Y. Chi and Y. Chen, "Compressive two-dimensional harmonic retrieval via atomic norm minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1030–1042, 2014. (Cited on pages 20 and 57.)

[80] Y. Yang, Z. Chu, Z. Xu, and G. Ping, "Two-dimensional grid-free compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 618–629, 2017. (Cited on pages 20 and 57.)

[81] Y. Yang, Z. Chu, G. G. Ping, and Z. Xu, "Resolution enhancement of two-dimensional grid-free compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3860–3872, 2018. (Cited on pages 20 and 57.)

[82] Y. Zhang, Y. Wang, Z. Tian, G. Leus, and G. Zhang, "Efficient super-resolution two-dimensional harmonic retrieval with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1224–1240, 2022. (Cited on pages 20 and 57.)

[83] Z. Yang, L. Xie, and P. Stoica, "Vandermonde decomposition of multilevel toeplitz matrices with application to multidimensional super-resolution," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3685–3701, 2016. (Cited on pages 20 and 57.)

[84] X. Wu, Z. Yang, P. Stoica, and Z. Xu, "Maximum likelihood line spectral estimation in the signal domain: A rank-constrained structured matrix recovery approach," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4156–4169, 2022. (Cited on pages 20 and 57.)

[85] Z. Yang and L. Xie, "On gridless sparse methods for multi-snapshot DOA estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3236–3240. (Cited on page 20.)

[86] S. Semper, F. Roemer, T. Hotz, and G. D. Galdo, "Grid-free direction-of-arrival estimation with compressed sensing and arbitrary antenna arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3251–3255. (Cited on pages 20 and 57.)

[87] M. Wagner, Y. Park, and P. Gerstoft, "Gridless DOA estimation and root-music for non-uniform linear arrays," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2144–2157, 2021. (Cited on pages 20 and 57.)

[88] Y. Wu, M. B. Wakin, and P. Gerstoft, "Gridless DOA estimation with multiple frequencies," *arXiv preprint arXiv:2207.06159*, 2022. (Cited on pages 20 and 57.)

[89] Y. Jiang, D. Li, X. Wu, and W. P. Zhu, "A gridless wideband DOA estimation based on atomic norm minimization," in *Sensor Array and Multimedia Signal Processing Workshop (SAM)*. IEEE, 2020, pp. 1–5. (Cited on pages 20 and 57.)

[90] Y. Y. Ang, N. Nguyen, and W. S. Gan, "Multiband grid-free compressive beamforming," *Mechanical Systems and Signal Processing*, vol. 135, p. 106425, 2020. (Cited on pages 20 and 57.)

[91] G. Chardon and U. Boureau, "Gridless three-dimensional compressive beamforming with the sliding frank-wolfe algorithm," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 3139–3148, 2021. (Cited on pages 20, 22, 57, and 58.)

[92] Y. Yang, Z. Chu, Y. Yang, and S. Yin, "Two-dimensional newtonized orthogonal matching pursuit compressive beamforming," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1337–1348, 2020. (Cited on pages 21 and 59.)

[93] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The sliding frank–wolfe algorithm and its application to super-resolution microscopy," *Inverse Problems*, vol. 36, no. 1, p. 014001, 2019. (Cited on pages 21, 22, 57, 58, and 59.)

[94] C. Evers, H. W. Löllmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020. (Cited on pages 23, 24, and 26.)

[95] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. Naylor, and W. Kellermann, "LOCATA challenge-evaluation tasks and measures," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 565–569. (Cited on pages 23 and 24.)

[96] C. Evers *et al.*, "Data Corpus for the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA)," Jan. 2020. (Cited on pages 24 and 70.)

[97] H. W. Lollmann *et al.*, "IEEE-AASP Challenge on Source Localization and Tracking: Documentation for Participants," Apr. 2018. [Online]. Available: www.locata-challenge.org (Cited on page 24.)

[98] M. V. Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice." in *INTERSPEECH*, 2013, pp. 704–708. (Cited on pages 24, 25, and 34.)

[99] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust ocean acoustic localization with sparse Bayesian learning," *IEEE Journal of Selected Topics Signal Processing*, vol. 13, no. 1, pp. 49–60, 2019. (Cited on pages 25, 29, and 51.)

[100] S. Nannuru, K. L. Gemba, P. Gerstoft, W. S. Hodgkiss, and C. F. Mecklenbräuker, "Sparse Bayesian learning with multiple dictionaries," *Signal Processing*, vol. 159, pp. 159–170, 2019. (Cited on pages 29, 30, 31, 32, 51, 61, and 70.)

[101] S. Nannuru, P. Gerstoft, G. Ping, and E. Fernandez-Grande, "Sparse planar arrays for azimuth and elevation using experimental data," *Journal of Acoustical Society of America*, vol. 159, no. 1, pp. 167–178, 2021. (Cited on pages 29, 30, 31, and 51.)

[102] J. Bohme, "Source-parameter estimation by approximate maximum likelihood and nonlinear regression," *IEEE Journal of Oceanic Engineering*, vol. 10, no. 3, pp. 206–212, 1985. (Cited on pages 32 and 61.)

[103] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020. (Cited on page 38.)

[104] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "Xsleepnet: Multiview sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2021. (Cited on page 40.)

[105] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 40.)

[106] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337. (Cited on page 41.)

[107] R. Pandey, S. Nannuru, and P. Gerstoft, "Experimental validation of wideband SBL models for DOA estimation," in *European Signal Processing Conference (EUSIPCO)*, 2022, pp. 219–223. (Cited on page 41.)

[108] IEEE AASP Challenge, "Detection and classification of acoustic scenes and events (DCASE)." [Online]. Available: https://dcase.community/challenge2020/task-sound-event-localization-and-detection#metrics (Cited on page 41.)

[109] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1684–1696, 2005. (Cited on page 46.)

[110] R. Pandey, S. Nannuru, and A. Siripuram, "Sparse Bayesian learning for acoustic source localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 4670–4674. (Cited on page 46.)

[111] Y. Park, F. Meyer, and P. Gerstoft, "Sequential sparse Bayesian learning for time-varying direction of arrival," *Journal of Acoustical Society of America*, vol. 149, no. 3, pp. 2089–2099, 2021. (Cited on page 46.)

[112] R. Pandey and S. Nannuru, "Parametric models for DOA trajectory localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022. (Cited on pages 47, 50, and 51.)

[113] S. Peleg and B. Porat, "Estimation and classification of polynomial-phase signals," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 422–430, 1991. (Cited on page 48.)

[114] B. Völcker, "Performance analysis of parametric spectral estimators," Ph.D. dissertation, Signaler, sensorer och system, 2002. (Cited on page 48.)

[115] A. W. Rihaczek, "Principles of high-resolution radar," *(No Title)*, 1969. (Cited on page 48.)

[116] B. Porat, *Digital processing of random signals: Theory and methods*. Courier Dover Publications, 2008. (Cited on page 48.)

[117] S. Peleg and B. Porat, "The cramer-rao lower bound for signals with constant amplitude and polynomial phase," *IEEE Transactions on Signal Processing*, vol. 39, no. 3, pp. 749–752, 1991. (Cited on page 48.)

[118] G. Zhou, G. Giannakis, and A. Swami, "On polynomial phase signals with time-varying amplitudes," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 848–861, 1996. (Cited on page 48.)

[119] M. Adjrad and A. Belouchrani, "Estimation of multicomponent polynomial-phase signals impinging on a multisensor array using state–space modeling," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 32–45, 2006. (Cited on page 48.)

[120] M. M. Ghogho, A. Nandi, and A. Swami, "Cramer-rao bounds and maximum likelihood estimation for random amplitude phase-modulated signals," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 2905–2916, 1999. (Cited on page 48.)

[121] B. Friedlander and J. M. Francos, "Estimation of amplitude and phase parameters of multicomponent signals," *IEEE Transactions on Signal Processing*, vol. 43, no. 4, pp. 917–926, 1995. (Cited on page 48.)

[122] Y. Zhang, W. Mu, and M. G. Amin, "Time–frequency maximum likelihood methods for direction finding," *Journal of the Franklin Institute*, vol. 337, no. 4, pp. 483–497, 2000. (Cited on page 48.)

[123] M. Adjrad, A. Beloucharni, and A. Ouldali, "Estimation of chirp signal parameters using state space modelization by incorporating spatial information," in *International Symposium on Signal Processing and Its Applications.*, vol. 2. IEEE, 2003, pp. 531–534. (Cited on page 48.)

[124] P. Tabaghi, I. Dokmanić, and M. Vetterli, "Kinetic euclidean distance matrices," *IEEE Transactions on Signal Processing*, vol. 68, pp. 452–465, 2019. (Cited on page 49.)

[125] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, 2011. (Cited on pages 50, 51, and 52.)

[126] ——, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2009–2015, 2013. (Cited on pages 50, 51, and 52.)

[127] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977. (Cited on page 52.)

[128] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993. (Cited on page 52.)

[129] T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011. (Cited on page 52.)

[130] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999. (Cited on page 59.)

[131] G. Chardon, "gilleschardon/sfwcb: Sfwcb 1.1," Sep. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5528801 (Cited on page 60.)

[132] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008. (Cited on page 61.)

[133] B. Ristic, B.-N. Vo, D. Clark, and B. T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3452–3457, 2011. (Cited on page 61.)

[134] R. Pandey, S. Nannuru, and P. Gerstoft, "Experimental validation of wideband sbl models for DOA estimation," in *30th European Signal Processing Conference*. IEEE, 2022, pp. 219–223. (Cited on page 70.)