



Signal Processing
Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
<https://sps.ewi.tudelft.nl/>

SPS-2026-6181821

M.Sc. Thesis

A Covariance-Based Simultaneous Orthogonal Matching Pursuit Framework for Acoustic Parameter Estimation and Beamforming

Kunal Kashyap, B.Tech.

Abstract

Binaural speech enhancement in hearing aids requires accurate estimates of the acoustic transfer functions (ATFs) of the active sources and the noise covariance matrix to drive spatial filters that simultaneously suppress interference and preserve binaural cues. Existing methods estimate these quantities sequentially, introducing error propagation and a dependency on noise-only intervals that limits robustness in adverse acoustic conditions. This thesis proposes a covariance-domain extension of Simultaneous Orthogonal Matching Pursuit that jointly estimates the active head-related transfer function (HRTF) subspace, source signal energies, and noise covariance from the observed cross-power spectral density matrix within a single iterative procedure. Controlled simulation experiments against a GEVD-based baseline demonstrate consistently lower HRTF estimation error across a wide range of SNR levels, time block sizes, and source azimuth configurations, with particular advantage at low SNR and short block lengths. When used to drive a joint binaural LCMV beamformer, the framework yields near-zero and stable target interaural time difference errors across the full SNR range, confirming that the joint estimation strategy produces internally consistent beamformer inputs that support robust binaural cue preservation.



A Covariance-Based Simultaneous Orthogonal Matching Pursuit Framework for Acoustic Parameter Estimation and Beamforming

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING

by

Kunal Kashyap, B.Tech.
born in Noida, India

This work was performed in:

Signal Processing Systems Group
Department of Microelectronics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology



Delft University of Technology

Copyright © 2026 Signal Processing Systems Group
All rights reserved.

DELFT UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF
MICROELECTRONICS

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled “**A Covariance-Based Simultaneous Orthogonal Matching Pursuit Framework for Acoustic Parameter Estimation and Beamforming**” by **Kunal Kashyap, B.Tech.** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: July 2, 2026

Chairman:

dr.ir. R.C. Hendriks

Advisors:

ir. J.W. de Vries

dr.ir. R.C. Hendriks

Committee Members:

dr. F. Fioranelli

Abstract

Binaural speech enhancement in hearing aids requires accurate estimates of the acoustic transfer functions (ATFs) of the active sources and the noise covariance matrix to drive spatial filters that simultaneously suppress interference and preserve binaural cues. Existing methods estimate these quantities sequentially, introducing error propagation and a dependency on noise-only intervals that limits robustness in adverse acoustic conditions. This thesis proposes a covariance-domain extension of Simultaneous Orthogonal Matching Pursuit that jointly estimates the active head-related transfer function (HRTF) subspace, source signal energies, and noise covariance from the observed cross-power spectral density matrix within a single iterative procedure. Controlled simulation experiments against a GEVD-based baseline demonstrate consistently lower HRTF estimation error across a wide range of SNR levels, time block sizes, and source azimuth configurations, with particular advantage at low SNR and short block lengths. When used to drive a joint binaural LCMV beamformer, the framework yields near-zero and stable target interaural time difference errors across the full SNR range, confirming that the joint estimation strategy produces internally consistent beamformer inputs that support robust binaural cue preservation.

Acknowledgments

The completion of this thesis marks the end of an important chapter in my academic journey, and I would like to express my sincere gratitude to the many people who have supported me along the way.

First and foremost, I would like to thank my supervisors, ir. J.W. de Vries and dr.ir. R.C. Hendriks, for their guidance, encouragement, and invaluable feedback throughout this project. Their expertise, patience, and willingness to discuss ideas have been instrumental in shaping both this thesis and my growth as a researcher. I am grateful for the opportunity to learn from them and for the trust they placed in me during this work.

I would also like to thank Ma and Pa for their unwavering love, support, and belief in me. Their sacrifices, encouragement, and confidence have made it possible for me to pursue my goals, even when they seemed far away. No matter the distance, their support has always been a constant source of strength and motivation.

To Mugdha, my girlfriend, thank you for your patience, understanding, and companionship throughout this journey. Your encouragement during challenging times and your ability to celebrate every small success made this experience far more meaningful and enjoyable. I am deeply grateful to have shared this chapter of my life with you.

I am also thankful to the friends I made during my master's studies. The countless discussions, group projects, study sessions, coffee breaks, and moments of laughter helped make this experience memorable. Beyond academics, your friendship turned challenges into shared experiences and made Delft feel like home.

Finally, I would like to thank everyone who has been part of my life since I moved to the Netherlands. Whether through friendship, support, advice, or simply shared moments, each of you has contributed to my personal and professional growth in ways both large and small. Moving to a new country is never easy, but the kindness and connections I found here made the journey truly rewarding.

As I conclude this thesis and my master's degree, I look back with gratitude for all the experiences, opportunities, and people who have shaped these years. This achievement is not mine alone; it reflects the support, encouragement, and generosity of everyone mentioned above.

Thank you.

Kunal Kashyap, B.Tech.
Delft, The Netherlands
July 2, 2026

Contents

Abstract	v
Acknowledgments	vii
Nomenclature	xv
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	1
1.3 Contributions	2
1.4 Thesis Outline	3
2 Background	5
2.1 Binaural Hearing and Spatial Audio	5
2.2 Acoustic Signal Model	6
2.3 Head-Related Transfer Functions	9
2.4 The HRTF Dictionary Framework	10
2.5 Binaural Beamforming	13
2.6 Sparse Signal Recovery	16
3 Related Work	19
3.1 ATF and RTF Estimation	19
3.2 Noise Covariance Estimation	21
3.3 Dictionary-Based Spatial Processing	22
3.4 Sparse Recovery for Source Localization	23
3.5 Joint and Unified Estimation Frameworks	25
3.6 Gap Analysis and Positioning	27
4 Covariance-Based Simultaneous Orthogonal Matching Pursuit	31
4.1 Inputs, Outputs and Design Choices	31
4.2 Proposed Algorithm	32
4.3 Atom Selection	33
4.4 Signal Energy Update	35
4.5 Residual Update	36
4.6 Stopping Criterion	38
4.7 Computational Complexity	40
5 Experimental Evaluations	43
5.1 Experimental Setup	43
5.2 Experiment 1: HRTF Estimation Accuracy	47
5.2.1 Effect of SNR	47
5.2.2 Effect of Time Block Size	48

5.2.3	Effect of Source Azimuth Configuration	49
5.3	Experiment 2: Binaural Cue Preservation	51
5.3.1	ILD Preservation	51
5.3.2	ITD Preservation	52
5.3.3	Binaural Cue Asymmetry	53
5.4	Discussion	53
5.5	Limitations	55
6	Conclusion	59
6.1	Summary of Contributions	59
6.2	Conclusions	60
6.3	Future Work	60
	References	63
A	Derivations	69
A.1	The Signal Energy Update Closed-Form Solution	69
A.2	The Projection Matrix and Its Properties	69
B	Preliminary Experiment	71
C	Code Snippets	75
C.1	Covariance-Based Simultaneous Orthogonal Matching Pursuit	75
C.2	Joint Binaural Linearly Constrained Minimum Variance (JBLCMV) Beamformer	77
D	Submitted Paper	79

List of Figures

2.1	Example HRTF dictionary with $N = 24$ virtual source locations at zero elevation, uniformly distributed in azimuth.	11
4.1	A beamforming framework using the proposed algorithm, offering an alternative to sequential estimation of quantities required by a beamformer.	33
5.1	BTE microphone placement with two sources at 0° and 30°	44
5.2	HRTF estimation error under varying SNR and time block size.	48
5.3	HRTF estimation error vs. source azimuth (S_1 fixed at 0° , SNR = 30 dB, TBS = 50).	50
5.4	Binaural cue preservation errors for S_1 and S_2 under varying SNR.	51
B.1	Virtual source tracking across time blocks.	72

List of Tables

3.1	Comparison of existing methods against the five properties of an ideal joint estimation framework. ✓: satisfied. ○: partially satisfied. ✗: not satisfied.	28
4.1	Per-iteration cost	42

Nomenclature

Abbreviations and Acronyms

AIC	An information theoretic criterion
ATF	Acoustic Transfer Function
BASE	Baseline method/algorithm
BLCMV	Binaural Linearly Constrained Minimum Variance
BTE	Behind-The-Ear
CPSDM	Cross-Power Spectral Density Matrix
DOA	Direction of Arrival
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Techniques
FFT	Fast Fourier Transform
GEVD	Generalized Eigenvalue Decomposition
HA	Hermitian Angle
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
ITF	Interaural Transfer Function
JBLCMV	Joint Binaural Linearly Constrained Minimum Variance
LCMV	Linearly Constrained Minimum Variance
MDL	Minimum Description Length
MMV	Multiple Measurement Vector
MUSIC	Multiple Signal Classification
MVDR	Minimum Variance Distortionless Response
MWF	Multichannel Wiener Filter
OMP	Orthogonal Matching Pursuit
PROP	Proposed method/algorithm
PSD	Power Spectral Density
RATF	Relative Acoustic Transfer Function
RTF	Relative Transfer Function
SNR	Signal-to-Noise Ratio
S-OMP	Simultaneous Orthogonal Matching Pursuit
STFT	Short-Time Fourier Transform
TBS	Time Block Size
VAD	Voice Activity Detection

Mathematical Notation

Scalars

d	Number of active sources
k	Frequency bin index, $k \in \{1, \dots, K\}$
K	Total number of frequency bins
l	STFT time frame index
L	Time block size (number of STFT frames)
M	Number of microphone channels
N	Number of atoms in the HRTF dictionary
p	Pursuit iteration index
$s_i(k, l)$	STFT coefficient of source i at bin k , frame l
$\sigma_{s_i}^2(k)$	Power spectral density of source i at bin k
$\mu(\cdot)$	Mutual coherence of a dictionary
ϕ	Source azimuth angle
θ	Source elevation angle

Vectors

$\mathbf{a}_i(k)$	ATF vector of source i at frequency bin k
$\mathbf{c}^{(p)}(k)$	Per-bin atom correlation score vector at iteration p
$\mathbf{c}_{agg}^{(p)}$	Frequency-aggregated atom correlation score vector at iteration p
$\mathbf{h}_n(k)$	n^{th} HRTF dictionary atom at frequency bin k
$\tilde{\mathbf{h}}_n(k)$	Normalized n^{th} HRTF dictionary atom at bin k
$\mathbf{n}(k, l)$	Noise vector at bin k , frame l
$\mathbf{s}(k, l)$	Source signal vector at bin k , frame l
$\mathbf{x}(k, l)$	Observed microphone signal vector at bin k , frame l
$\mathbf{y}(k, l)$	Desired signal component at bin k , frame l

Matrices

$\mathbf{A}(k)$	ATF matrix at frequency bin k , $\mathbf{A}(k) \in \mathbb{C}^{M \times d}$
$\mathcal{A}(k)$	HRTF dictionary matrix at bin k , $\mathcal{A}(k) \in \mathbb{C}^{M \times N}$
$\tilde{\mathcal{A}}(k)$	Normalized HRTF dictionary matrix at bin k , $\tilde{\mathcal{A}}(k) \in \mathbb{C}^{M \times N}$
$\mathcal{A}_{\Lambda}(k)$	Sub-dictionary of selected atoms at bin k , $\mathcal{A}_{\Lambda}(k) \in \mathbb{C}^{M \times p}$
$\tilde{\mathcal{A}}_{\Lambda}(k)$	Normalized sub-dictionary of selected atoms at bin k , $\tilde{\mathcal{A}}_{\Lambda}(k) \in \mathbb{C}^{M \times p}$
$\mathbf{P}_{\Lambda}(k)$	Orthogonal projector onto the column space of $\tilde{\mathcal{A}}_{\Lambda}(k)$
$\mathbf{R}_{\mathbf{x}}(k)$	Observed CPSDM at bin k , $\mathbf{R}_{\mathbf{x}}(k) \in \mathbb{C}^{M \times M}$
$\hat{\mathbf{R}}_{\mathbf{x}}(k)$	Sample CPSDM estimate at bin k
$\mathbf{R}_{\mathbf{s}}(k)$	Source signal covariance matrix at bin k , $\mathbf{R}_{\mathbf{s}}(k) \in \mathbb{C}^{d \times d}$

$\hat{\mathbf{R}}_{\mathbf{s}}(k)$	Estimated source signal covariance at bin k
$\mathbf{R}_{\mathbf{n}}(k)$	Noise covariance matrix at bin k , $\mathbf{R}_{\mathbf{n}}(k) \in \mathbb{C}^{M \times M}$
$\mathbf{R}_{\mathbf{y}}(k)$	Signal covariance contribution at bin k
$\mathbf{R}_{res}^{(p)}(k)$	Residual covariance matrix at iteration p and bin k
$\hat{\mathbf{R}}_{\mathbf{y}}^{(p)}(k)$	Estimated explained covariance at iteration p and bin k
$\underline{\mathcal{A}}$	Full HRTF dictionary tensor, $\underline{\mathcal{A}} \in \mathbb{C}^{M \times N \times K}$
$\tilde{\underline{\mathcal{A}}}$	Full normalized HRTF dictionary tensor, $\tilde{\underline{\mathcal{A}}} \in \mathbb{C}^{M \times N \times K}$
$\underline{\mathbf{R}}_{\mathbf{x}}$	Observed CPSDM tensor, $\underline{\mathbf{R}}_{\mathbf{x}} \in \mathbb{C}^{M \times M \times K}$

Sets and Indices

Λ	Support set of active dictionary indices
$\Lambda^{(p)}$	Support set after p iterations of the pursuit
$\lambda^{(p)}$	Dictionary index selected at iteration p
$\mathcal{L}$	Set of STFT frame indices in the current processing block
$\mathcal{L}_{\mathbf{n}}$	Set of noise-only frame indices identified by VAD

Operators and Functions

$(\cdot)^T$	Transpose
$(\cdot)^H$	Hermitian (conjugate) transpose
$(\cdot)^\dagger$	Moore-Penrose pseudoinverse
$(\cdot)^{-1}$	Matrix inverse
$\mathbb{E}[\cdot]$	Expectation operator
$\ \cdot\ _F$	Frobenius norm
$\ \cdot\ _2$	Euclidean (ℓ_2) norm
$\ \cdot\ _0$	ℓ_0 pseudo-norm (number of non-zero entries)
$\text{diag}(\cdot)$	Diagonal extraction operator (extracts main diagonal of a matrix as a vector)
$\text{diag}(\mathbf{v})$	Diagonal matrix with vector $\mathbf{v}$ on its main diagonal
$\text{tr}(\cdot)$	Matrix trace
$\text{vec}(\cdot)$	Vectorization operator
$\otimes$	Kronecker product
$\odot$	Element-wise (Hadamard) product
$\mathcal{F}\{\cdot\}$	Fourier transform
$\arg \max$	Argument of the maximum
$\arg \min$	Argument of the minimum
$\succeq \mathbf{0}$	Positive semi-definite
$\subset$	Subset membership (used for sub-dictionary notation)

Introduction

1.1 Motivation

Hearing loss is among the most prevalent chronic health conditions globally, affecting over 1.5 billion people worldwide and representing a significant burden on individual quality of life and public health systems [39]. For the majority of those affected, hearing aid devices are the primary clinical intervention, amplifying and processing sound to restore speech intelligibility in everyday listening environments. Yet amplification alone is insufficient in the acoustic conditions where hearing difficulty is most acute: noisy, reverberant, multi-talker environments where multiple sound sources compete for the listener’s attention. In these conditions, the fundamental challenge is not loudness but selectivity, i.e., the ability to attend to a desired speaker while suppressing competing sources and background noise.

The human auditory system addresses this challenge through binaural processing, exploiting the small but perceptually significant differences between the signals arriving at the two ears to localize, segregate, and selectively attend to sound sources in space [62, 63]. This spatial hearing capability depends critically on the preservation of binaural cues (particularly interaural time differences and interaural level differences) that encode the directional information of each source. When hearing aid processing distorts these cues, even a user with restored audibility may lose the ability to localize sounds or to exploit spatial separation between sources for improved intelligibility, undermining the perceptual mechanisms that binaural hearing provides [57].

Binaural hearing aid systems therefore face a dual objective: suppressing noise and interference sufficiently to improve speech intelligibility, while preserving the binaural cues of the sources in the acoustic scene sufficiently to maintain the listener’s spatial auditory experience. Meeting both objectives simultaneously is the central design challenge of binaural speech enhancement, and it is the problem that this thesis addresses.

1.2 Problem Statement

Binaural beamforming is the principal signal processing technique for achieving this dual objective. Minimum variance-based binaural beamformers, including the binaural minimum variance distortionless response (BMVDR) beamformer and the joint binaural linearly constrained minimum variance (JBLCMV) beamformer [26, 21], produce output signal pairs for the left and right ears by applying a spatial filter to the microphone array signals. These filters are designed to enhance the target source, suppress interference and noise, and impose constraints that preserve the binaural cues of the active sources at the output. The design of these filters requires accurate knowledge of two quantities: the acoustic transfer function (ATF) vectors of all active sources, which

serve as the steering constraints of the beamformer, and the noise cross-power spectral density matrix (CPSDM), which enters the minimum variance objective function.

In practice, both quantities are unknown and must be estimated from the observed microphone signals. This estimation problem is fundamentally challenging for two reasons. First, the ATF vectors and noise covariance are entangled in the observed CPSDM; in other words, the observation is a sum of signal and noise contributions, and separating them without additional constraints is ill-posed. Second, the acoustic scene is dynamic: sources move, noise statistics change, and the estimates must be updated continuously to track the current acoustic environment. Existing approaches address these challenges by estimating the ATF vectors and noise covariance separately, i.e., ATFs from the observed mixture using subspace methods such as the generalized eigenvalue decomposition [52, 34], and noise covariance from noise-only intervals identified by a voice activity detector [46, 7]. This sequential approach has two structural limitations. First, the noise covariance estimate depends on the availability of noise-only intervals, which may be scarce in continuously active speech scenarios. Second, any inaccuracy in the noise estimate propagates directly into the ATF estimation step, since subspace methods require an accurate noise covariance as input. The resulting estimates are therefore mutually inconsistent, potentially degrading beamformer performance in conditions where robust spatial processing is most needed.

Joint estimation approaches have been explored to address this inconsistency [28, 50, 30], estimating ATF vectors and noise covariance simultaneously from the observed CPSDM. These methods reduce error propagation and produce more consistent estimates, but they estimate ATF vectors as continuous-valued parameters without constraining them to a physically meaningful set. In the binaural hearing aid setting, head-related transfer functions (HRTFs) measured at predetermined directions provide exactly such a constraint: a finite dictionary of physically realizable ATF vectors that captures the directional acoustic filtering of the listener’s anatomy. Doing so simultaneously regularizes the estimation problem, ensures that the recovered vectors are directly usable as beamformer steering constraints, and introduces the structural prior that active sources occupy only a sparse subset of the dictionary directions. Existing dictionary-based methods [25, 24] use the dictionary as a fixed constraint set for the beamformer rather than as a basis for estimation, constraining all dictionary entries simultaneously and thereby consuming all available degrees of freedom for noise reduction. No existing method jointly estimates the active HRTF subspace, the source signal energies, and the noise covariance from the observed CPSDM within a unified framework that exploits the sparsity of the source support and produces beamformer-ready outputs without a separate noise estimation stage.

1.3 Contributions

This thesis proposes a covariance-domain sparse pursuit framework that addresses the joint estimation problem described above. The framework extends Simultaneous Orthogonal Matching Pursuit to the covariance domain, operating on the cross-power spectral density matrix estimated from a short block of STFT frames rather than on instantaneous signal vectors. By replacing the standard inner product matching step

with a quadratic form evaluated against the residual covariance matrix, the algorithm identifies the HRTF vectors stored as dictionary atoms most consistent with the observed spatial covariance structure, one per iteration, deflating the residual after each selection to ensure that previously identified sources do not contaminate subsequent selections. The signal energies of the active sources are estimated at each iteration using the solution to a matrix least-squares problem, and the residual covariance at termination constitutes a noise covariance estimate that requires no separate estimation stage, no VAD, and no noise-only intervals. These three outputs directly supply all inputs required by a binaural beamformer within a single, non-sequential procedure.

The framework is evaluated through controlled simulation experiments against established baselines representing the sequential estimation approach, across a range of SNR levels, time block sizes, and source azimuth configurations. The proposed algorithm achieves consistently lower HRTF estimation error than the GEVD-based baseline across all tested conditions, with particular advantage at low SNR and short block sizes where the covariance domain formulation and frequency-aggregated atom selection provide noise robustness that coefficient-domain methods cannot match. When used to drive a JBLCMV beamformer, the framework yields near-zero and stable target interaural time difference errors as well as interferer level difference errors across the full SNR range, confirming that the directional accuracy of the dictionary-constrained HRTF estimates translates into better preserved binaural cues at the beamformer output. The internal consistency of the jointly produced estimates is identified as the key structural advantage of the joint framework, producing a beamformer whose constraints and minimum variance criterion are mutually compatible by construction. Note that parts of this work were previously submitted as a workshop paper to the 19th International Workshop on Audio Signal Enhancement (IWAENC), 2026.

1.4 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 establishes the theoretical foundations of the work, covering binaural hearing and spatial audio, the acoustic signal model and CPSDM decomposition, HRTFs and their role as steering vectors, the HRTF dictionary framework and its assumptions, binaural beamforming and its estimation requirements, and the OMP and S-OMP algorithms that form the algorithmic ancestry of the proposed method. Chapter 3 surveys the relevant literature across ATF and RTF estimation, noise covariance estimation, dictionary-based spatial processing, sparse recovery for source localization, and joint estimation frameworks, closing with a structured gap analysis that identifies the specific combination of properties unaddressed by existing methods. Chapter 4 presents the proposed covariance-based S-OMP algorithm in full, deriving each step from first principles, establishing connections to existing methods, and analyzing the computational complexity of the framework. Chapter 5 evaluates the proposed framework through controlled simulation experiments, reporting HRTF estimation accuracy and binaural cue preservation results against sequential baselines across a range of acoustic conditions, followed by a synthesis discussion and a critical analysis of the framework’s limitations. Chapter 6 summarizes the contributions of the thesis, states the overarching conclusions supported

by the experimental evidence, and identifies directions for future work that address the current framework's limitations and extend it toward real-world deployment.

Background

2.1 Binaural Hearing and Spatial Audio

The human auditory system is remarkably adept at parsing complex acoustic environments. In everyday listening situations, multiple simultaneously active sources can be selectively attended to by exploiting spatial differences in the signals arriving at the two ears, a perceptual capability commonly referred to as the cocktail party effect [20, 63]. Hearing loss degrades the ability to resolve these spatial cues, leading to difficulties in noisy, multi-talker environments that amplification alone cannot fully compensate [39]. Binaural speech enhancement systems aim to restore this capability by operating on signals from both ears jointly, rather than treating each ear independently.

Spatial Hearing Cues

Sound source localization relies primarily on two classes of binaural cue: interaural time differences (ITDs) and interaural level differences (ILDs), both arising from the physical separation of the ears and the acoustic filtering effects of the head, pinnae, and torso.

The ITD is the interaural arrival time difference, approximately proportional to $\sin(\phi)$ for a source at azimuth ϕ under a spherical head model. Below approximately 1.5 kHz, it manifests as an interaural phase difference (IPD) and is the dominant azimuthal cue [62]; above this frequency, it becomes phase-ambiguous and the auditory system transitions to envelope ITD and ILD.

The ILD is the interaural level difference arising from head shadow. It becomes a strong directional cue above approximately 3 kHz, where the acoustic wavelength is comparable to the head diameter [65], but is typically weak at low frequencies. The two cues are thus complementary, with ITD dominating at low frequencies and ILD at high frequencies.

Beyond azimuth, the direction-dependent spectral coloration introduced by the pinnae provides cues for elevation and front-back disambiguation [29]. These spectral shape cues are encoded in the head-related transfer function, discussed in Section 2.3.

Binaural Cue Preservation in Hearing Aids

Hearing aid processing introduces a fundamental tension between noise reduction and spatial hearing. Noise reduction algorithms improve the signal-to-noise ratio of the target source but inevitably alter the binaural cues the listener relies on for localization and source segregation [57]. This tension is particularly acute for multi-microphone systems, which achieve greater noise reduction by exploiting spatial diversity but apply different filters to the left and right ear signals. If binaural cue preservation is not

an explicit design objective, the processed outputs might no longer carry consistent binaural cues, causing a perceptual collapse of the spatial scene: the target may be heard more clearly, but the listener loses the ability to localize it or exploit spatial release from masking [9].

Binaural speech enhancement systems must therefore simultaneously suppress noise sufficiently to improve intelligibility and preserve binaural cues sufficiently to maintain spatial auditory perception. This tradeoff is a central design consideration in the binaural hearing aid literature [53, 9] and directly motivates the beamforming framework described in Section 2.5.

The Interaural Transfer Function

A convenient encapsulation of the binaural cues carried by a source is the interaural transfer function (ITF), defined as the ratio of the acoustic transfer functions from source i to the left and right reference microphones at bin k ,

$$\text{ITF}_i(k) = \frac{a_{i,L}(k)}{a_{i,R}(k)}, \quad (2.1)$$

whose squared magnitude encodes the ILD and whose phase encodes the IPD [9] (the ITD essentially being the IPD across frequency),

$$\text{ILD}_i(k) = |\text{ITF}_i(k)|^2, \quad \text{IPD}_i(k) = \angle \text{ITF}_i(k), \quad \text{ITD}_i(k) = \frac{\text{IPD}_i(k)}{2\pi k}. \quad (2.2)$$

A system that preserves the ITF of each source at its output therefore simultaneously preserves both level- and phase-based spatial cues [27]. Moreover, for a simple setup with one microphone on each side, the ITF is equivalent to the relative transfer function (RTF) between the left and right reference microphones, linking binaural cue preservation directly to the RTF estimation problem discussed further in Chapter 3.

2.2 Acoustic Signal Model

The acoustic signal model formalizes the relationship between the sound sources present in a scene and the signals observed at the microphone array. Establishing this model precisely is essential, as it underpins every subsequent processing step. In other words, the beamformer design, the dictionary-based pursuit, and the interpretation of the estimated quantities all follow directly from the assumptions embedded in the model. This section develops the model from the time domain through to the second-order statistical representation that the proposed algorithm operates on.

The Time Domain Representation

Consider a hearing aid microphone array of M channels observing d simultaneously active point sources. The signal $x_m[n]$ at microphone m is a superposition of convolutions

of each source signal with its corresponding acoustic impulse response, plus additive noise,

$$x_m[n] = (h_{s_1 \rightarrow m} * s_1)[n] + \sum_{i=2}^d (h_{s_i \rightarrow m} * s_i)[n] + n_m[n], \quad (2.3)$$

where $s_i[n]$ is the i^{th} source signal, $h_{s_i \rightarrow m}[n]$ is the corresponding acoustic impulse response to microphone m , $*$ is the convolution operator, and $n_m[n]$ captures late reverberation, ambient noise, and microphone self-noise. Source 1 is designated the target of interest, with the remaining $d - 1$ sources treated as interferers. Since direct processing of time-domain convolutions is computationally expensive and does not exploit the frequency-dependent structure of the acoustic scene, it is standard practice to work in the short-time Fourier transform (STFT) domain. Using short time frames also allows for treating speech, a statistically unstationary signal, as roughly wide-sense stationary (WSS).

The STFT Domain Representation

The STFT of a discrete-time signal $x[n]$ is defined as

$$x(k, l) = \sum_{n=0}^{N_{FFT}-1} x[lR + n] w[n] e^{-j2\pi kn/N_{FFT}}, \quad (2.4)$$

where $w[n]$ is an analysis window of length N_{FFT} , R is the frame shift, $k \in \{0, \dots, N_{FFT}-1\}$ indexes frequency bins, and l indexes time frames. Under the narrow-band assumption, i.e., each acoustic impulse response is shorter than N_{FFT} , convolution reduces to multiplication by a complex scalar (the ATF) within each bin. The $M \times 1$ observation vector at bin k , frame l then becomes

$$\mathbf{x}(k, l) = \mathbf{a}_1(k, l)s_1(k, l) + \sum_{i=2}^d \mathbf{a}_i(k, l)s_i(k, l) + \mathbf{n}(k, l) \quad (2.5)$$

$$= \underbrace{\mathbf{A}(k, l)\mathbf{s}(k, l)}_{\mathbf{y}(k, l)} + \mathbf{n}(k, l), \quad (2.6)$$

where $\mathbf{A}(k, l) = [\mathbf{a}_1(k, l) \cdots \mathbf{a}_d(k, l)]$ is the ATF matrix, $\mathbf{s}(k, l) = [s_1(k, l) \cdots s_d(k, l)]^T$ is the source signal vector, $\mathbf{y}(k, l) = \mathbf{A}(k, l)\mathbf{s}(k, l)$ is the desired signal component, and $\mathbf{n}(k, l)$ captures late reverberation, ambient noise, and microphone self-noise.

The Cross-Power Spectral Density Matrix

Since individual STFT frames are highly stochastic, it is standard to work with the second-order statistics of the observation vector, captured by the cross-power spectral density matrix (CPSDM). Assuming mutually independent sources and noise [17], the

theoretical CPSDM of $\mathbf{x}$ is

$$\begin{aligned}\mathbf{R}_{\mathbf{x}}(k, l) &= \mathbb{E} [\mathbf{x}(k, l)\mathbf{x}^H(k, l)] \\ &= \underbrace{\mathbf{A}(k, l)\mathbf{R}_{\mathbf{s}}(k, l)\mathbf{A}^H(k, l)}_{\mathbf{R}_{\mathbf{y}}(k, l)} + \mathbf{R}_{\mathbf{n}}(k, l),\end{aligned}\quad (2.7)$$

where $\mathbb{E}$ is the expectation operator, $\mathbf{R}_{\mathbf{s}}(k, l) \in \mathbb{C}^{d \times d}$ is the source covariance matrix and $\mathbf{R}_{\mathbf{n}}(k, l) \in \mathbb{C}^{M \times M}$ is the noise covariance matrix. Under source independence, $\mathbf{R}_{\mathbf{s}}(k, l)$ is diagonal with source power entries $\sigma_{s_i}^2(k, l)$, reducing the signal covariance contribution to

$$\mathbf{R}_{\mathbf{y}}(k) = \sum_{i=1}^d \sigma_{s_i}^2(k) \mathbf{a}_i(k) \mathbf{a}_i^H(k), \quad (2.8)$$

a weighted sum of rank-one outer products of the ATF vectors with source powers as weights. This makes explicit that the spatial structure of $\mathbf{R}_{\mathbf{x}}$ is entirely determined by the ATF vectors and source powers, and recovering them from the observed CPSDM is the central estimation problem addressed in this thesis.

Sample Covariance Estimation

In practice, the theoretical CPSDM $\mathbf{R}_{\mathbf{x}}(k, l)$ is estimated by averaging outer products of the observation vectors over a block of L consecutive STFT frames [61],

$$\hat{\mathbf{R}}_{\mathbf{x}}(k) = \frac{1}{L} \sum_{\ell \in \mathcal{L}} \mathbf{x}(k, \ell) \mathbf{x}^H(k, \ell), \quad (2.9)$$

where $\mathcal{L}$ denotes the set of L frame indices in the current block. It converges to the true CPSDM as $L \rightarrow \infty$ provided the acoustic scene is approximately stationary over the time block, i.e., $\mathbf{A}(k)$ and the source and noise statistics remain approximately constant, a reasonable assumption for slowly moving sources and quasi-stationary noise over block durations of a few hundred milliseconds. The estimator exhibits a well-characterized bias-variance tradeoff: small L yields low bias and rapid tracking but high variance, while large L reduces variance at the cost of potential stationarity violations. The choice of L therefore represents a fundamental tradeoff between statistical reliability and temporal resolution, partly influenced by the intended use case.

The Estimation Problem

The signal model (2.7) decomposes the observed CPSDM into signal and noise covariance contributions, both depending on quantities unknown in practice: the ATF matrix $\mathbf{A}(k)$, the source powers $\sigma_{s_i}^2(k)$, and the noise covariance $\mathbf{R}_{\mathbf{n}}(k)$. Accurate knowledge of these quantities is required for optimal spatial filtering: minimum variance beamformers such as the MVDR, MWF, and LCMV require the ATF vectors as steering constraints and the noise covariance as the matrix to be inverted, and mismatches in either can degrade performance severely, potentially amplifying rather than suppressing noise [10, 5, 47].

The problem is fundamentally ill-posed without additional constraints, as infinitely many combinations of $\mathbf{A}(k)$, $\mathbf{R}_s(k)$, and $\mathbf{R}_n(k)$ are consistent with a given $\hat{\mathbf{R}}_x(k)$. The structural constraints adopted in this thesis are introduced in Section 2.4, following the formal definition of the HRTF in Section 2.3.

2.3 Head-Related Transfer Functions

The acoustic path from a sound source to the ear canal of a listener is shaped by a complex chain of physical interactions involving the torso, shoulders, head, and pinnae. These structures scatter, reflect, and diffract the incident sound field in ways that are highly dependent on the direction of the source, introducing direction-specific spectral and temporal modifications to the signal before it reaches the eardrum. The head-related transfer function (HRTF) is the frequency-domain characterization of this direction-dependent filtering, and constitutes the fundamental acoustic representation of spatial hearing used throughout this thesis.

Definition and Physical Interpretation

The HRTF for a source at azimuth ϕ , elevation θ and distance r is defined as the ratio of the sound pressure at the ear canal entrance with the listener present to the free-field pressure at the head center without the listener,

$$h(\phi, \theta, r, f) = \frac{P_{ear}(\phi, \theta, r, f)}{P_{free}(f)}, \quad (2.10)$$

capturing the direction-dependent transformation imposed by the listener’s anatomy on an incoming sound. Its time-domain counterpart is the head-related impulse response (HRIR), related by $H(\phi, \theta, r, f) = \mathcal{F}^{-1}\{h(\phi, \theta, r, t)\}$, where $\mathcal{F}^{-1}$ is the inverse Fourier transform.

The HRTF encodes all three classes of spatial hearing cues discussed in Section 2.1: the ILD from the interaural level difference between left and right ear HRTFs, the IPD and ITD from their interaural phase difference, and elevation and front-back cues from the direction-dependent spectral envelope shaped by the pinna geometry [29]. It therefore constitutes a complete acoustic description of the binaural cues associated with a source at a given direction, making its accurate measurement a prerequisite for binaural speech enhancement.

The Far-Field Assumption

Beyond a critical distance (typically on the order of one meter for human-scale heads), the HRTF becomes approximately independent of distance as the incident wavefront approximates a plane wave, varying only with direction (ϕ, θ) . Under this far-field assumption, the distance dependence in (2.10) is dropped and the HRTF written as $h(\phi, \theta, f)$. This assumption is standard in both HRTF measurement and binaural processing, and is adopted throughout this thesis; the near-field case, where source distance materially alters the HRTF shape and introduces distance-dependent ILD cues, is not considered.

HRTF Measurement

Since the HRTF encodes the full acoustic effect of an individual listener’s anatomy, it is in principle unique to each person. However, a dense spherical HRTF measurement requires specialized equipment and is impractical for routine applications, making publicly available databases measured from dummy heads or subject pools the standard surrogate. The OlHeadHRTF dataset [11], used in this thesis, provides measurements at the microphone positions of seven hearing device placements on a Brüel & Kjær 4128C HATS dummy head, covering 91 directions at a sampling rate of 48 kHz.

Individualizing HRTFs to a specific listener (via anthropometric scaling [66] or database selection [40]) remains an open research problem. The use of a generic dummy head HRTF in this thesis is a deliberate simplification that isolates the estimation framework’s performance from individualization error, consistent with the controlled evaluation design of Chapter 5.

HRTFs in the STFT Domain

For microphone array processing, the HRTF at direction (ϕ, θ) is evaluated at each discrete frequency bin k , yielding an M -dimensional complex vector

$$\mathbf{h}(\phi, \theta, k) = [h_1(\phi, \theta, k), h_2(\phi, \theta, k), \dots, h_M(\phi, \theta, k)]^T \in \mathbb{C}^M, \quad (2.11)$$

where $h_m(\phi, \theta, k)$ is the HRTF from the source at direction (ϕ, θ) to microphone m at bin k . Under the narrowband approximation of Section 2.2, this vector is exactly the ATF vector $\mathbf{a}_i(k)$ in (2.6) when source i is located at (ϕ, θ) , and therefore serves simultaneously as the beamformer steering vector and the dictionary atom for that direction in the sparse pursuit framework.

Relationship Between HRTFs and ATFs

An ATF describes the acoustic transfer function between a source and microphone in any environment, including room reflections, while an HRTF is a specific ATF measured under anechoic conditions capturing only the direct-path anatomical contribution. The ATF of a real room can be approximated by [2]

$$a_{i,m}(k) = h_m(\phi_i, \theta_i, k) \cdot h_{room,i,m}(k), \quad (2.12)$$

where $h_{room,i,m}(k)$ is the room transfer function; in an anechoic environment, $h_{room,i,m}(k) = 1$ and the ATF reduces exactly to the HRTF.

2.4 The HRTF Dictionary Framework

The estimation problem identified at the end of Section 2.2 is ill-posed without additional constraints. The approach adopted in this thesis addresses this ill-posedness by restricting the set of admissible ATF vectors to a finite collection of pre-measured HRTFs, forming an overcomplete dictionary that discretizes the continuous space of possible source directions. This section defines the dictionary, states the assumptions

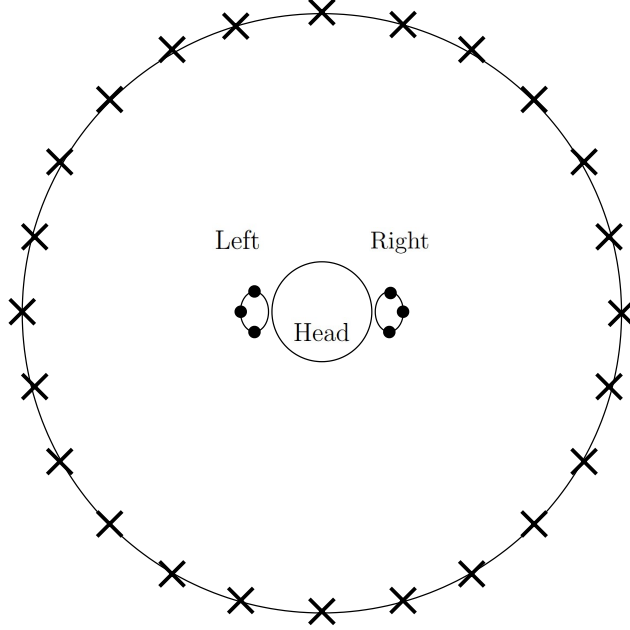


Figure 2.1: Example HRTF dictionary with $N = 24$ virtual source locations at zero elevation, uniformly distributed in azimuth.

under which it is used, and establishes the sparse recovery interpretation that motivates the algorithm of Chapter 4.

Dictionary Construction

Let $\mathcal{D} = \{(\phi_1, 0), (\phi_2, 0), \dots, (\phi_N, 0)\}$ denote N candidate source directions at zero elevation, uniformly distributed in azimuth. The HRTF vectors across all directions form the dictionary at bin k ,

$$\mathcal{A}(k) = [\mathbf{h}_1(k) \ \mathbf{h}_2(k) \ \cdots \ \mathbf{h}_N(k)] \in \mathbb{C}^{M \times N}, \quad (2.13)$$

with each column $\mathbf{h}_n(k)$ referred to as a dictionary atom. The frequency-stacked dictionary tensor is $\underline{\mathcal{A}} \in \mathbb{C}^{M \times N \times K}$, with the k^{th} slice $\mathcal{A}(k)$. Fig. 2.1 illustrates an example dictionary with $N = 24$ uniformly spaced azimuths.

The Sparse Source Model

The dictionary framework transforms ATF estimation into a support recovery problem. If the d active sources coincide with dictionary entries, the ATF matrix is a column submatrix of $\mathbf{A}(k)$,

$$\mathbf{A}(k) = \mathcal{A}_{\mathbf{\Lambda}}(k), \quad (2.14)$$

where $\mathbf{\Lambda} = \{\lambda_1, \dots, \lambda_d\} \subset \{1, \dots, N\}$ is the active support set. Substituting into (2.7),

$$\mathbf{R}_{\mathbf{x}}(k) = \mathcal{A}_{\mathbf{\Lambda}}(k) \mathbf{R}_{\mathbf{s}}(k) \mathcal{A}_{\mathbf{\Lambda}}^H(k) + \mathbf{R}_{\mathbf{n}}(k), \quad (2.15)$$

expressing the observed covariance as a sparse decomposition over $\mathcal{A}(k)$: only $d \ll N$ atoms contribute, with the remainder having zero contribution. The support recovery problem is then: given $\hat{\mathbf{R}}_{\mathbf{x}}(k)$ and $\mathcal{A}(k)$, identify $\mathbf{\Lambda}$ and jointly estimate the source powers $\mathbf{R}_{\mathbf{s}}(k)$ and noise covariance $\mathbf{R}_{\mathbf{n}}(k)$. In general, the HRTF vectors selected through $\mathbf{\Lambda}$ provide an approximation to the true ATF vectors of the acoustic scene. However, since the experimental evaluation in this thesis (Chapter 5) is conducted under anechoic conditions, the ATF vectors estimated using the proposed method will hereafter also be referred to as HRTF vectors.

Dictionary Assumptions

The dictionary framework rests on three assumptions as they define the operating conditions and validity boundary of the proposed algorithm.

A1: Source sparsity. $d \ll N$ and $d \leq M$. The first condition ensures a sparse CPSDM representation over the dictionary; the second ensures that $\mathbf{R}_{\mathbf{y}}(k)$ has rank at most $M-1$, leaving sufficient degrees of freedom for the residual covariance to constitute a meaningful noise estimate and for the beamformer to remain well-conditioned.

A2: Source stationarity. The sources are static or slowly moving, implying $\mathbf{A}(k)$ is approximately constant over the estimation block, i.e., $\mathbf{A}(k, l) \approx \mathbf{A}(k)$ for all $l \in \mathcal{L}$, justifying the block-based sample CPSDM estimator (2.9). This assumption may be violated for rapidly moving sources or highly dynamic acoustic scenes.

A3: On-grid source directions. When source i lies at a dictionary direction ϕ_{λ_i} , the true ATF satisfies $\mathbf{a}_i(k) = \mathbf{h}_{\lambda_i}(k) \in \mathcal{A}(k)$. Off-grid sources are approximated by the nearest dictionary atom.

Dictionary Normalization

Prior to being used in the algorithm, each dictionary atom $\mathbf{h}_n(k)$ is normalized to unit norm along the channel dimension. As a result, we have

$$\tilde{\mathbf{h}}_n(k) = \frac{\mathbf{h}_n(k)}{\|\mathbf{h}_n(k)\|_2}, \quad (2.16)$$

so that $\|\tilde{\mathbf{h}}_n(k)\|_2 = 1$ for all n and k . Analogously, the dictionary of these normalized atoms is represented by $\tilde{\mathcal{A}}(k)$. The motivation for this normalization and its consequences for the atom selection step are discussed in detail in Section 4.1.

Dictionary Coherence

The mutual coherence of the dictionary, defined as the maximum absolute normalized inner product between any two distinct atoms,

$$\mu(\tilde{\mathcal{A}}(k)) = \max_{i \neq j} \left| \tilde{\mathbf{h}}_i^H(k) \tilde{\mathbf{h}}_j(k) \right|, \quad (2.17)$$

takes values in $[0, 1]$ and has significant implications for sparse recovery performance. For HRTF dictionaries, coherence is governed by the angular spacing between entries

and the spatial selectivity of the array: atoms corresponding to nearby directions have similar spatial signatures and therefore high coherence. Under the condition $\mu < 1/(2d - 1)$, Orthogonal Matching Pursuit (OMP) is guaranteed to recover the correct d -sparse support in the noiseless case [55]; for fine-resolution HRTF dictionaries, this condition may not hold for closely spaced sources, setting a practical resolution limit on the algorithm’s ability to distinguish them.

2.5 Binaural Beamforming

A beamformer is a spatial filter that combines the signals from multiple microphones to enhance a desired source while suppressing interference and noise. In the binaural hearing aid context, the beamformer must satisfy an additional constraint beyond noise reduction: it must produce a pair of output signals (one for the left and right ear each) that preserve the binaural cues of the sources in the acoustic scene, so that the listener retains the ability to localize and segregate sounds spatially. This section introduces the beamforming framework used in this thesis, progressing from the single-channel filter to the binaural multichannel formulation, and identifies the quantities that must be estimated to implement it in practice.

The MVDR Beamformer

The MVDR beamformer minimizes the output noise power subject to a distortionless constraint on the target source [47]. Formally, for a target with ATF vector $\mathbf{a}_1(k)$, the MVDR filter is the solution to

$$\mathbf{w}_{\text{MVDR}}(k) = \arg \min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_n(k) \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{a}_1(k) = 1, \quad (2.18)$$

with closed-form solution

$$\mathbf{w}_{\text{MVDR}}(k) = \frac{\mathbf{R}_n^{-1}(k) \mathbf{a}_1(k)}{\mathbf{a}_1^H(k) \mathbf{R}_n^{-1}(k) \mathbf{a}_1(k)}. \quad (2.19)$$

It passes the target signal without distortion while minimizing the contribution of noise and interference at the output.

The sensitivity of the MVDR to steering vector mismatch is well documented: when the assumed steering vector $\mathbf{a}_1(k)$ differs from the true ATF, the distortionless constraint is applied at the wrong direction and the filter may perceptually suppress the target rather than enhance it [10]. This sensitivity makes accurate ATF estimation a prerequisite for reliable MVDR performance, directly motivating the HRTF dictionary-based estimation approach of this thesis.

The Multichannel Wiener Filter

The multichannel Wiener filter (MWF) is a widely used spatial filter that minimizes the mean squared error between the filter output and a reference signal, typically taken

as the target source at a reference microphone [17]. For a target at source index $i = 1$, the MWF output at the left reference microphone is

$$\hat{s}_1(k, l) = \mathbf{w}_L^H(k) \mathbf{x}(k, l), \quad (2.20)$$

where $\mathbf{w}_L(k) \in \mathbb{C}^M$ is the filter weight vector. Minimizing the mean squared error $\mathbb{E}[|\hat{s}_1(k, l) - s_1(k, l)|^2]$ with respect to $\mathbf{w}_L(k)$ yields the Wiener solution:

$$\mathbf{w}_{\text{MWF}}(k) = \mathbf{R}_x^{-1}(k) \mathbf{R}_y(k) \mathbf{e}_1, \quad (2.21)$$

where $\mathbf{e}_1 \in \{0, 1\}^d$ is the first standard basis vector selecting the target source, and $\mathbf{R}_y(k) = \mathbf{A}(k) \mathbf{R}_s(k) \mathbf{A}^H(k)$ is the signal covariance. Using $\mathbf{R}_x^{-1} = (\mathbf{R}_y + \mathbf{R}_n)^{-1}$ and the Woodbury lemma [42], the MWF solution can be rewritten as

$$\mathbf{w}_{\text{MWF}}(k) = (\mathbf{R}_y(k) + \mathbf{R}_n(k))^{-1} \mathbf{R}_y(k) \mathbf{e}_1, \quad (2.22)$$

making explicit that the filter requires knowledge of both the signal covariance $\mathbf{R}_y(k)$ and the noise covariance $\mathbf{R}_n(k)$. Equivalently, the MWF can be expressed in terms of the ATF vector of the target and the noise covariance as

$$\mathbf{w}_{\text{MWF}}(k) = \frac{\mathbf{R}_n^{-1}(k) \mathbf{a}_1(k) \sigma_{s_1}^2(k)}{\mathbf{a}_1^H(k) \mathbf{R}_n^{-1}(k) \mathbf{a}_1(k) + 1/\text{SNR}(k)}, \quad (2.23)$$

connecting the MWF to the MVDR beamformer in the high-SNR limit.

The Linearly Constrained Minimum Variance Beamformer

The LCMV beamformer generalizes the MVDR by imposing multiple linear constraints simultaneously, allowing the filter to satisfy distortionless or cue-preservation constraints for multiple sources [34]. For a set of d sources with ATF vectors $\mathbf{a}_1(k), \dots, \mathbf{a}_d(k)$ and a matrix of constraint vectors $\mathbf{C}(k) \in \mathbb{C}^{M \times d}$ and response vector $\mathbf{f}(k) \in \mathbb{C}^d$, the LCMV problem is

$$\mathbf{w}_{\text{LCMV}}(k) = \arg \min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_n(k) \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H(k) \mathbf{C}(k) = \mathbf{f}^H(k), \quad (2.24)$$

with solution

$$\mathbf{w}_{\text{LCMV}}(k) = \mathbf{R}_n^{-1}(k) \mathbf{C}(k) (\mathbf{C}^H(k) \mathbf{R}_n^{-1}(k) \mathbf{C}(k))^{-1} \mathbf{f}(k). \quad (2.25)$$

Each constraint in $\mathbf{C}(k)$ consumes one degree of freedom from the available M microphone channels. The remaining $M - d$ degrees of freedom are used for noise suppression. This tradeoff between constraint satisfaction and noise reduction is fundamental to the LCMV framework and motivates the strategy of selecting only the d most relevant dictionary atoms rather than constraining all N entries simultaneously.

The Joint Binaural LCMV Beamformer

For binaural beamformers, it is more convenient to express the left and weights as one vector $\mathbf{w}(k) = [\mathbf{w}_L^T(k) \mathbf{w}_R^T(k)]^T \in \mathbb{C}^{2M \times 1}$, and also introduce a binaural noise CPSDM

$$\tilde{\mathbf{R}}_{\mathbf{n}}(k) = \begin{bmatrix} \mathbf{R}_{\mathbf{n}}(k) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\mathbf{n}}(k) \end{bmatrix} \in \mathbb{C}^{2M \times 2M}. \quad (2.26)$$

The constraints matrix $\mathbf{C}(k)$ also accounts for this concatenation, i.e., $\mathbf{C}(k) \in \mathbb{C}^{2M \times d}$.

The General Binaural LCMV (GBLCMV) beamformer [27] provides, as the name suggests, a generalized extension of the LCMV framework to produce a binaural output pair while simultaneously preserving the binaural cues of both the target and the interfering sources. Using the above notation, the GBLCMV problem becomes

$$\mathbf{w}_{\text{GBLCMV}}(k) = \arg \min_{\mathbf{w}} \mathbf{w}^H \tilde{\mathbf{R}}_{\mathbf{n}}(k) \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H(k) \mathbf{C}(k) = \mathbf{f}^H(k), \quad (2.27)$$

with solution

$$\mathbf{w}_{\text{GBLCMV}}(k) = \tilde{\mathbf{R}}_{\mathbf{n}}^{-1}(k) \mathbf{C}(k) \left(\mathbf{C}^H(k) \tilde{\mathbf{R}}_{\mathbf{n}}^{-1}(k) \mathbf{C}(k) \right)^{-1} \mathbf{f}(k). \quad (2.28)$$

The Joint Binaural LCMV (JBLCMV) beamformer [26] is a special case of the GBLCMV where $\mathbf{C}(k)$ and $\mathbf{f}(k)$ are modified so that more interfering sources can be accommodated within the same number of degrees of freedom, when compared to the simpler Binaural LCMV (BLCMV). The problem is the same as (2.27), where

$$\begin{aligned} \mathbf{C}(k) &= \begin{bmatrix} \mathbf{C}_t(k) & | & \mathbf{C}_i(k) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{a}_1(k) & \mathbf{0} & | & \mathbf{a}_2(k)a_{2,R}(k) & \cdots & \mathbf{a}_d(k)a_{d,R}(k) \\ \mathbf{0} & \mathbf{a}_1(k) & | & -\mathbf{a}_2(k)a_{2,L}(k) & \cdots & -\mathbf{a}_d(k)a_{d,L}(k) \end{bmatrix} \in \mathbb{C}^{2M \times (d+1)}, \end{aligned} \quad (2.29)$$

where $\mathbf{C}_t(k)$ and $\mathbf{C}_i(k)$ are constraint submatrices for the target and interferers respectively. Moreover, by the same analogy,

$$\begin{aligned} \mathbf{f}^T(k) &= \begin{bmatrix} \mathbf{f}_t^T(k) & | & \mathbf{f}_i^T(k) \end{bmatrix} \\ &= \begin{bmatrix} a_{1,L}^* & a_{1,R}^* & | & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{1 \times (d+1)}. \end{aligned} \quad (2.30)$$

Requirements for Beamformer Implementation

All minimum variance-based beamformers discussed above share a common set of requirements that define the estimation targets and motivate the design of the proposed framework.

ATF accuracy. Beamformer constraints are formulated in terms of the ATF vectors of the active sources. Constraint mismatch from ATF estimation errors can cause the filter to suppress rather than preserve the intended sources [10], making ATF accuracy the most critical requirement.

Noise covariance accuracy. The minimum variance criterion requires inverting the noise covariance matrix; an inaccurate estimate degrades noise suppression, while

a rank-deficient or ill-conditioned estimate can destabilize the inversion entirely. Interfering source contributions are incorporated by adding their rank-one covariance terms to the noise CPSDM.

Internal consistency. When ATF vectors and noise covariance are estimated by different methods at different times, the resulting estimates may be mutually inconsistent, causing the beamformer constraints and minimum variance criterion to pull in different directions. Joint estimation of all required quantities from the same observation block ensures internal consistency [28], and is a key motivation for the framework proposed in Chapter 4.

2.6 Sparse Signal Recovery

Sparse signal recovery addresses the problem of representing an observed signal as a linear combination of a small number of elements drawn from a large, overcomplete collection. This framework has found widespread application in signal processing, statistics, and machine learning, underpinning methods for compressed sensing, dictionary learning, and source separation [3, 6]. In the context of this thesis, sparse recovery provides the algorithmic foundation for the dictionary-based ATF estimation framework introduced in Section 2.4: the problem of identifying which d atoms from the HRTF dictionary best explain the observed CPSDM is a support recovery problem, and the proposed algorithm is derived as an extension of the greedy pursuit methods reviewed here.

Overcomplete Dictionaries and Sparse Representations

Let $\mathbf{y} \in \mathbb{C}^M$ be an observed signal vector and $\mathbf{D} \in \mathbb{C}^{M \times N}$ an overcomplete dictionary with $N > M$ columns, referred to as atoms. Unless stated otherwise, the atoms are assumed to be ℓ_2 -normalized, i.e., $\|\mathbf{d}_n\|_2 = 1$ for all n . A sparse representation of $\mathbf{y}$ over $\mathbf{D}$ is a coefficient vector $\mathbf{x} \in \mathbb{C}^N$ such that $\mathbf{y} \approx \mathbf{D}\mathbf{x}$ and $\|\mathbf{x}\|_0 \ll N$, where $\|\cdot\|_0$ denotes the ℓ_0 norm, or number of non-zero entries. The support of $\mathbf{x}$ is the index set $\Lambda = \{n : x_n \neq 0\}$, and the sparsity level is defined as the cardinality of its support $|\Lambda| = d = \|\mathbf{x}\|_0$. The sparse recovery problem seeks to find the sparsest coefficient vector consistent with the observation:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \epsilon, \quad (2.31)$$

where $\epsilon \geq 0$ is a noise tolerance. This problem is NP-hard in general due to the combinatorial nature of the ℓ_0 norm, motivating the use of tractable approximations. Two principal classes of approximation exist: convex relaxation methods, which replace the ℓ_0 norm with the convex ℓ_1 norm [54], and greedy pursuit methods, which iteratively build up the support set one atom at a time. This thesis employs greedy pursuit, reviewed in the subsections that follow.

Orthogonal Matching Pursuit

Orthogonal Matching Pursuit (OMP) [55] is one of the most widely used greedy algorithms for sparse recovery. Given an observation $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{n}$ where $\mathbf{x}$ is d -sparse and $\mathbf{n}$ is noise, OMP recovers the support Λ and the non-zero coefficients of $\mathbf{x}$ through a sequence of greedy selections and orthogonal projections.

The algorithm is initialized with a residual $\mathbf{r}^{(0)} = \mathbf{y}$ and an empty support $\Lambda^{(0)} = \emptyset$. At each iteration p , three steps are performed.

Matching. The atom most correlated with the current residual is identified:

$$\lambda^{(p)} = \arg \max_{1 \leq n \leq N} |\mathbf{d}_n^H \mathbf{r}^{(p-1)}|, \quad (2.32)$$

The selected atom is the one pointing most in the direction of the current residual, measuring alignment via the inner product.

Support update. The selected index is added to the support:

$$\Lambda^{(p)} = \Lambda^{(p-1)} \cup \{\lambda^{(p)}\}. \quad (2.33)$$

Orthogonal projection. The coefficients of all selected atoms are re-estimated jointly by solving a least-squares problem over the current sub-dictionary $\mathbf{D}_{\Lambda}^{(p)}$:

$$\hat{\mathbf{x}}_{\Lambda}^{(p)} = \arg \min_{\mathbf{x}_{\Lambda}} \left\| \mathbf{y} - \mathbf{D}_{\Lambda}^{(p)} \mathbf{x}_{\Lambda} \right\|_2^2 = \mathbf{D}_{\Lambda}^{(p)\dagger} \mathbf{y}, \quad (2.34)$$

and the residual is updated as the component of $\mathbf{y}$ orthogonal to the selected subspace:

$$\mathbf{r}^{(p)} = \mathbf{y} - \mathbf{D}_{\Lambda}^{(p)} \hat{\mathbf{x}}_{\Lambda}^{(p)} = \left(\mathbf{I} - \mathbf{P}_{\Lambda}^{(p)} \right) \mathbf{y}, \quad (2.35)$$

where $\mathbf{P}_{\Lambda}^{(p)} = \mathbf{D}_{\Lambda}^{(p)} \mathbf{D}_{\Lambda}^{(p)\dagger}$ is the orthogonal projector onto the column space of $\mathbf{D}_{\Lambda}^{(p)}$. By re-estimating all coefficients at every iteration rather than fixing previously selected ones, OMP ensures that the residual is always orthogonal to all selected atoms, which is the defining property that distinguishes it from simpler greedy methods such as Matching Pursuit.

The algorithm terminates after d iterations when the known sparsity level is reached, or when the residual energy falls below a prescribed threshold. Under suitable conditions on the dictionary coherence (specifically when $\mu(\mathbf{D}) < 1/(2d-1)$), OMP recovers the correct support exactly in the noiseless case [55]. In the noisy case, recovery is approximate and the probability of support error depends on the SNR and the coherence structure of the dictionary [13].

Simultaneous Orthogonal Matching Pursuit

Simultaneous OMP (S-OMP) [56] extends OMP to the multiple measurement vector (MMV) setting, where L observations share a common sparse support. The MMV model is

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{N}, \quad (2.36)$$

where $\mathbf{Y} \in \mathbb{C}^{M \times L}$ is the matrix of observations, $\mathbf{X} \in \mathbb{C}^{N \times L}$ is the matrix of sparse coefficients sharing a common row support Λ , and $\mathbf{N} \in \mathbb{C}^{M \times L}$ is noise. The joint

sparsity assumption states that the support of $\mathbf{X}$ is identical, i.e., all columns share the same support set Λ ; the non-zero coefficient values are not required to be equal across all columns.

S-OMP exploits this joint structure by modifying the matching step to aggregate correlation scores across all L measurement vectors simultaneously. At iteration p , the residual matrix is $\mathbf{R}^{(p-1)} \in \mathbb{C}^{M \times L}$ and the atom selection criterion is

$$\lambda^{(p)} = \arg \max_{1 \leq n \leq N} \|\mathbf{d}_n^H \mathbf{R}^{(p-1)}\|_2, \quad (2.37)$$

selecting the atom whose projections onto all L residual vectors have the largest combined energy. The support update and orthogonal projection steps proceed identically to OMP, with the coefficient matrix updated as

$$\hat{\mathbf{X}}_{\Lambda}^{(p)} = \mathbf{D}_{\Lambda}^{(p)\dagger} \mathbf{Y}, \quad (2.38)$$

and the residual matrix updated as

$$\mathbf{R}^{(p)} = \mathbf{Y} - \mathbf{D}_{\Lambda}^{(p)} \hat{\mathbf{X}}_{\Lambda}^{(p)} = \left(\mathbf{I} - \mathbf{P}_{\Lambda}^{(p)} \right) \mathbf{Y}. \quad (2.39)$$

The aggregation of scores across L measurement vectors in (2.37) promotes support selections that are consistent across all observations, improving robustness over single-vector OMP when $L > 1$. Thus, an atom that is genuinely in the support will produce large projections across all L residual vectors, while a spurious atom driven by noise in a single vector will produce inconsistent projections that are suppressed by the aggregation. The exact recovery conditions for S-OMP mirror those of OMP, with the probability of correct support recovery improving as L increases [12].

Chapter Summary

This chapter established the theoretical foundations for the proposed framework. Starting from the binaural hearing context and its dual objective of noise reduction and spatial cue preservation, the acoustic signal model was developed in the STFT domain, yielding the CPSDM decomposition that separates observed spatial covariance into signal and noise contributions. The HRTF was introduced as the direction-dependent steering vector encoding binaural localization cues, and the HRTF dictionary framework was shown to recast ill-posed ATF estimation as a tractable sparse support recovery problem. The beamforming section established that joint, internally consistent estimation of ATF vectors, source covariance, and noise covariance is a prerequisite for reliable performance. Finally, OMP and S-OMP were derived as the algorithmic ancestors of the proposed method, bridging coefficient-domain greedy pursuit and the covariance-domain extension developed in Chapter 4.

Related Work

The problem addressed in this thesis sits at the intersection of three active research areas: acoustic transfer function and relative transfer function estimation, noise covariance estimation, and sparse signal recovery applied to spatial audio. Each of these areas has produced a rich body of methods that address subsets of the joint estimation problem, and understanding their individual contributions and limitations is essential for appreciating both the motivation for and the novelty of the proposed framework.

Throughout this chapter, methods are evaluated not only on their individual merits but on the assumptions they require, the quantities they produce, and the degree to which their outputs are internally consistent and suitable as direct inputs to a binaural beamformer.

3.1 ATF and RTF Estimation

Accurate knowledge of the acoustic or relative transfer functions (ATFs/RTFs) of the active sources is a prerequisite for minimum variance beamforming, as established in Section 2.5. In practice, these functions are unknown and must be estimated from the observed microphone signals, typically without access to clean source signals or noise-only intervals. This section reviews the principal approaches to this estimation problem, organized by their underlying methodology.

Covariance-Based RTF Estimation

The most widely studied RTF estimators exploit the relationship between the observed CPSDM and the signal model (2.7) to extract ATF or RTF vectors without requiring clean source observations. Two foundational methods are covariance subtraction (CS) [16, 32] and covariance whitening (CW) [8, 59]. CS estimates the RTF as the dominant eigenvector of the difference between speech-active and noise-only CPSDMs, i.e.,

$$\mathbf{R}_{\mathbf{x}}(k) - \mathbf{R}_{\mathbf{n}}(k) = \sigma_s^2(k) \mathbf{a}(k) \mathbf{a}^H(k), \quad (3.1)$$

while CW instead takes the dominant generalized eigenvector of the pair $(\mathbf{R}_{\mathbf{x}}(k), \mathbf{R}_{\mathbf{n}}(k))$ by

$$\mathbf{R}_{\mathbf{x}}(k) \mathbf{a}(k) = \lambda(k) \mathbf{R}_{\mathbf{n}}(k) \mathbf{a}(k). \quad (3.2)$$

In both cases the dominant eigenvector yields an estimate of the ATF vector $\mathbf{a}(k)$, up to a scalar ambiguity. A comparative study [35] demonstrates that CW consistently outperforms CS, as noise covariance whitening normalizes channel-dependent signal-to-noise ratio (SNR) prior to eigendecomposition.

Both methods, however, share a fundamental dependence on a separately estimated noise covariance $\mathbf{R}_n(k)$, typically obtained from voice-activity-detected noise-only intervals. This requirement is a practical limitation in continuous speech or non-stationary noise environments, where reliable noise-only intervals may be scarce and estimation errors propagate directly into RTF accuracy.

Subspace and GEVD-Based Methods

Taseska and Habets [52] estimate RTFs without disjoint activity periods by exploiting single-source-dominant time-frequency bins under the W-disjoint orthogonality assumption, extracting per-bin estimates from the dominant left singular vector and aggregating them into a broadband MVDR steering vector. While effective at moderate SNR, accuracy degrades in densely overlapping multi-source scenarios. This method serves as a baseline in Chapter 5.

Markovich et al. [34] proposed an eigenspace LCMV beamformer that applies generalized eigenvalue decomposition (GEVD) to the pair $(\mathbf{R}_x(k), \mathbf{R}_n(k))$, but retains multiple generalized eigenvectors rather than only the dominant one. The leading eigenvectors estimate the RTFs of the desired sources, while the remaining eigenvectors span an interference subspace basis, enabling null-steering toward multiple interferers without requiring their individual RTFs. However, the need for disjoint activity periods for noise covariance and subspace estimation limits applicability in continuously active scenes.

Varzandeh et al. [58] achieve comparable performance to GEVD-based estimators at lower complexity through iterative covariance subtraction, exploiting the low-rank signal covariance structure to refine RTF estimates without full eigendecomposition. This work also introduces the Hermitian angle metric for RTF estimation accuracy, adopted as the primary HRTF estimation metric in Chapter 5.

Binaural RTF Estimation

Fejgin and Doclo [15] compare covariance whitening, covariance subtraction, and spatial coherence methods for RTF-based direction-of-arrival (DOA) estimation in a binaural hearing aid setup augmented with an external microphone. DOA is estimated by minimizing the frequency-averaged Hermitian angle between the estimated RTF vector and a database of prototype head-mounted RTFs. This is a dictionary matching step structurally similar to that of the proposed algorithm, but performed as a decoupled post-hoc lookup rather than an integrated procedure. The spatial coherence method achieves the best performance in reverberant diffuse-noise conditions by exploiting the external microphone as a noise reference, while CW performs comparably without requiring one. This work motivates the use of HRTF dictionaries as a constraint on the estimation problem, and illustrates the key limitation of post-hoc matching: decoupling RTF estimation from dictionary projection cannot guarantee consistency with any dictionary atom, introducing approximation errors that propagate into the downstream beamformer.

3.2 Noise Covariance Estimation

Accurate estimation of the noise covariance matrix $\mathbf{R}_n(k)$ is a prerequisite for minimum variance beamforming, as established in Section 2.5. In practice, the noise covariance is unknown and potentially nonstationary, requiring estimation from the observed microphone signals even when the target source is active. This section reviews the principal approaches to noise covariance estimation, with emphasis on the methods most relevant to the binaural hearing aid setting and to the baseline used in Chapter 5.

VAD-Based Noise Estimation

The most straightforward approach to noise covariance estimation accumulates the sample covariance of observed signals over noise-only frames identified by voice activity detection (VAD), via

$$\hat{\mathbf{R}}_n(k) = \frac{1}{|\mathcal{L}_n|} \sum_{l \in \mathcal{L}_n} \mathbf{x}(k, l) \mathbf{x}^H(k, l), \quad (3.3)$$

where $\mathcal{L}_n$ denotes the noise-only frame set. This estimator is asymptotically unbiased and consistent under noise stationarity. Serizel et al. [46] employ this scheme within a GEVD-based low-rank MWF for cochlear implant noise reduction, using rank- R signal covariance approximations to improve robustness at low SNR. This work provides the VAD-based noise estimation component of the baseline in Chapter 5.

The fundamental limitations of VAD-based methods are their dependence on the availability and accuracy of noise-only intervals: continuously active speech scenes yield sparse updates, non-stationary noise fields may not be trackable even with frequent updates, and VAD errors introduce bias that propagates directly into beamformer performance. These limitations motivate approaches that estimate the noise covariance without requiring explicit noise-only intervals.

Statistical Noise Estimation

Cohen [7] proposes the improved minima controlled recursive averaging (IMCRA) algorithm for single-channel noise spectrum estimation, replacing hard VAD decisions with minimum-statistics tracking over a sliding periodogram window and a soft speech-presence probability that adapts smoothing to non-stationary conditions. While limited to single-channel power spectra, IMCRA establishes the principle of VAD-free noise estimation that motivates later multichannel extensions. Kamkar-Parsi and Bouchard [23] extend this to the binaural multichannel setting, exploiting the known spatial coherence structure of a diffuse noise field to estimate per-microphone noise power spectral densities (PSDs) without VAD or knowledge of the target direction, enabling estimation during speech activity. However, the assumed diffuse noise model limits accuracy in the presence of directional noise sources.

3.3 Dictionary-Based Spatial Processing

The use of pre-measured or pre-computed acoustic transfer functions as a dictionary for spatial processing is a relatively recent development in the binaural hearing aid literature, motivated by the observation that individualized HRTFs capture the full directional filtering of the listener’s anatomy and can serve as accurate steering vectors without requiring real-time ATF estimation. This section reviews the principal methods that adopt this dictionary-based approach.

Pre-Determined RATF Beamforming

Koutrouvelis et al. [25] propose avoiding real-time RTF estimation entirely by constraining the binaural beamformer to preserve cues across a fixed set of pre-determined directions, using a measured RTF dictionary as the constraint set. The JBLCMV [26] and relaxed BLCMV [27] beamformers are shown to approximately preserve the binaural cues of the full acoustic scene in this setting without requiring source-specific RTF estimates.

This work is the most direct precursor to the dictionary framework adopted in this thesis, establishing that a pre-measured HRTF dictionary is a viable constraint set for binaural spatial filtering. Its key limitation, however, is that all N dictionary entries are enforced simultaneously as beamformer constraints, consuming all available degrees of freedom and leaving none for noise reduction. As N grows relative to the number of microphones M , the beamformer becomes severely over-constrained and noise reduction degrades proportionally. This directly motivates the selective approach of this thesis, i.e., rather than constraining all N entries, the proposed framework identifies only the d active-source entries, preserving $M - d$ degrees of freedom for noise suppression.

Automatic Interferer Selection

Kokke et al. [24] address the degree-of-freedom limitation of dictionary-constrained beamformers by proposing a convex optimization scheme that selects the subset of interferers whose binaural cues are most perceptually important to preserve, trading cue fidelity for less critical sources against additional degrees of freedom for noise reduction. This work shares the core motivation of the proposed framework in that selective dictionary use is preferable to exhaustive use, but approaches selection from a beamformer design perspective rather than an estimation perspective. Crucially, the selection is driven by perceptual importance criteria applied to a known or pre-assumed source set, whereas the proposed framework selects dictionary entries based on their consistency with the observed CPSDM, jointly determining which sources are active and their relative energy contributions.

HRTF-Based Source Localization and Beamforming

Talagala et al. [49] propose a composite binaural DOA estimator based on signal subspace decomposition that retains the full frequency-domain structure of the HRTF rather than reducing it to scalar ITD or ILD cues, aggregating subspace alignment

evidence across frequency subbands to produce robust estimates. This frequency aggregation strategy is conceptually related to the proposed algorithm’s atom selection step, though applied to a subspace alignment criterion rather than a quadratic form on the residual covariance.

Maazaoui et al. [33] apply pre-measured HRTFs to robot audition, using them to design fixed beamforming filters that account for head-induced array manifold distortions as a preprocessing stage for blind source separation. This work demonstrates that pre-measured HRTFs serve as effective steering vectors even for geometrically irregular arrays, directly supporting their use as dictionary atoms for behind-the-ear (BTE) hearing aid configurations.

Binaural Speech Enhancement with Spatial Scene Preservation

Thiemann et al. [53] propose a selective binaural beamformer that uses an SNR-controlled binary mask to switch between a binaural MVDR output and scaled reference microphone signals, demonstrating that target enhancement and spatial scene preservation can be achieved simultaneously, though binary switching introduces transition artifacts. Cornelis et al. [9] provide a theoretical analysis of binaural MWF variants, establishing that exact binaural cue preservation for both target and noise sources requires knowledge of the ITF of each source and, by extension, accurate ATF estimates. This result formally justifies the emphasis on ATF estimation accuracy in this thesis: without it, exact cue preservation is structurally unachievable regardless of beamformer design.

Van Den Bogaert et al. [57] provide perceptual evidence that noise reduction algorithms can significantly distort spatial hearing even at good noise reduction performance, and that augmenting the binaural MWF with a partial noise estimate restores correct localization of both target and noise components. This motivates the dual objective of noise reduction and spatial cue preservation that underpins the proposed framework.

3.4 Sparse Recovery for Source Localization

Sparse signal recovery methods have also been applied to source localization and speech enhancement, motivated by the observation that acoustic scenes are typically sparse in the directional domain: only a small number of sources are active simultaneously relative to the number of candidate directions in a spatial dictionary. This sparsity prior provides a powerful constraint that can be exploited to improve localization accuracy, reduce computational cost, and extend the range of conditions under which reliable spatial processing is achievable. This section reviews the principal applications of sparse recovery to source localization and speech enhancement, tracing the development from single-vector OMP-based methods to multi-vector and covariance-domain approaches.

OMP-Based Direction of Arrival Estimation

Aich and Palanisamy [1] formulate single-snapshot DOA estimation as a sparse recovery problem over a steering vector dictionary, applying OMP to recover the active source support. The method outperforms subspace methods such as MUSIC and ESPRIT in coherent source scenarios where spatial smoothing would otherwise be required, though its single-snapshot formulation limits robustness at low SNR due to the absence of temporal averaging.

Yang et al. [64] apply OMP to single-channel speech enhancement by sparsifying noisy speech frames in the discrete cosine transform (DCT) domain and recovering clean coefficients from compressive measurements, achieving competitive PESQ and MOS scores against subspace and spectral subtraction baselines. While distinct in domain, this work establishes the broader utility of OMP as a recovery tool in speech processing and demonstrates that transform-domain sparsity of speech can be exploited effectively under adverse noise conditions.

Multi-Source Localization with Matching Pursuit

Pavlidis et al. [41] propose a real-time source localization and counting method for circular arrays that applies matching pursuit to a histogram of per-frame DOA estimates accumulated over single-source-dominant time-frequency zones, jointly recovering active source directions and count. The method imposes relaxed sparsity requirements relative to standard compressed sensing formulations, requiring only the existence of single-source zones rather than global observation sparsity, and demonstrates strong performance in reverberant environments at low computational cost. This work shares the greedy pursuit philosophy of the proposed algorithm but operates on a histogram of instantaneous DOA estimates rather than directly on the CPSDM, and does not produce the source energy or noise covariance estimates required as inputs to a minimum variance beamformer.

Frequency-Weighted Subspace Methods

Gao et al. [18] propose a modified frequency-weighted MUSIC algorithm that replaces maximum eigenvalue weighting with SNR-based frequency band weighting, reducing the contribution of heavily overlapping bands and improving localization accuracy in multi-source scenarios where W-disjoint orthogonality is violated, without requiring prior knowledge of the source count. Beit-On et al. [4] address coherent broadband source localization within the coherent signal subspace framework, proposing a weighted frequency smoothing procedure with weights set inversely proportional to per-frequency source power, achieving superior source decorrelation over standard smoothing for equivalent bandwidth. The frequency aggregation concept in both works is conceptually related to the frequency-aggregated atom selection step of the proposed algorithm, though instantiated within a subspace framework rather than a greedy pursuit.

Compressed Sensing for DOA Estimation

Stockle et al. [48] extend compressed sensing-based DOA estimation to the 1-bit quantization setting relevant to massive MIMO systems, adapting Binary Iterative Hard Thresholding to complex-valued multiple measurement vector (MMV) problems and exploiting joint sparsity across snapshots to recover DOAs under extreme quantization. While remote from the binaural hearing aid context, this work demonstrates the generality of the MMV sparse recovery framework and the practical value of joint sparsity when individual measurements are severely degraded.

3.5 Joint and Unified Estimation Frameworks

The preceding sections have reviewed methods that address subsets of the joint estimation problem: ATF estimation, noise covariance estimation, dictionary-based spatial processing, and sparse source localization. A recurring theme across all of these is that methods addressing one sub-problem typically assume the others are solved or known, creating dependencies that introduce inconsistencies and error propagation in practice. This section reviews the most closely related family of methods that attempt to address two or more sub-problems simultaneously, examining how far each goes toward the fully joint estimation goal and where each falls short.

Alternating Optimization for Joint RTF and PSD Estimation

Tammen et al. [50] jointly estimate the relative early transfer function (RETF) vector and all source and noise PSDs via alternating least squares minimization of a Frobenius norm objective, decoupling the joint problem into closed-form sub-problems that converge to a locally consistent solution outperforming covariance whitening in dynamic scenarios. Two limitations constrain its applicability in the present context. First, the method addresses a single-source scenario with diffuse noise; extension to d simultaneously active point sources increases optimization complexity and local minima risk. Second, the estimated RETF is unconstrained and continuous-valued, making it incompatible with dictionary-referenced beamformer constraints that require consistency with pre-measured HRTF atoms.

Confirmatory Factor Analysis for Joint Signal Model Estimation

Koutrouvelis et al. [28] propose the most comprehensive existing treatment of joint signal model parameter estimation, using confirmatory factor analysis (CFA) to simultaneously estimate source PSDs, ATFs, late reverberation PSD, and microphone self-noise PSDs from the observed CPSDM via structured covariance fitting, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \left\| \hat{\mathbf{R}}_{\mathbf{x}}(k) - \mathbf{R}_{\mathbf{x}}(k; \boldsymbol{\theta}) \right\|_F^2, \quad (3.4)$$

solved by alternating iterative updates over parameter subsets. CFA is the closest existing precursor to the proposed framework: both operate on the observed CPSDM

and jointly recover ATF vectors, source energies, and noise covariance without separate noise estimation or disjoint activity periods. The critical distinction lies in how ill-posedness is resolved. CFA imposes a parametric noise model and estimates ATF vectors as continuous-valued parameters, whereas the proposed framework constrains ATF vectors to a finite pre-measured HRTF dictionary, recasting continuous parameter estimation as discrete support recovery amenable to greedy pursuit. This dictionary constraint simultaneously reduces the degrees of freedom of the problem, ensures recovered ATF vectors are directly usable as beamformer steering constraints, and yields a non-iterative solution free from convergence and initialization sensitivity.

Structured Linear Algebraic Methods

Li et al. [31] address the computational burden of CFA-based approaches by proposing a multi-source RTF estimator grounded in linear algebraic operations on observed covariance matrices. Rather than solving a non-convex joint likelihood, the method exploits the algebraic structure of the multi-source covariance model to decouple RTF estimation into a sequence of tractable linear operations, avoiding the cost of full statistical optimization while retaining joint multi-source estimation capability absent in single-source methods. Experimental results demonstrate accuracy comparable to Koutrouvelis et al. [28] at significantly reduced computational load, suggesting that the algebraic structure of the covariance model is sufficient for accurate RTF estimation without full probabilistic modeling.

Covariance Learning with Greedy Pursuit

Ollila [38, 37] proposes a covariance learning (CL) framework for sparse signal recovery in the MMV model, which is the closest existing method to the proposed algorithm in terms of problem formulation and algorithmic strategy. The CL framework treats signal and noise as independent Gaussians with unknown signal covariance and noise variance, recovering the sparse support by maximizing the likelihood over the support set via a greedy pursuit analogous to OMP. The atom selection criterion in these works reduces under isotropic noise to a quantity structurally similar to the quadratic form $\mathbf{h}_n^H \mathbf{R}_{res} \mathbf{h}_n$, and the residual is updated by deflating the covariance contribution of the selected atom, mirroring (4.14).

Despite these structural parallels, three key distinctions separate the two approaches. First, the CL framework assumes scalar isotropic noise $\mathbf{R}_n(k) = \sigma^2 \mathbf{I}$, whereas the proposed algorithm estimates an unstructured residual noise covariance, making it applicable to spatially correlated noise fields. Second, the CL framework is not specialized to the binaural HRTF estimation and speech enhancement setting. Third, and most critically, the CL framework does not yield the full noise covariance matrix required by the JBLCMV beamformer, whereas the proposed framework directly produces an unstructured residual covariance suitable as beamformer input.

Maximum Likelihood Joint Estimation

Hoang et al. [22] derive joint ML estimates of the RTFs and PSDs of all sources in a multi-source scene, as introduced in Section 3.1. The work is notable for establishing the statistical optimality of covariance whitening as an ML estimator under known noise covariance, and for integrating VAD-based activity detection within the ML framework. However, the noise covariance is treated as a known input rather than as a jointly estimated quantity, which is precisely the gap the proposed framework addresses.

Li and Hendriks [30] propose a robust joint estimator for source RTFs and PSDs in multi-source reverberant scenarios, exploiting cross-frame variance of sample covariance matrices to match the performance of confirmatory factor analysis at substantially lower computational cost. Like the CFA approach, however, the method estimates continuous-valued RTF vectors without a dictionary constraint and requires a parametric noise model, producing no unstructured noise covariance estimate. Its primary contribution over existing methods is computational efficiency rather than a fundamentally different estimation strategy.

3.6 Gap Analysis and Positioning

The preceding sections have surveyed the state of the art across four areas relevant to the proposed framework: ATF and RTF estimation, noise covariance estimation, dictionary-based spatial processing, and joint estimation frameworks. This section synthesizes the findings of that survey into a structured gap analysis, identifying the specific combination of properties that no existing method provides and that the proposed framework is designed to address.

Properties of an Ideal Framework

Drawing on the beamformer requirements of Section 2.5 and the limitations identified throughout this chapter, an ideal joint estimation framework for binaural speech enhancement should simultaneously satisfy the following properties.

P1: Joint estimation without separate noise reference. ATF vectors, source energies, and noise covariance should be estimated jointly from the observed mixture CPSDM, eliminating the error propagation that arises from sequential estimation across different data segments.

P2: Dictionary-constrained ATF recovery. Estimated ATF vectors should be constrained to a pre-measured HRTF dictionary, ensuring direct compatibility with the beamformer’s constraint set while simultaneously regularizing the ill-posed estimation problem.

P3: General noise covariance output. The noise covariance estimate should be a full Hermitian positive semi-definite matrix, accommodating spatially correlated, directional, and non-stationary noise fields without imposing isotropic or diffuse parametric assumptions.

P4: Multi-source capability. The framework should jointly recover ATF vectors and energies for $d > 1$ simultaneously active sources, rather than decomposing multi-source scenarios into sequential single-source problems.

Table 3.1: Comparison of existing methods against the five properties of an ideal joint estimation framework. ✓: satisfied. ○: partially satisfied. ✗: not satisfied.

Method	Parameter(s) Estimated	P1	P2	P3	P4	P5
Taseska & Habets [52]	ATF/RTF	✗	✗	✗	○	✓
Serizel et al. [46]	Noise CPSDM	✗	✗	✓	○	✓
Tammen et al. [50]	ATF/RTF, CPSDM	○	✗	○	✗	✓
Koutrouvelis et al. [28]	ATF/RTF, Noise CPSDM	✓	✗	○	✓	✗
Koutrouvelis et al. [25]	ATF/RTF	✗	✓	✗	✓	✓
Kokke et al. [24]	ATF/RTF	✗	✓	✗	○	✓
Ollila [38]	CPSDM	✓	✗	✗	✓	✓
Li & Hendriks [30]	ATF/RTF, CPSDM	○	✗	○	✓	✓
Hoang et al. [22]	ATF/RTF, CPSDM	○	✗	✗	✓	✓
Proposed		✓	✓	✓	✓	✓

P5: Computational efficiency. The framework should be tractable for real-time or near-real-time processing in a hearing aid context, avoiding the initialization sensitivity and convergence overhead of iterative optimization-based approaches.

Table 3.1 summarizes how the most closely related existing methods satisfy each of the five properties identified above. A method is marked as satisfying a property if it addresses it as a primary contribution, partially satisfying it if it addresses it under restrictive assumptions, and not satisfying it if it explicitly assumes the property holds or does not address it.

Identified Gaps

Table 3.1 reveals a consistent pattern: no existing method satisfies all five properties simultaneously. Methods that achieve joint estimation without a separate noise reference (P1) do so at the cost of a parametric noise model (P3) or without constraining ATF vectors to a dictionary (P2). Methods that use a pre-measured HRTF dictionary (P2) require a separately estimated noise covariance (P1) and do not estimate source energies or noise covariance as outputs (P3). Methods that are computationally efficient (P5) generally sacrifice either joint estimation (P1) or dictionary constraint (P2), while the most comprehensive joint estimator requires iterative convergence that may be costly for real-time processing (P5).

The Proposed Framework in Context

The proposed framework addresses all five properties simultaneously through a single design decision: extending the greedy S-OMP pursuit to the covariance domain with a pre-measured HRTF dictionary as the atom set. Joint estimation without a separate noise reference (P1) follows because the noise covariance emerges as the pursuit residual. Dictionary-constrained ATF recovery (P2) is enforced by construction, since only dictionary atoms are selected. An unstructured noise covariance (P3) is produced because the residual carries no parametric constraint on its spatial structure.

Multi-source capability (P4) follows from iterating the pursuit over d selections, and computational efficiency (P5) from terminating the algorithm after exactly d iterations, with closed-form updates at each step.

The closest existing method is the covariance learning pursuit of Ollila [37], which shares the covariance-domain greedy strategy but assumes isotropic noise (P3), imposes no dictionary constraint (P2), and targets generic sparse recovery rather than binaural HRTF estimation. The proposed framework can thus be understood as extending the covariance learning pursuit to the binaural setting, augmenting it with a dictionary constraint that simultaneously regularizes estimation and guarantees beamformer-ready outputs, and replacing the isotropic noise assumption with an unstructured residual covariance suited to the spatially correlated noise fields encountered in practice.

Chapter Summary

This chapter surveyed the state of the art in acoustic transfer function estimation, noise covariance estimation, dictionary-based spatial processing, sparse recovery, and joint estimation frameworks, identifying the contributions and limitations of each method family with respect to the proposed framework’s requirements. The survey established that no existing method simultaneously satisfies all five ideal properties: joint estimation without a separate noise reference, dictionary-constrained ATF recovery, unstructured noise covariance output, multi-source capability, and computational efficiency. Methods achieving joint estimation typically rely on parametric noise models or unconstrained continuous-valued ATF estimates, while dictionary-based methods do not jointly estimate noise covariance or source energies. The covariance learning pursuit of Ollila [37] is the closest algorithmic analogue, sharing the covariance-domain greedy strategy but differing in noise model, dictionary constraint, and application context.

This gap directly motivates the proposed framework of Chapter 4, which addresses all five properties simultaneously through a coherent design grounded in the theoretical foundations of Chapter 2, with practical performance evaluated against sequential baselines in Chapter 5.

Covariance-Based Simultaneous Orthogonal Matching Pursuit

4

As introduced in Chapter 2, OMP [55] is a greedy algorithm for sparse signal recovery that iteratively selects dictionary atoms by correlation with the current residual, augments the active support, and updates the residual via orthogonal projection. S-OMP [56] extends this to the MMV setting where observations share a common sparse support, by aggregating correlation scores across measurement vectors to promote jointly sparse solutions.

However, they operate directly on signal coefficients, an assumption difficult to satisfy in the binaural speech enhancement setting where individual time-frequency realizations of the microphone mixture are highly stochastic. A more statistically robust alternative is to work with second-order statistics: estimating the CPSDM $\mathbf{R}_\mathbf{x}(k) \in \mathbb{C}^{M \times M}$ over a block of L frames suppresses noise perturbations via temporal averaging, yielding a stable representation of the spatial covariance structure on which the pursuit can operate.

This shift from the coefficient domain to the covariance domain has well-established precedent in array processing. MUSIC [45] and ESPRIT [44] exploit the eigenstructure of the sample covariance to separate signal and noise subspaces; in particular, the MUSIC pseudospectrum evaluates the quadratic form $\mathbf{h}^H \mathbf{P}_\mathbf{n} \mathbf{h}$ for each candidate steering vector against the noise projector. The atom selection step proposed here is structurally analogous: we evaluate $\tilde{\mathbf{h}}_n^H \mathbf{R}_{res}^{(p-1)}(k) \tilde{\mathbf{h}}_n$, for each dictionary atom, measuring the beamformer output power of atom n applied to the residual covariance. Unlike MUSIC, however, this scan is embedded within an iterative pursuit: at each step the selected source contribution is projected out of the residual, progressively deflating the covariance until the stopping criterion is met.

The proposed algorithm takes as input $\mathbf{R}_\mathbf{x}(k)$, the normalized HRTF dictionary $\tilde{\mathcal{A}}(k)$, and a stopping criterion, and returns the estimated source energy matrix $\hat{\mathbf{R}}_\mathbf{s}(k)$, residual covariance $\mathbf{R}_{res}(k)$, support set $\mathbf{\Lambda}$, and iteration count p . These outputs directly supply the quantities required by a binaural beamformer: $\hat{\mathbf{R}}_\mathbf{s}(k)$ provides the source covariance, $\mathbf{R}_{res}(k)$ the noise covariance, and $\mathbf{\Lambda}$ the active steering vectors. The algorithm is initialized with $\mathbf{R}_{res}^{(0)}(k) = \mathbf{R}_\mathbf{x}(k)$ and $\mathbf{\Lambda}^{(0)} = \emptyset$, and iterates through atom selection, signal energy update, and residual update steps described in the subsections that follow.

4.1 Inputs, Outputs and Design Choices

The proposed algorithm operates on three inputs: the measured CPSDM $\mathbf{R}_\mathbf{x}(k) \in \mathbb{C}^{M \times M}$, the dictionary of HRTFs $\mathcal{A}(k) \in \mathbb{C}^{M \times N}$, and a stopping criterion. Each of these carries design decisions that are worth making explicit, as they directly influence

the robustness and interpretability of the algorithm’s outputs.

The CPSDM as the Input Observation

Rather than operating on raw time-frequency signal vectors, the algorithm takes as input the CPSDM estimated from a block of L consecutive STFT frames, as also discussed in Section 2.2. Therefore, we will be using (2.9) to calculate the sample estimate and consequently the starting point for the proposed algorithm.

Dictionary Normalization

As discussed in Section 2.4, each atom in $\mathcal{A}(k)$ is normalized along the channel dimension (2.16) to form $\tilde{\mathcal{A}}(k)$, i.e., we treat $\tilde{\mathbf{h}}_n = \mathbf{h}_n / \|\mathbf{h}_n\|_2$ as a dictionary atom. Without normalization, the quadratic form $\mathbf{h}_n^H \mathbf{R}_{res} \mathbf{h}_n$ would conflate directional energy concentration with raw atom magnitude, biasing selection irrespective of alignment with the residual covariance, analogously to the unit-norm requirement in classical OMP [55]. Although HRTF magnitudes vary across directions and frequencies, discarding this information is an accepted tradeoff: source energies are recovered separately (after undoing normalization) in the signal energy update step, which explicitly estimates the scaling factors required to reconstruct the observed covariance.

Outputs and Their Role in the Beamforming Framework

The algorithm outputs together constitute a complete parameterization of the acoustic scene and map directly onto the inputs required by a binaural beamformer such as GBLCMV [27]: $\mathbf{\Lambda}$ identifies the active-source columns of $\tilde{\mathcal{A}}(k)$ (and thus $\mathcal{A}(k)$ as well), supplying the requisite steering vectors; the diagonal elements of $\hat{\mathbf{R}}_s(k)$ provide per-source energy estimates, yielding spatial covariance contributions of the form $\sigma_n^2(k) \tilde{\mathbf{h}}_n(k) \tilde{\mathbf{h}}_n^H(k)$; and $\mathbf{R}_{res}(k)$ captures the unexplained residual, interpretable as the noise covariance under the assumed signal model. Fig. 4.1 shows such a framework in action. The algorithm thus acts as a self-contained front-end that simultaneously performs source selection and furnishes all inputs required for subsequent spatial filtering, without separate HRTF estimation or noise covariance measurement.

4.2 Proposed Algorithm

We will consider (2.7) as the base of the proposed algorithm, which solves the problem defined in Chapters 2 and 3 through a greedy iterative pursuit extending S-OMP to the covariance domain. Each iteration comprises three steps: atom selection, identifying the dictionary entry most consistent with the current residual covariance; signal energy update, re-estimating all source energies jointly; and residual update, deflating the covariance by the contribution of the selected sources. Re-estimating all coefficients at each iteration ensures the residual remains orthogonal to the subspace spanned by the selected atoms, preventing redundant selections analogously to the projection mechanism of classical S-OMP. The key departure is that all steps operate on covariance

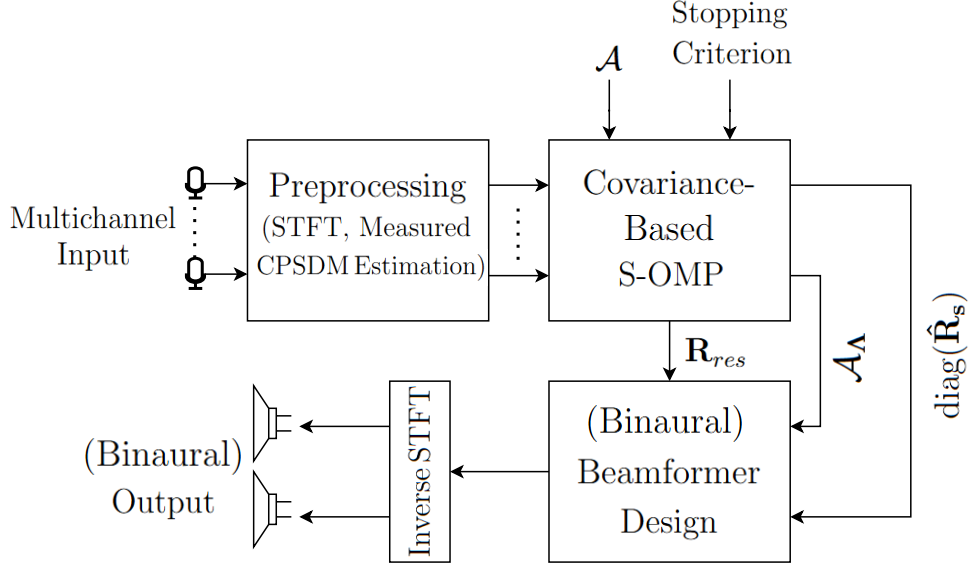


Figure 4.1: A beamforming framework using the proposed algorithm, offering an alternative to sequential estimation of quantities required by a beamformer.

matrices rather than signal vectors: the residual is a matrix initialized as $\mathbf{R}_{res}^{(0)}(k) = \mathbf{R}_x(k)$, atom selection uses a quadratic form in place of an inner product, and coefficient estimation becomes a matrix least-squares problem. The full procedure is summarized in Algorithm 1, with each step derived in the following sections.

4.3 Atom Selection

At iteration p , the goal of the atom selection step is to identify the dictionary atom $\mathbf{h}_n(k)$ whose direction best explains the spatial structure of the current residual covariance $\mathbf{R}_{res}^{(p-1)}(k)$. We seek the index

$$\lambda^{(p)} = \arg \max_{1 \leq n \leq N} f_n(\mathbf{R}_{res}^{(p-1)}), \quad (4.1)$$

for some scoring function $f_n(\cdot)$ that measures the alignment of atom n with the residual. The choice of this scoring function is the central design decision of the matching step, and we derive it from first principles below.

Derivation of the Quadratic Form

The scalar quantity

$$f_n(k) = \tilde{\mathbf{h}}_n^H(k) \mathbf{R}_{res}^{(p-1)}(k) \tilde{\mathbf{h}}_n(k) \quad (4.2)$$

is the output power of a data-independent beamformer with weight vector $\tilde{\mathbf{h}}_n(k)$ applied to the residual covariance. Since $\mathbf{R}_{res}^{(p-1)}(k)$ is Hermitian positive semi-definite, $f_n(k) \geq 0$, with large values when $\tilde{\mathbf{h}}_n(k)$ aligns with a direction of high residual energy and

Algorithm 1 Covariance-Based Simultaneous Orthogonal Matching Pursuit

Require: Measured CPSD tensor $\underline{\mathbf{R}}_{\mathbf{x}} \in \mathbb{C}^{M \times M \times K}$, (normalized) dictionary $\tilde{\mathbf{A}} \in \mathbb{C}^{M \times N \times K}$, stopping criterion

Ensure: Estimated signal CPSD tensor $\hat{\underline{\mathbf{R}}}_{\mathbf{s}} \in \mathbb{C}^{d \times d \times K}$, residual CPSD tensor $\underline{\mathbf{R}}_{res} \in \mathbb{C}^{M \times M \times K}$, support set $\mathbf{\Lambda}$, iteration count p

- 1: $p \leftarrow 0$
- 2: $\underline{\mathbf{R}}_{res} \leftarrow \underline{\mathbf{R}}_{\mathbf{x}}$
- 3: $\mathbf{\Lambda} \leftarrow \emptyset$
- 4: **while** stopping criterion not met **do**
- 5: $p \leftarrow p + 1$
- 6: **Matching I:**
- 7: **for** $k = 1$ to K **do**
- 8: Calculate correlations as per (4.5)
- 9: **end for**
- 10: **Matching II:** Summing up and identifying support as per (4.6), (4.7)
- 11: **Updating signal CPSD tensor:**
- 12: **for** $k = 1$ to K **do**
- 13: Calculate as per (4.10)
- 14: **end for**
- 15: **Updating residual CPSD tensor:**
- 16: **for** $k = 1$ to K **do**
- 17: Calculate as per (4.12), (4.14)
- 18: **end for**
- 19: **end while**

small values otherwise, precisely the behavior desired of a matching criterion. This is the complement of the MUSIC criterion [45], which evaluates $\mathbf{h}^H \mathbf{P}_{\mathbf{n}} \mathbf{h}$ and selects directions of minimum noise projection; here we instead maximize residual energy. The two are dual: both exploit the covariance structure of the observed mixture, and in the noiseless case would yield identical selections. In the presence of noise, however, the two diverge: MUSIC explicitly isolates the noise subspace via eigendecomposition and nulls it, so signal directions remain sharp zeros of $\mathbf{h}^H \mathbf{P}_{\mathbf{n}} \mathbf{h}$ even under additive noise. Meanwhile, $f_n(k)$ is evaluated directly against $\underline{\mathbf{R}}_{res}^{(p-1)}(k)$, which conflates residual signal and noise energy, raising the baseline uniformly across directions and broadening rather than sharpening the peak at the dominant source.

Frequency Aggregation

In a multi-frequency setting, a broadband source manifests consistently across all K bins, whereas a spurious noise-driven match produces elevated scores only in isolated bins. Scores are therefore aggregated before maximization, i.e.,

$$\lambda^{(p)} = \arg \max_{1 \leq n \leq N} \left| \sum_{k=1}^K \tilde{\mathbf{h}}_n^H(k) \underline{\mathbf{R}}_{res}^{(p-1)}(k) \tilde{\mathbf{h}}_n(k) \right|. \quad (4.3)$$

This is the covariance-domain analogue of the joint sparsity assumption in S-OMP: atoms are selected whose beamformer output power is consistently high across frequency bins, favoring true broadband sources over spurious per-bin matches.

Alternate Notation via Matrix Projection

All N per-bin scores can be obtained simultaneously by noting that

$$\left[\tilde{\mathbf{h}}_1^H \mathbf{R}_{res}^{(p-1)} \tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_N^H \mathbf{R}_{res}^{(p-1)} \tilde{\mathbf{h}}_N \right] = \text{diag} \left(\tilde{\mathcal{A}}^H(k) \mathbf{R}_{res}^{(p-1)}(k) \tilde{\mathcal{A}}(k) \right), \quad (4.4)$$

rewriting N separate $\mathcal{O}(M^2)$ evaluations to a single matrix product of cost $\mathcal{O}(M^2N)$ followed by an $\mathcal{O}(N)$ diagonal extraction. The aggregated scores and support update are then

$$\mathbf{c}^{(p)}(k) = \text{diag} \left(\tilde{\mathcal{A}}^H(k) \mathbf{R}_{res}^{(p-1)}(k) \tilde{\mathcal{A}}(k) \right), \quad (4.5)$$

$$\mathbf{c}_{agg}^{(p)} = \sum_{k=1}^K \mathbf{c}^{(p)}(k), \quad \lambda^{(p)} = \arg \max_n |(c_{agg}^{(p)})_n|. \quad (4.6)$$

This expresses the entire matching step as a sequence of standard linear algebra operations, amenable to optimized matrix multiplication routines in matrix-oriented environments.

Support Augmentation

Once $\lambda^{(p)}$ is identified, the support set is augmented as

$$\mathbf{\Lambda}^{(p)} = \mathbf{\Lambda}^{(p-1)} \cup \{\lambda^{(p)}\}, \quad (4.7)$$

and the sub-dictionary of selected atoms $\tilde{\mathcal{A}}_{\mathbf{\Lambda}}(k) \in \mathbb{C}^{M \times p}$ is formed by extracting the columns of $\tilde{\mathcal{A}}(k)$ indexed by $\mathbf{\Lambda}^{(p)}$. This sub-dictionary is passed to the signal energy update step, described next.

4.4 Signal Energy Update

Having identified the support set $\mathbf{\Lambda}^{(p)}$ at iteration p , the algorithm re-estimates the source signal covariance matrix using all selected atoms jointly. This is performed from scratch at every iteration rather than incrementally, which is the defining characteristic of the orthogonal projection mechanism inherited from OMP: by solving a fresh least-squares problem at each step, the estimates of previously selected sources are corrected in light of newly added atoms, preventing the accumulation of greedy errors.

Problem Formulation

Substituting the sub-dictionary $\tilde{\mathcal{A}}_{\mathbf{\Lambda}}(k) \in \mathbb{C}^{M \times p}$ of p selected atoms for the full ATF matrix $\mathbf{A}(k)$ in (2.7), the signal model restricted to the current support is

$$\mathbf{R}_{\mathbf{x}}(k) \approx \tilde{\mathcal{A}}_{\mathbf{\Lambda}}(k) \mathbf{R}_{\mathbf{s}}(k) \tilde{\mathcal{A}}_{\mathbf{\Lambda}}^H(k) + \mathbf{R}_{\mathbf{n}}(k), \quad (4.8)$$

where $\mathbf{R}_s(k) \in \mathbb{C}^{p \times p}$ is now the source covariance restricted to the p selected directions. Since $\mathbf{R}_n(k)$ is unknown, the signal energy update minimizes the Frobenius-norm discrepancy between the observed CPSDM and the noiseless model, yielding the least-squares problem

$$\hat{\mathbf{R}}_s^{(p)}(k) = \arg \min_{\mathbf{R}_s(k)} \left\| \mathbf{R}_x(k) - \tilde{\mathcal{A}}_\Lambda(k) \mathbf{R}_s(k) \tilde{\mathcal{A}}_\Lambda^H(k) \right\|_F^2. \quad (4.9)$$

The Closed-Form Solution

(4.9) can be solved by exploiting the structure of the objective, the derivation of which is shown in Appendix A.1. The closed-form solution derived from there is

$$\hat{\mathbf{R}}_s^{(p)}(k) = \tilde{\mathcal{A}}_\Lambda^\dagger(k) \mathbf{R}_x(k) \tilde{\mathcal{A}}_\Lambda^{\dagger H}(k), \quad (4.10)$$

where $\tilde{\mathcal{A}}_\Lambda^\dagger(k) = \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k) \in \mathbb{C}^{p \times M}$ is the Moore-Penrose pseudoinverse of $\tilde{\mathcal{A}}_\Lambda(k)$.

Interpretation

The pseudoinverse $\tilde{\mathcal{A}}_\Lambda^\dagger(k)$ implements an oblique projection from the M -dimensional observation space onto the p -dimensional source space, recovering the source covariance up to the noise contribution projected through the pseudoinverse. In the noiseless case, (4.10) recovers $\mathbf{R}_s(k)$ exactly when the true ATFs lie in the dictionary. Under source independence, $\hat{\mathbf{R}}_s^{(p)}(k)$ should be diagonal, with $[\hat{\mathbf{R}}_s^{(p)}(k)]_{ii}$ estimating the power spectral density of source i at bin k ; off-diagonal elements, arising from noise, finite block length, or dictionary mismatch, represent estimated cross-source coherence and serve as a diagnostic of model fit but are not used directly by the beamformer. Note that the normalization of HRTFs done during initialization has to be undone to recover the true estimates of signal energies; not doing so would lead to a physical irrelevance of these quantities.

Full Column Rank and Dictionary Coherence

The derivation above assumes that the selected atoms are linearly independent, i.e. that $\tilde{\mathcal{A}}_\Lambda(k)$ has full column rank p at every bin k . This condition is satisfied when no two selected atoms are collinear, which in turn depends on the coherence of the HRTF dictionary (2.17). The interpretation of $\mu(\tilde{\mathcal{A}}(k))$ from Section 2.4 holds true here. In practice, high coherence sets a lower bound on the resolvable angular separation between sources, which is analogous to the Rayleigh resolution limit in array processing. This limitation is acknowledged and its practical impact is assessed in the experimental evaluation of Chapter 5.

4.5 Residual Update

Following the signal energy update, the algorithm computes a new residual covariance matrix that reflects the portion of the observed CPSDM not explained by the p sources

selected so far, which serves a dual purpose: it drives atom selection in the next iteration by deflating the covariance contribution of already-selected sources, and at termination it constitutes the “noise” covariance estimate supplied to the beamformer; the “noise” being all components of the signal mixture not attributed to the p selected sources. We will use the orthogonal projector onto the column space of $\tilde{\mathcal{A}}_{\Lambda}(k)$ for the remainder of this section, defined as

$$\mathbf{P}_{\Lambda}(k) = \tilde{\mathcal{A}}_{\Lambda}(k) \tilde{\mathcal{A}}_{\Lambda}^{\dagger}(k) = \tilde{\mathcal{A}}_{\Lambda}(k) \left(\tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k) \tilde{\mathcal{A}}_{\Lambda}(k) \right)^{-1} \tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k). \quad (4.11)$$

The verification of $\mathbf{P}_{\Lambda}(k)$ as an orthogonal projector is shown in Appendix A.2.

Estimating the Explained Covariance

The covariance contribution of the p selected sources to the observed CPSDM is estimated by projecting $\mathbf{R}_{\mathbf{x}}(k)$ onto the column space of $\tilde{\mathcal{A}}_{\Lambda}(k)$ from both sides:

$$\hat{\mathbf{R}}_{\mathbf{y}}^{(p)}(k) = \mathbf{P}_{\Lambda}(k) \mathbf{R}_{\mathbf{x}}(k) \mathbf{P}_{\Lambda}^{\text{H}}(k). \quad (4.12)$$

This two-sided projection is not arbitrary. Substituting the restricted signal model (4.8) into (4.12) and noting that $\mathbf{P}_{\Lambda}(k) \tilde{\mathcal{A}}_{\Lambda}(k) = \tilde{\mathcal{A}}_{\Lambda}(k)$ by definition of the projector, we obtain

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{y}}^{(p)}(k) &= \mathbf{P}_{\Lambda}(k) \left(\tilde{\mathcal{A}}_{\Lambda}(k) \mathbf{R}_{\mathbf{s}}(k) \tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k) + \mathbf{R}_{\mathbf{n}}(k) \right) \mathbf{P}_{\Lambda}^{\text{H}}(k) \\ &= \tilde{\mathcal{A}}_{\Lambda}(k) \mathbf{R}_{\mathbf{s}}(k) \tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k) + \mathbf{P}_{\Lambda}(k) \mathbf{R}_{\mathbf{n}}(k) \mathbf{P}_{\Lambda}^{\text{H}}(k). \end{aligned} \quad (4.13)$$

In the noiseless case, $\hat{\mathbf{R}}_{\mathbf{y}}^{(p)}(k)$ recovers the true source covariance contribution exactly. In the noisy case, it additionally contains the noise covariance projected onto the signal subspace, $\mathbf{P}_{\Lambda}(k) \mathbf{R}_{\mathbf{n}}(k) \mathbf{P}_{\Lambda}^{\text{H}}(k)$, which represents noise energy leaking into the selected subspace. This leakage is unavoidable when the signal and noise subspaces overlap, and is a fundamental limitation of subspace-based methods operating at finite SNR.

Residual Covariance Update

The residual covariance is defined as the component of the observed CPSDM not captured by the projection:

$$\mathbf{R}_{res}^{(p)}(k) = \mathbf{R}_{\mathbf{x}}(k) - \hat{\mathbf{R}}_{\mathbf{y}}^{(p)}(k) = (\mathbf{I} - \mathbf{P}_{\Lambda}(k)) \mathbf{R}_{\mathbf{x}}(k) (\mathbf{I} - \mathbf{P}_{\Lambda}(k))^{\text{H}}. \quad (4.14)$$

Substituting (4.13), the residual expands to

$$\begin{aligned} \mathbf{R}_{res}^{(p)}(k) &= \mathbf{R}_{\mathbf{x}}(k) - \tilde{\mathcal{A}}_{\Lambda}(k) \mathbf{R}_{\mathbf{s}}(k) \tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k) - \mathbf{P}_{\Lambda}(k) \mathbf{R}_{\mathbf{n}}(k) \mathbf{P}_{\Lambda}^{\text{H}}(k) \\ &= \mathbf{R}_{\mathbf{n}}(k) + (\mathbf{I} - \mathbf{P}_{\Lambda}(k)) \mathbf{R}_{\mathbf{n}}(k) (\mathbf{I} - \mathbf{P}_{\Lambda}(k))^{\text{H}} - \mathbf{R}_{\mathbf{n}}(k) \\ &= (\mathbf{I} - \mathbf{P}_{\Lambda}(k)) \mathbf{R}_{\mathbf{n}}(k) (\mathbf{I} - \mathbf{P}_{\Lambda}(k))^{\text{H}}, \end{aligned} \quad (4.15)$$

revealing that in the noiseless case $\mathbf{R}_{res}^{(p)}(k) \rightarrow \mathbf{0}$ as $p \rightarrow d$, and in the noisy case the residual converges to the noise covariance projected onto the complement of the signal subspace. This confirms the interpretability of the residual as a noise covariance estimate.

Positive Semi-Definiteness of the Residual

A necessary condition for $\mathbf{R}_{res}^{(p)}(k)$ to be interpretable as a covariance matrix is that it remains positive semi-definite (PSD) throughout the iteration. This is guaranteed by construction. Since $\mathbf{R}_x(k) \succeq \mathbf{0}$ and $\hat{\mathbf{R}}_y^{(p)}(k) = \mathbf{P}_\Lambda(k)\mathbf{R}_x(k)\mathbf{P}_\Lambda^H(k)$ is a congruence transformation of a PSD matrix, $\hat{\mathbf{R}}_y^{(p)}(k) \succeq \mathbf{0}$. It then follows from (4.14) that for any vector $\mathbf{v} \in \mathbb{C}^M$,

$$\mathbf{v}^H \mathbf{R}_{res}^{(p)}(k) \mathbf{v} = \mathbf{v}^H \mathbf{R}_x(k) \mathbf{v} - \mathbf{v}^H \hat{\mathbf{R}}_y^{(p)}(k) \mathbf{v} \geq 0 \quad (4.16)$$

provided that $\hat{\mathbf{R}}_y^{(p)}(k) \preceq \mathbf{R}_x(k)$, i.e. the explained covariance does not exceed the observed covariance. This condition holds as long as $\mathbf{R}_x(k)$ is consistent with the signal model and the selected atoms span a subspace of the observation space, both of which are satisfied by construction.

Interpretation as a Noise Covariance Estimate

Unlike conventional beamformers requiring a separate noise-only measurement period or a parametric noise model, the proposed algorithm estimates the noise covariance jointly with source selection, requiring no input beyond the observed mixture. If interferers are to be included to form an “interference plus noise” covariance, we can add source covariances of the form $\sigma^2 \mathbf{h} \mathbf{h}^H$ to the residual. At the same time, the quality of this estimate depends critically on support recovery accuracy: premature termination leaves residual source energy in $\mathbf{R}_{res}^{(p)}(k)$, overestimating the noise covariance, while excess iterations absorb noise energy into the signal subspace, underestimating it. The stopping criterion therefore directly governs noise covariance quality, and its design is discussed in Section 4.6.

4.6 Stopping Criterion

The stopping criterion governs when the iterative pursuit terminates, and therefore determines how many atoms are selected, how many source energies are estimated, and crucially, what portion of the observed CPSDM is assigned to the residual noise covariance estimate. Its design is a fundamental parameter of the framework whose consequences propagate directly into beamformer performance.

Known Sparsity Criterion

The algorithm is terminated after exactly d iterations, where it is assumed to be known, i.e.,

$$p = d, \quad (4.17)$$

yielding a support set Λ of exactly d indices, source energy matrix $\hat{\mathbf{R}}_s^{(d)}(k) \in \mathbb{C}^{d \times d}$, and noise covariance estimate $\mathbf{R}_{res}^{(d)}(k)$. This is a direct comparison to the sparsity-constrained stopping rule in classical OMP and S-OMP. The assumption of known d is standard in both the sparse recovery literature and array processing evaluations, where it isolates recovery performance from model order estimation errors.

Consequences of Incorrect Sparsity

Although d is assumed to be known in this work, it is instructive to characterize the two failure modes arising from model order errors.

Underfitting ($p < d$). Premature termination leaves residual source energy in $\mathbf{R}_{res}^{(p)}(k)$, overestimating the noise covariance. The beamformer consequently suppresses unmodeled sources as interference, and the incomplete support set $\mathbf{\Lambda}^{(p)}$ causes the spatial filter to steer toward only a subset of true sources.

Overfitting ($p > d$). Iterating beyond the true source count causes the pursuit to select noise-aligned atoms, depleting $\mathbf{R}_{res}^{(p)}(k)$ of noise energy and underestimating $\mathbf{R}_n(k)$. In the extreme case $p = M$, the residual is driven to zero, yielding an ill-conditioned spatial filter; spurious energies are additionally assigned to noise directions, corrupting $\hat{\mathbf{R}}_s^{(p)}(k)$.

The two failure modes are asymmetric: underfitting degrades signal enhancement by missing sources, while overfitting destabilizes the spatial filter by corrupting the noise covariance estimate.

Relation to Beamformer Rank Constraints

After $p = d$ iterations, the estimated signal covariance $\tilde{\mathbf{A}}_{\mathbf{\Lambda}}(k)\hat{\mathbf{R}}_s^{(d)}(k)\tilde{\mathbf{A}}_{\mathbf{\Lambda}}^H(k)$ has rank at most d , and $\mathbf{R}_{res}^{(d)}(k)$ has rank at most $M - d$. For beamformers such as LCMV requiring an invertible noise covariance, this necessitates $d < M$, a standard array processing constraint satisfied in all scenarios considered here. When rank deficiency is a concern, diagonal loading restores invertibility, i.e.,

$$\tilde{\mathbf{R}}_{res}^{(d)}(k) = \mathbf{R}_{res}^{(d)}(k) + \delta \mathbf{I}, \quad (4.18)$$

where $\delta > 0$ introduces a mild bias toward the identity covariance, analogous to white noise gain regularization in robust beamforming [5].

Toward Data-Driven Stopping

The known sparsity assumption is a limitation in real-world deployment where the number of active sources is unknown. Several principled alternatives could replace the fixed stopping criterion without modifying the remaining algorithmic steps. Residual energy criteria terminate when $\|\mathbf{R}_{res}^{(p)}(k)\|_F < \epsilon$, though calibrating ϵ to an unknown noise floor is non-trivial. Information-theoretic criteria such as MDL [60] and AIC select the model order by penalized likelihood over the eigenvalues of the sample covariance, and are well-established for source enumeration in array processing. Eigenvalue gap criteria offer a computationally simpler alternative based on successive eigenvalue ratios. A systematic evaluation of such data-driven stopping criteria is deferred to future work (Chapter 6).

4.7 Computational Complexity

This section analyzes the per-block computational cost of the proposed algorithm in terms of the number of complex multiply-accumulate operations, expressed using big- $\mathcal{O}$ notation in the parametric sense to characterize how cost scales with each parameter independently. The relevant parameters are: M the number of microphone channels, N the number of dictionary atoms, K the number of frequency bins, and d the number of sources, with the algorithm running for exactly d iterations under the known sparsity criterion, out of which N and K can most notably vary across experimental conditions or future extensions of the framework. In the experimental setting of this work, $M = 6$, $N = 48$, and $d = 2$, generally placing the analysis in the regime $N \gg M \geq d$.

CPSDM Estimation

Before the pursuit begins, the sample CPSDM must be estimated from a block of L time frames as per (2.9). For each frequency bin k , computing $\mathbf{x}(k, l)\mathbf{x}^H(k, l) \in \mathbb{C}^{M \times M}$ costs $\mathcal{O}(M^2)$ and accumulating over L frames costs $\mathcal{O}(LM^2)$. Over all K bins, the total cost of CPSDM estimation is

$$\mathcal{C}_{\text{CPSDM}} = \mathcal{O}(KLM^2). \quad (4.19)$$

This cost is incurred once per block and is independent of the pursuit iterations. It is shared by any algorithm that operates on the sample covariance matrix and is therefore not a cost unique to the proposed method.

Dictionary Normalization

Normalizing each of the N atoms across K bins requires computing $\|\mathbf{h}_n(k)\|_2$ for each (n, k) pair, at cost $\mathcal{O}(M)$ per pair, giving a total of $\mathcal{O}(KNM)$. Since the dictionary is fixed for a given HRTF dataset and microphone geometry, normalization is performed offline and does not contribute to the online per-block cost.

Per-Iteration Cost

Each iteration of the pursuit consists of three steps: atom selection, signal energy update, and residual update. We analyze each in turn.

Atom selection. The matching step computes the matrix product $\tilde{\mathbf{A}}^H(k)\mathbf{R}_{res}^{(p-1)}(k)\tilde{\mathbf{A}}(k) \in \mathbb{C}^{N \times N}$ at each bin k . This involves two matrix multiplications: first $\tilde{\mathbf{A}}^H(k)\mathbf{R}_{res}^{(p-1)}(k) \in \mathbb{C}^{N \times M}$ at cost $\mathcal{O}(NM^2)$, then right-multiplying by $\tilde{\mathbf{A}}(k) \in \mathbb{C}^{M \times N}$ at cost $\mathcal{O}(N^2M)$. Only the diagonal of the result is needed, so the full $N \times N$ product need not be formed explicitly; instead, the n^{th} diagonal element is obtained as $\tilde{\mathbf{h}}_n^H(k)\mathbf{R}_{res}^{(p-1)}(k)\tilde{\mathbf{h}}_n(k)$ where $\tilde{\mathbf{h}}_n(k)$ is the n^{th} column of $\tilde{\mathbf{A}}(k)$, at cost $\mathcal{O}(M^2)$ per atom. Computing all N diagonal entries therefore costs $\mathcal{O}(NM^2)$ per bin, and aggregating across K bins and taking the argmax costs $\mathcal{O}(KNM^2 + N)$. Dropping the lower-order term, the atom selection cost per iteration is

$$\mathcal{C}_{\text{sel}} = \mathcal{O}(KNM^2). \quad (4.20)$$

Signal energy update. The pseudoinverse $\tilde{\mathcal{A}}_{\Lambda}^{\dagger}(k) = \left(\tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k)\tilde{\mathcal{A}}_{\Lambda}(k)\right)^{-1}\tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k)$ involves a Gram matrix computation $\tilde{\mathcal{A}}_{\Lambda}^{\text{H}}(k)\tilde{\mathcal{A}}_{\Lambda}(k) \in \mathbb{C}^{p \times p}$ at cost $\mathcal{O}(pM^2)$. Note this is cheap since $p \leq d \ll M$. This is followed by inversion of a $p \times p$ matrix at cost $\mathcal{O}(p^3)$, and a matrix multiplication to form the pseudoinverse at cost $\mathcal{O}(p^2M)$. Applying the pseudoinverse on both sides of $\mathbf{R}_{\mathbf{x}}(k)$ as per (4.10) costs $\mathcal{O}(pM^2)$. Summing over K bins and noting that $p \leq d \ll M$ so that p -dependent terms are dominated by M -dependent ones, the signal energy update cost per iteration is

$$\mathcal{C}_{\text{sig}} = \mathcal{O}(KpM^2) \subseteq \mathcal{O}(KdM^2). \quad (4.21)$$

Residual update. The projection matrix $\mathbf{P}_{\Lambda}(k) = \tilde{\mathcal{A}}_{\Lambda}(k)\tilde{\mathcal{A}}_{\Lambda}^{\dagger}(k) \in \mathbb{C}^{M \times M}$ is formed from the already-computed pseudoinverse at cost $\mathcal{O}(pM^2)$. The two-sided projection $\mathbf{P}_{\Lambda}(k)\mathbf{R}_{\mathbf{x}}(k)\mathbf{P}_{\Lambda}^{\text{H}}(k)$ then costs $\mathcal{O}(M^3)$, and subtracting from $\mathbf{R}_{\mathbf{x}}(k)$ costs $\mathcal{O}(M^2)$. Over K bins, the residual update cost per iteration is

$$\mathcal{C}_{\text{res}} = \mathcal{O}(KM^3). \quad (4.22)$$

Total Cost

Summing the per-iteration costs over d iterations, the total online cost of the pursuit is

$$\begin{aligned} \mathcal{C}_{\text{total}} &= d \cdot (\mathcal{C}_{\text{sel}} + \mathcal{C}_{\text{sig}} + \mathcal{C}_{\text{res}}) \\ &= \mathcal{O}\left(Kd(NM^2 + dM^2 + M^3)\right) \\ &= \mathcal{O}\left(KdM^2(N + d + M)\right). \end{aligned} \quad (4.23)$$

For the experimental parameters of this work ($M = 6$, $N = 48$, $d = 2$), the term $N + d + M$ is numerically dominated by N , which reduces the total cost in practice to

$$\mathcal{C}_{\text{total}} \approx \mathcal{O}(KdNM^2). \quad (4.24)$$

This reveals that the dominant cost of the algorithm is the atom selection step, which scales linearly with the dictionary size N and the number of iterations d . The algorithm is thus computationally lightweight: the small values of M and d ensure that per-iteration costs remain modest even as K grows with the STFT resolution.

Dominant Cost and Scaling Behavior

Table 4.1 summarizes the cost of each step and its scaling with the key parameters.

Two observations follow from this analysis. First, the algorithm scales linearly with the number of frequency bins K , which is favorable since each bin is processed independently and the computation is trivially parallelizable across bins. Second, the quadratic scaling in M means that extending the framework to arrays with significantly more channels would increase the cost substantially, though for small microphone arrays $M \leq 8$ this remains entirely tractable. The linear scaling in N implies that enlarging the dictionary to a finer angular grid increases cost proportionally, providing a direct computational tradeoff against angular resolution that can be tuned to the requirements of a specific application.

Table 4.1: Per-iteration computational cost. Total cost accumulated over d iterations.

Step	Per-iteration cost	Dominant parameter
CPSDM estimation ¹	$\mathcal{O}(KLM^2)$	L, K
Atom selection	$\mathcal{O}(KNM^2)$	N
Signal energy update	$\mathcal{O}(KdM^2)$	d
Residual update	$\mathcal{O}(KM^3)$	M
Total (d iterations)	$\mathcal{O}(KdNM^2)$	N, d

¹ One-time cost for a time block.

Chapter Summary

This chapter presented a greedy iterative framework for joint estimation of active source directions, source energies, and noise covariance from binaural microphone observations. By extending S-OMP to the covariance domain and operating on the sample CPSDM rather than instantaneous signal coefficients, the framework exploits second-order spatial statistics to achieve robust atom selection across acoustic conditions. The algorithm was derived in full, covering the HRTF dictionary design, the three iterative steps (quadratic form atom selection, matrix least-squares energy estimation, and orthogonal residual update), and the positive semi-definiteness of the residual covariance throughout. The known sparsity stopping criterion was analyzed, consequences of model order mismatch characterized, and data-driven alternatives identified for future work. Complexity analysis established a dominant cost of $\mathcal{O}(KdNM^2)$, tractable for the binaural setting considered.

Experimental Evaluations

The preceding chapter established the theoretical foundations of the proposed joint estimation framework. This chapter evaluates its practical performance through controlled simulation experiments across two complementary metrics. HRTF estimation accuracy, measured by the Hermitian angle between estimated and true HRTFs, quantifies how well the algorithm recovers source spatial signatures. Binaural cue preservation, assessed via interaural level difference and time difference errors at the beamformer output, captures the end-to-end perceptual consequence of estimation quality.

The proposed algorithm (PROP) is evaluated against a sequential baseline (BASE) combining a GEVD-based HRTF estimator [52] with a VAD-based noise covariance estimator [46]. The performance advantages of PROP over BASE could therefore be attributable to either worse estimation via the methods used in BASE or to the joint estimation strategy employed by PROP.

Three conditions are varied systematically over 30 Monte Carlo trials: signal-to-noise ratio (SNR) (-10 to 30 dB), time block size (TBS) (10 to 100 STFT frames), and source azimuth separation (one source fixed at 0° , the second swept across the full azimuth range).

The experimental setup, including the acoustic scene geometry, signal generation procedure, noise model, STFT parameters, HRTF dictionary, and beamformer configuration, is described in full in Section 5.1. HRTF estimation results are presented in Section 5.2 and binaural cue preservation results in Section 5.3. A synthesis of findings across both experiments is provided in the Discussion (Section 5.4), followed by an assessment of the framework’s limitations in Section 5.5.

5.1 Experimental Setup

This section describes all simulation parameters and processing configurations used in the experiments reported in this chapter. All experiments share a common setup unless explicitly stated otherwise in the relevant experiment section.

Acoustic Scene and Signal Generation

The simulated scene comprises a behind-the-ear (BTE) hearing aid array of $M = 6$ microphones in total (3 on either side) and $d = 2$ simultaneously active point sources, as illustrated in Fig. 5.1. Each source signal is a 30 s sequence of ten concatenated 3 s speech snippets drawn from distinct speakers in the EARS dataset [43] at $f_s = 48$ kHz, ensuring speaker diversity across the trial and preventing any single speaker from dominating the spatial covariance. Sources are convolved with head-related impulse responses (HRIRs) from a Brüel & Kjær 4128C HATS dummy head (OIHeadHRTF dataset [11]) at the

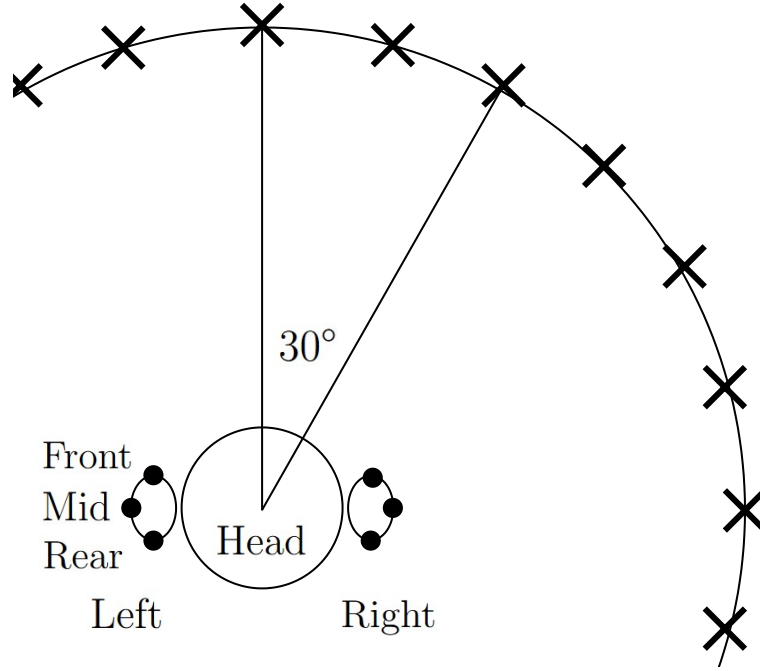


Figure 5.1: BTE microphone placement with two sources at 0° and 30° .

same f_s , yielding anechoic binaural microphone signals across all M channels. Both sources are placed at zero elevation, with azimuths specified per experiment.

Noise Model

Two independent noise components are simulated per channel: microphone self-noise and diffuse ambient noise, both modeled as zero-mean white Gaussian processes, independent across channels. The combined noise is therefore zero-mean Gaussian with a diagonal cross power spectral density matrix, consistent with the incoherent noise assumption of (2.7). It is worth acknowledging here that interpreting the diffuse noise in this way is a simplification; a physically realistic diffuse field would have spatially coherent noise whose CPSDM follows the sinc model.

SNR is defined as the ratio of total source power to total noise power at the reference microphone, swept from -10 dB to 30 dB in SNR experiments and fixed at 30 dB otherwise. Low-SNR conditions reflect a combination of self-noise and elevated diffuse ambient noise; high-SNR conditions are dominated by self-noise alone.

STFT and Block Processing Parameters

Signals are processed in the STFT domain using a 32 ms Hann window with 50% overlap (16 ms frame shift) at $f_s = 48$ kHz, with an FFT length of 2048 points yielding 1025 unique frequency bins. The sample CPSDM is estimated over non-overlapping blocks of L STFT frames per (2.9), with adjacent blocks overlapping by 50% for temporal smoothing. In SNR and azimuth sweep experiments, L is fixed at 50 frames (≈ 800 ms);

in TBS sweep experiments, L is varied from 10 to 100 frames (≈ 160 ms – 1.6 s). Both sources are placed on-grid with respect to the HRTF dictionary, isolating estimation performance from off-grid mismatch error, which is deferred to future work.

HRTF Dictionary

The dictionary $\mathcal{A}(k) \in \mathbb{C}^{M \times N}$ is constructed from the same OlHeadHRTF dataset used for signal generation [11], covering the full 360° azimuthal range at 7.5° spacing and zero elevation, yielding $N = 48$ atoms. Each atom $\mathbf{h}_n(k) \in \mathbb{C}^M$ contains the complex HRTF coefficients at bin k for the n^{th} grid directions, normalized per (2.16).

The Baseline

The baseline (BASE) combines two established methods applied sequentially to the same observed microphone signals, with resulting estimates passed to the same beamformer as PROP for a fair comparison.

GEVD-based HRTF estimation [52] decomposes the observed CPSDM via generalized eigenvalue decomposition, extracting dominant generalized eigenvectors as HRTF estimates. It requires a separate noise covariance estimate as input, provided here by the VAD-based estimator described below. It represents a state-of-the-art approach to sequential noise-then-HRTF estimation.

VAD-based noise covariance estimation [46] accumulates the sample covariance during VAD-identified noise-only frames — a requirement absent in PROP — supplying the resulting estimate to both the GEVD estimator and the beamformer directly. Performance is inherently limited by VAD accuracy and noise stationarity.

Together, BASE instantiates the conventional sequential pipeline. Any performance advantage of PROP can be attributable to either suboptimal estimation of parameters using BASE, or to the joint coupling of source selection, energy estimation, and noise covariance recovery within a single iterative procedure.

Evaluation Metrics

Three metrics assess performance: the Hermitian angle (HA) for HRTF estimation accuracy, and interaural level and time difference (ILD, ITD) errors for binaural cue preservation.

Hermitian angle (HA). The HA measures the angular distance between the estimated and true HRTF vectors [58],

$$\text{HA}_i = \frac{1}{T(K/2 + 1)} \sum_{t=1}^T \sum_{k=1}^{K/2+1} \arccos \left(\frac{|\mathbf{a}_i^H(t, k) \mathbf{h}_i(t, k)|}{\|\mathbf{a}_i(t, k)\| \|\mathbf{h}_i(t, k)\|} \right), \quad (5.1)$$

measured over even K and where T is the total number of time blocks, $\mathbf{a}_i$ is the true ATF/HRTF and $\mathbf{h}_i$ is the ATF/HRTF estimated using either method. It takes values in $[0, \pi/2]$, with 0 indicating perfect alignment. Scale-invariance makes it appropriate for normalized dictionary atoms.

Interaural transfer function (ITF). The ITF is the ratio of left and right ear HRTFs, encoding the binaural localization cues. Input and output ITFs for source i are [9]

$$\text{ITF}^{\text{in}}(t, k) = \frac{a_{iL}(t, k)}{a_{iR}(t, k)}, \quad \text{ITF}^{\text{out}}(t, k) = \frac{\hat{\mathbf{w}}_L^H(t, k)\mathbf{h}_i(t, k)}{\hat{\mathbf{w}}_R^H(t, k)\mathbf{h}_i(t, k)}. \quad (5.2)$$

ILD error. The ILD error is the mean absolute difference between output and reference ILD magnitudes, computed in the logarithmic domain, over 3–8 kHz [27], defined as

$$\text{ILD}^{\text{err}} = \frac{1}{(K/2 + 2 - k_{\text{ILD}})T} \sum_{k=k_{\text{ILD}}}^{K/2+1} \sum_t^T |20 \log_{10} (|\text{ITF}^{\text{out}}(t, k)|) - 20 \log_{10} (|\text{ITF}^{\text{in}}(t, k)|)|, \quad (5.3)$$

defined for even K and where k_{ILD} is the first frequency bin (lower bound) of the 3–8 kHz range (center frequency = 3 kHz).

ITD error. The ITD error is the mean absolute time difference between output and reference ITFs over 0–1.5 kHz [27], defined as

$$\text{IPD}^{\text{err}} = \frac{1}{k_{\text{IPD}}T} \sum_{k=1}^{k_{\text{IPD}}} \sum_t^T \frac{|\angle \text{ITF}^{\text{out}}(t, k) - \angle \text{ITF}^{\text{in}}(t, k)|}{2\pi f_k}, \quad (5.4)$$

where f_k is the center frequency of bin k , and k_{IPD} is the last frequency bin (upper bound) of the 0–1.5 kHz range (center frequency = 1.5 kHz). The resulting metric is in seconds and is consistent across frequency, as a given time delay produces the same ITD value regardless of f_k .

Both binaural cue metrics are computed separately for target S_1 and interferer S_2 , reflecting the asymmetric requirements of hearing aid-based beamforming.

Beamformer Configuration

The estimated HRTFs and noise covariance from each method are passed to a Joint Binaural Linearly Constrained Minimum Variance (JBLCMV) beamformer [26], which enhances the target while suppressing the interferer and preserving binaural cues. The beamformer requires steering vectors, source covariances, and a noise covariance matrix. Source covariances can be reconstructed by the $\sigma^2 \mathbf{h}\mathbf{h}^H$ form in both algorithms; moreover, the interferer covariance can be added to the noise covariance to give the “interferer plus noise” covariance (referred to as $\hat{\mathbf{R}}_{\mathbf{n}}(k)$ for simplicity). Therefore, following (2.26–2.30), we can replace $\mathbf{R}_{\mathbf{n}}(k)$ with $\hat{\mathbf{R}}_{\mathbf{n}}(k)$ and get the JBLCMV beamformer solution

$$\mathbf{w}_{\text{JBLCMV}}(k) = \tilde{\mathbf{R}}_{\mathbf{n}}^{-1}(k) \mathbf{C}(k) \left(\mathbf{C}^H(k) \tilde{\mathbf{R}}_{\mathbf{n}}^{-1}(k) \mathbf{C}(k) \right)^{-1} \mathbf{f}(k), \quad (5.5)$$

$$\tilde{\mathbf{R}}_{\mathbf{n}}(k) = \begin{bmatrix} \hat{\mathbf{R}}_{\mathbf{n}}(k) & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_{\mathbf{n}}(k) \end{bmatrix}, \quad (5.6)$$

$$\mathbf{C}(k) = \left[\begin{array}{cc|cc} \mathbf{h}_1(k) & \mathbf{0} & \mathbf{h}_2(k)h_{2,R}(k) & \\ \mathbf{0} & \mathbf{h}_1(k) & -\mathbf{h}_2(k)h_{2,L}(k) & \end{array} \right], \quad \mathbf{f}^T(k) = [h_{1,L}^* \quad h_{1,R}^* \mid 0] \quad (5.7)$$

for the analysis.

For BASE, steering vectors and source energies are derived from the GEVD, and the noise covariance from the VAD estimator. $\hat{\mathbf{R}}_{\mathbf{n}}(k)$ can similarly be calculated by

$$\hat{\mathbf{R}}_{\mathbf{n}}(k) = \mathbf{R}_{\mathbf{n},\text{BASE}}(k) + \lambda_2 \mathbf{h}_{2,\text{BASE}}(k) \mathbf{h}_{2,\text{BASE}}^H(k), \quad (5.8)$$

where λ_2 is the generalized eigenvalue associated with $\mathbf{S}_2$ [52].

In PROP, the quantities are supplied directly by $\mathcal{A}_{\mathbf{\Lambda}}(k)$, the diagonal of $\hat{\mathbf{R}}_{\mathbf{s}}(k)$, and $\mathbf{R}_{res}(k)$ respectively. In particular, we have

$$\hat{\mathbf{R}}_{\mathbf{n}}(k) = \mathbf{R}_{res}(k) + \frac{\left[\hat{\mathbf{R}}_{\mathbf{s}}(k)\right]_{22}}{\|\mathbf{h}_{2,\text{PROP}}(k)\|_2^2} \mathbf{h}_{2,\text{PROP}}(k) \mathbf{h}_{2,\text{PROP}}^H(k), \quad (5.9)$$

where $\mathbf{h}_{2,\text{PROP}}(k) = [\mathcal{A}_{\mathbf{\Lambda}}(k)]_2$, column of the HRTF dictionary $\mathcal{A}(k)$ supported by $\mathbf{\Lambda}$ connected to $\mathbf{S}_2$.

The beamformer configuration is identical across all methods, ensuring that metric differences are attributable solely to differences in the quality of the estimated inputs and not to differences in the spatial filter design or the beamformer parameterization.

5.2 Experiment 1: HRTF Estimation Accuracy

This experiment evaluates the ability of the proposed algorithm to recover the HRTFs of the two active sources from the observed binaural mixture, as measured by the Hermitian angle defined in (5.1). Three conditions are examined sequentially: the effect of SNR on estimation accuracy, the effect of time block size on covariance quality and its downstream impact on HRTF recovery, and the geometric sensitivity of each method to source azimuth configuration, each averaged over 30 Monte Carlo trials. Unless otherwise stated, sources $\mathbf{S}_1$ and $\mathbf{S}_2$ are fixed at 0° and 30° respectively, SNR is fixed at 30 dB, and the time block size is fixed at $L = 50$ frames.

5.2.1 Effect of SNR

Fig. 5.2a reports the Hermitian angle for both sources as a function of SNR, ranging from -10 dB to 30 dB, for both PROP and BASE.

Overall Trend

PROP achieves consistently lower Hermitian angle than BASE across the full SNR range for both sources, demonstrating superior HRTF recovery under all tested noise conditions. The performance gap is most pronounced at low SNR: at -10 dB, PROP's Hermitian angle for $\mathbf{S}_2$ remains below BASE's high-SNR floor. This robustness follows from block-based CPSDM estimation, which averages noise fluctuations over $L = 50$ frames, and from frequency-aggregated atom selection per (4.6), which requires consistent energy across all K bins rather than a strong response in any single bin.

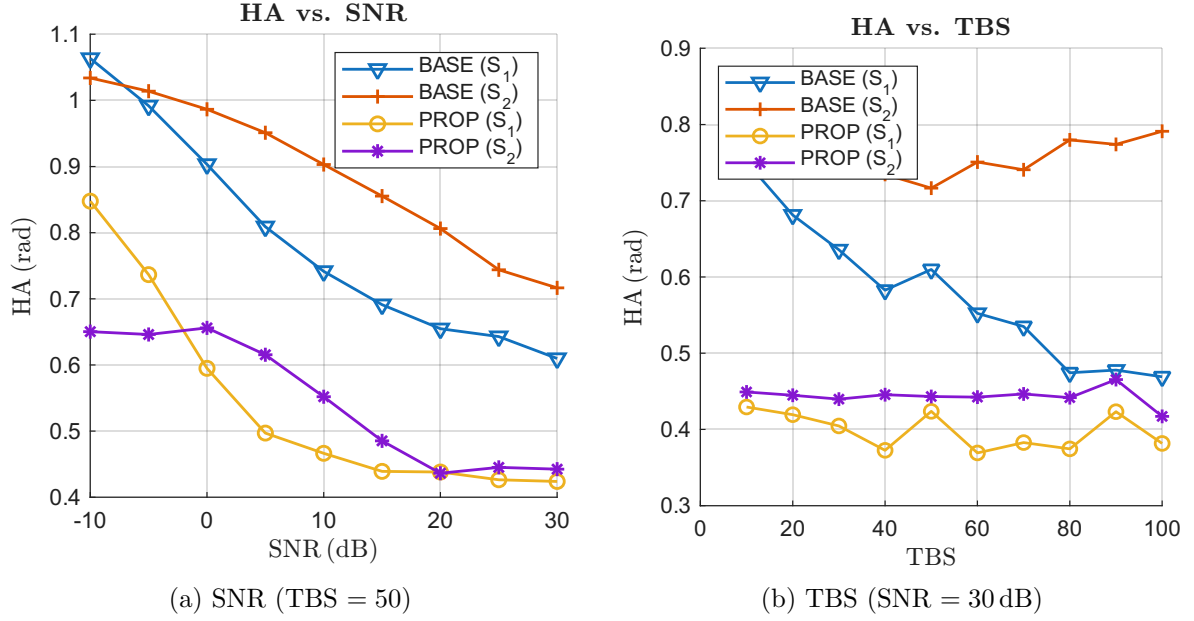


Figure 5.2: HRTF estimation error under varying SNR and time block size.

Source Asymmetry

BASE exhibits a persistent asymmetry between S_1 and S_2 , with the latter at 30° showing a substantially higher Hermitian angle across all SNR levels. This reflects a fundamental limitation of the GEVD: when one source occupies the frontal direction, its HRTF may dominate the signal subspace, leaving the off-frontal source poorly represented in the dominant eigenvectors. PROP maintains comparably low Hermitian angle for both sources, a consequence of iterative deflation: once S_1 is selected and projected out of the residual covariance, the pursuit recovers S_2 from a residual it dominates, yielding equitable accuracy regardless of frontal dominance.

SNR Saturation

Both methods exhibit a Hermitian angle floor beyond approximately 20 dB SNR. For PROP, this reflects an irreducible approximation error intrinsic to the estimator: at high SNR, noise contamination becomes negligible and the residual Hermitian angle could be attributed to a structural source, e.g., signal leakage or numerical precision. BASE's floor is higher, attributable to a combination of finite sample effects and structural GEVD limitations in the presence of two spatially close sources. This motivates the block size analysis in Section 5.2.2.

5.2.2 Effect of Time Block Size

Fig. 5.2b reports the Hermitian angle as a function of time block size L , varied from 10 to 100 frames (approximately 160 ms to 1.6 s), at a fixed SNR of 30 dB.

BASE: Data-Limited vs. Model-Limited Error

BASE exhibits qualitatively different block size dependence across the two sources. Hermitian angle for S_1 decreases meaningfully with increasing L , confirming data-limited error: more frames improve the CPSDM estimate and GEVD separation. For S_2 , increasing L yields no meaningful improvement, indicating model-limited error: the GEVD reaches an accuracy ceiling determined by its structural properties, not by data quality. This distinction implies that improving BASE’s off-frontal source recovery requires a fundamentally different estimation approach, not longer observation windows.

PROP: Robustness to Block Size

PROP maintains consistently low Hermitian angle across all tested block sizes, indicating that reliable atom selection requires relatively few frames. Frequency-aggregated matching pools information across K bins at each iteration, compensating for reduced statistical reliability at short blocks by exploiting spectral redundancy. This is practically significant in non-stationary scenes where source configurations change on a timescale of hundreds of milliseconds, as temporal adaptivity requires small L , a regime where PROP remains accurate.

Latency Implications

The contrast between PROP’s block-size robustness and BASE’s data-dependence has direct implications for system latency. A block of $L = 10$ frames at a 16 ms frame shift corresponds to 160 ms of processing delay, which is near the upper bound of what is perceptually acceptable in hearing aid applications [19]. PROP achieves near-optimal performance at this block size, whereas BASE requires substantially larger blocks to approach its best-case accuracy for S_1 . The proposed framework therefore offers a more favorable latency-accuracy tradeoff for real-time binaural processing.

5.2.3 Effect of Source Azimuth Configuration

Fig. 5.3 reports the Hermitian angle as a function of source azimuth, with S_1 fixed at 0° and S_2 swept across the full azimuth range from -180° to 180° at 7.5° spacing, at fixed SNR of 30 dB and $L = 50$ frames. Results are displayed as polar plots, with the radial axis representing Hermitian angle and the angular axis representing the azimuth of S_2 .

BASE: Uniform High Error

The polar plot for BASE reveals near-uniform, high Hermitian angle contours for both sources across almost all azimuth configurations of S_2 , which indicates that BASE’s estimation quality is largely insensitive to the spatial arrangement of the sources, maintaining a consistently poor performance floor regardless of the angular separation between S_1 and S_2 . The absence of a meaningful dependence on source configuration further

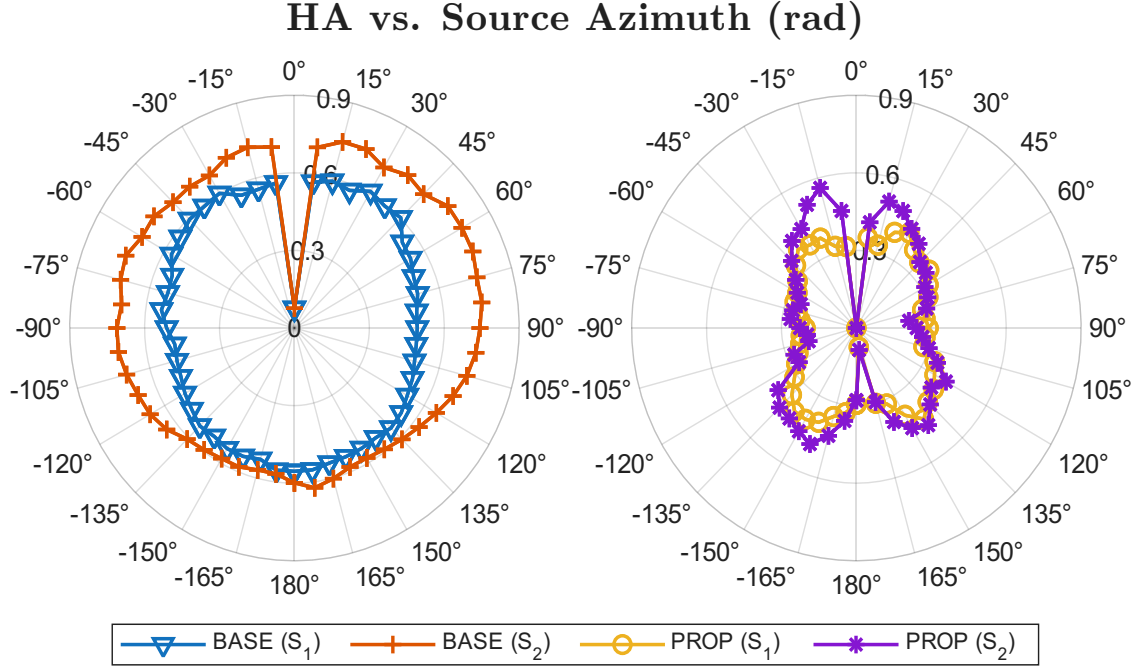


Figure 5.3: HRTF estimation error vs. source azimuth (S_1 fixed at 0° , $\text{SNR} = 30$ dB, $\text{TBS} = 50$).

supports the model-limited interpretation from Section 5.2.2: BASE’s errors are dominated by the structural limitations of the GEVD estimator rather than by the geometry of the scene.

PROP: Compact Contour with Geometric Sensitivity

PROP yields a compact, low Hermitian angle contour for S_1 across most azimuth configurations, consistent with iterative deflation: S_1 is selected first due to its frontal energy dominance, projected out, and S_2 recovered from the residual. The S_2 contour is more irregular, with elevated error near frontal and rear azimuths, due to two compounding geometric effects. First, S_2 near 0° or 180° increases mutual coherence $\mu(\mathcal{A}(k))$ between selected atoms, degrading the pseudoinverse conditioning and leaving residual S_1 energy that contaminates subsequent atom selection. Second, the 7.5° dictionary resolution concentrates similar atoms near frontal and rear directions, increasing the risk of selecting an incorrect neighbor when the true atom holds only a marginal energy advantage.

Angular Resolution and the Coherence Limit

The azimuth results reveal a practical resolution limit analogous to the Rayleigh criterion: for the BTE geometry and $N = 48$ atom dictionary, PROP achieves reliably low Hermitian angle for S_2 when angular separation from S_1 exceeds approximately 30° . Below this threshold, dictionary coherence becomes the limiting factor. Finer resolution

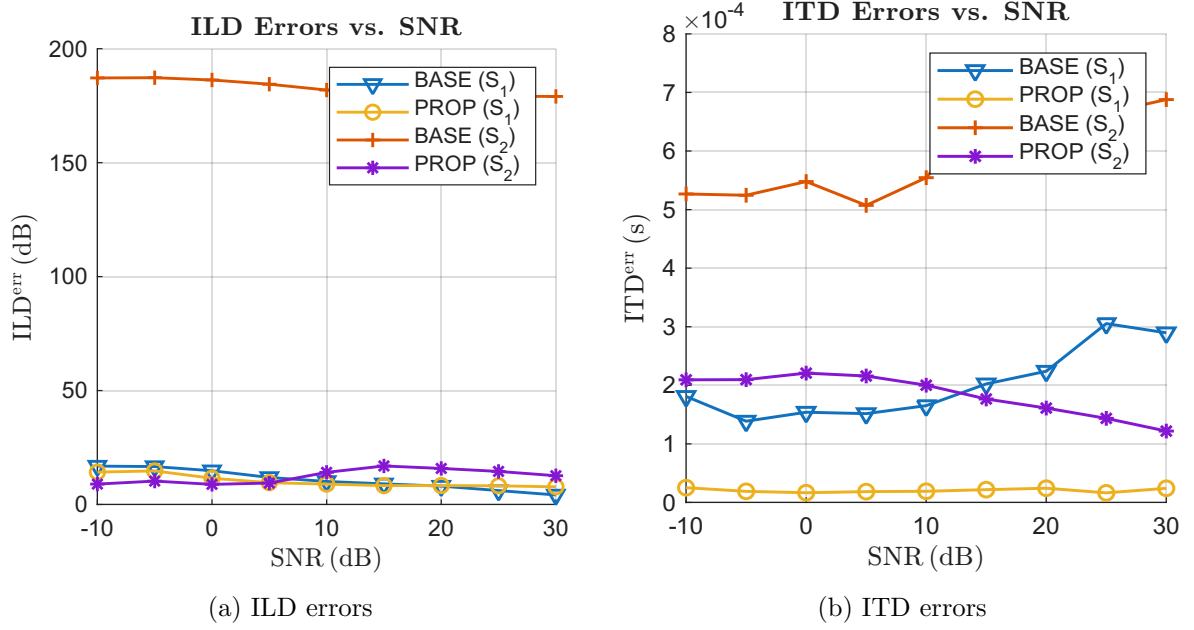


Figure 5.4: Binaural cue preservation errors for S_1 and S_2 under varying SNR.

could be achieved by increasing N or applying coherence-reducing dictionary pruning prior to the pursuit; both are identified as directions for future work in Chapter 6.

5.3 Experiment 2: Binaural Cue Preservation

This experiment evaluates the end-to-end perceptual consequence of the HRTF and noise covariance estimates produced by each method, as measured by the binaural cue preservation metrics defined in (5.3) and (5.4). Where Experiment 1 assessed the intermediate estimation outputs directly, this experiment assesses whether estimation quality translates into preserved spatial hearing cues at the beamformer output, one step closer to the perceived qualities ultimately experienced by the listener. Sources S_1 at 0° and S_2 at 30° serve as the target and interferer respectively, and SNR is varied from -10 dB to 30 dB with $L = 50$ frames fixed throughout. ILD and ITD errors are reported separately for target and interferer, reflecting the asymmetric perceptual requirements of a hearing aid system.

5.3.1 ILD Preservation

Fig. 5.4a reports the ILD error for the target S_1 and interferer S_2 as a function of SNR for both PROP and BASE.

Overall Trend

Across both sources and both algorithms, ILD error decreases monotonically as SNR increases from -10 dB to 30 dB. This is expected behavior: at higher SNRs the tar-

get signal dominates, making acoustic parameter estimation more reliable and thus reducing beamforming errors.

Target ILD

At low SNR (-10 dB), both BASE and PROP show their highest errors for the target S_1 , with BASE sitting slightly above PROP. As SNR increases, BASE converges steeply to near-zero error relative to PROP. Along the SNR range, however both algorithms are consistent with low ILD errors for the target. This suggests that, at least for S_1 , both methods produce reliable parameters to preserve high-frequency cues.

Interferer ILD

For the interferer S_2 , the curves tell a qualitatively similar but quantitatively distinct story. BASE starts at substantially elevated error at low SNR and decreases more slowly than either S_1 curve; this may be due to the presence of large outlier errors amplified by the logarithmic transformation. PROP again outperforms its counterpart, showing a faster decline with increasing SNR. However, compared to the S_1 curve, the S_2 curve remains at a higher level across most of the SNR range, particularly in the mid-SNR region (0 – 15 dB). This asymmetry is intuitive: beamformers designed to enhance target S_1 may inadvertently suppress or distort the ILD cues of interferer S_2 as a side effect.

5.3.2 ITD Preservation

Fig. 5.4b reports the ITD error for the target and interferer as a function of SNR for both methods.

Target ITD

BASE for the target S_1 maintains relatively low ITD errors across low to moderate SNRs (0 – 15 dB), but then degrades noticeably at higher SNRs before slightly recovering at 30 dB; it may suggest that at very high signal levels the BASE algorithm begins overfitting to direct-path energy or encounters a model mismatch that inflates its ITD estimates for the target. On the other hand, PROP is strikingly superior throughout the entire SNR range, with errors remaining essentially flat and near-zero. It indicates that PROP is highly robust to noise conditions for the target source and does not suffer from the SNR-dependent degradation seen in BASE.

From a perceptual standpoint, target ITD preservation is the more critical binaural cue metric as it operates in the 0 – 1.5 kHz frequency range, the dominant cue for azimuthal localization and spatial release from masking [62]. PROP’s near-zero target ITD error across all SNR levels therefore represents its most practically significant result.

Interferer ITD

BASE for interferer S_2 shows consistently elevated errors throughout with an increasing deterioration as SNR increases. This is the highest absolute error of any curve on the

plot, suggesting that BASE struggles significantly with the interferer’s ITD estimation. PROP performs markedly better than BASE across low to moderate SNRs, with a downward trend at higher SNRs. This contrast reflects the JBLCMV’s inherent constraint tradeoff: tighter spatial constraints on the target channel induce greater phase distortion on the interferer, particularly when SNR is insufficient to fully decouple the two sources in the covariance domain. This is perceptually acceptable in the hearing aid context, where accurate interferer localization is not required; phase distortion on the suppressed channel is an expected consequence of optimizing for target preservation.

5.3.3 Binaural Cue Asymmetry

Taken together, the ILD and ITD results demonstrate that PROP consistently outperforms BASE in binaural cue preservation across the entire SNR range, with the largest gains occurring at low to moderate SNR where BASE is most susceptible to noise-contaminated parameter estimates. ILD errors decrease monotonically with SNR for both methods, reflecting improved acoustic parameter estimation, with BASE and PROP both having consistently low target ILD errors; meanwhile, interferer ILD errors from PROP are substantially lower than BASE. The most significant result, however, is PROP’s near-zero and stable target ITD error across all SNR conditions, in contrast to the elevated and less stable ITD errors observed for BASE, particularly for the interferer. Given the perceptual importance of ITD for low-frequency localization and spatial release from masking [62], this robust target ITD preservation represents a key practical advantage of the proposed method. Overall, the results indicate that the joint estimation framework provides higher-quality beamformer inputs and more reliable preservation of both level- and time-based spatial cues than the sequential baseline.

5.4 Discussion

The two experiments presented in this chapter evaluate the proposed framework from complementary points of view: Experiment 1 assessed the quality of the intermediate estimation outputs directly, while Experiment 2 assessed whether that quality propagates into preserved binaural cues at the beamformer output. This section synthesizes the findings across both experiments, identifies the consistent themes that emerge, and interprets them in the context of the framework’s design.

Joint Estimation Confers Consistent Estimation Advantage

Across all conditions in Experiment 1, PROP achieves lower Hermitian angle than BASE for both sources across the full SNR range, all block sizes, and the majority of azimuth configurations. This consistent advantage is attributable to the joint nature of the estimation: source selection, energy estimation, and noise covariance recovery are coupled within a single iterative procedure, ensuring internal consistency between the source model and the noise covariance used for residual deflation. BASE estimates noise independently of source activity, allowing VAD inaccuracies to propagate into

the GEVD decomposition and degrade HRTF recovery. The breadth of conditions over which PROP prevails suggests this coupling is the primary driver of the estimation advantage.

Noise Robustness Through Covariance Domain Processing

At -10 dB, where instantaneous observations are dominated by noise, PROP’s Hermitian angle for S_2 remains below BASE’s high-SNR floor. This reflects two-level noise averaging inherent to the framework: temporal averaging over L frames in the CPSDM estimation step, and spectral averaging over K bins in the frequency-aggregated atom selection step. Together, these suppress noise contributions before any selection decision is made — a robustness that instantaneous signal processing cannot achieve — which is practically significant in hearing aid applications where acoustic conditions with SNR below 0 dB are common.

Iterative Deflation Produces Equitable Source Recovery

BASE exhibits persistent asymmetry between S_1 and S_2 , with the latter being off-frontal sustaining higher Hermitian angle regardless of SNR or block size. PROP largely eliminates this asymmetry via iterative deflation: projecting out the first selected source’s contribution ensures the residual covariance is dominated by the remaining source, providing an unobstructed basis for the second atom selection. This equitable recovery is a structural advantage of the greedy pursuit formulation that one-shot methods such as the GEVD cannot readily replicate.

Block Size Robustness Favors Real-Time Operation

The TBS experiment revealed that PROP achieves near-optimal estimation accuracy at block sizes as short as $L = 10$ frames, corresponding to approximately 160 ms of audio, while BASE requires substantially larger blocks to approach its best-case accuracy for the frontal source and remains model-limited for the off-frontal source regardless of block size. This asymmetry has direct implications for system design: a hearing aid processing pipeline must balance estimation accuracy against processing latency, and PROP’s robustness to short block sizes means it can operate near the perceptual latency boundary without sacrificing estimation quality. BASE’s data dependence, by contrast, forces a tradeoff between latency and frontal source accuracy that cannot be resolved within the GEVD framework.

Geometric Sensitivity as a Dictionary Coherence Effect

The source azimuth configuration experiment revealed that PROP’s estimation accuracy degrades when the two sources are placed near collinear or antipodal directions relative to the array. This geometric sensitivity was attributed in Section 5.2.3 to the mutual coherence of the HRTF dictionary: when two selected atoms are nearly collinear, the pseudoinverse becomes ill-conditioned and the deflation step leaves residual source

energy that contaminates subsequent atom selections. This is a known and principled limitation of greedy sparse recovery methods, analogous to the resolution limit of MUSIC-type spectral estimators, and it sets a practical lower bound on the resolvable angular separation between sources for the given array geometry and dictionary resolution. Importantly, this limitation is well-characterized and addressable: finer dictionary resolution, coherence-aware atom selection, or regularized pseudoinverse computation are all viable strategies for extending the framework’s resolution.

Estimation Quality Partially Propagates to Binaural Cues

Experiment 2 revealed that PROP’s superior HRTF estimation quality translates into consistent advantages in both binaural cue metrics across the SNR range. For ILD preservation, PROP is consistently close to near-zero interferer level difference error in comparison to BASE, confirming that the dictionary-constrained HRTF estimates produce more accurate level-based beamformer constraints. For ITD preservation, PROP achieves near-zero and stable target ITD errors throughout, while BASE degrades unexpectedly at high SNR, suggesting a model mismatch effect in the sequential estimation pipeline that the joint framework avoids. Together, these results confirm that the joint estimation strategy produces beamformer inputs of consistently higher quality across both the level-based and time-based spatial cues that govern the listener’s spatial auditory experience.

Internal Consistency of Joint Estimates Benefits Beamforming

A recurring theme across both experiments is that the internal consistency of jointly produced estimates confers advantages beyond the accuracy of any individual estimate. PROP’s stable target ITD across SNR exemplifies this: the beamformer receives mutually compatible steering vectors and noise covariance, yielding coherent spatial filter constraints that independently optimized estimators cannot replicate regardless of their individual accuracy. This structural property of the joint framework represents perhaps its strongest argument, with implications extending beyond the specific beamformer and metrics used here.

5.5 Limitations

The experiments presented in this chapter were conducted under a set of controlled assumptions that, while standard in the sparse recovery and binaural processing literature, do not fully reflect the complexity of real-world acoustic deployment. This section critically examines the four most consequential assumptions underlying the current framework, characterizes the nature and expected severity of the performance degradation each introduces, and identifies the specific extensions needed to address them.

On-Grid Source Assumption

All experiments place sources at azimuths coinciding exactly with dictionary entries, ensuring the true HRTF vectors are recoverable in the noiseless case. In practice, acoustic sources are continuously distributed, and an off-grid source at azimuth ϕ produces an HRTF vector lying between the adjacent atoms $\mathbf{h}(\phi_n, k)$ and $\mathbf{h}(\phi_{n+1}, k)$. The matching step in (4.6) will select the nearest on-grid atom, introducing a directional bias of up to half the grid spacing (3.75° for the 7.5° dictionary) that is irreducible by any improvement in SNR, block size, or algorithmic refinement. This mismatch propagates to the JBLCMV as an incorrect steering vector, potentially degrading both source enhancement and binaural cue preservation.

The severity of the bias depends on dictionary density and local HRTF sensitivity to azimuth perturbation, with sharper spectral variation near the ipsilateral side producing larger atom mismatch. Addressing this requires either finer dictionary resolution at the cost of higher complexity and increased coherence, or an off-grid recovery strategy such as continuous basis pursuit [14] or atomic norm minimization [51]. Incorporating such approaches is identified as the highest-priority extension of the current framework.

Free-Field HRTF Assumption

The dictionary consists of anechoic free-field measurements [11], whereas real listening environments introduce room reflections such that the effective acoustic transfer function becomes

$$\mathbf{h}_{room}(\phi, k) = \mathbf{h}_{ff}(\phi, k) \odot \mathbf{h}_{RTF}(k), \quad (5.10)$$

where $\mathbf{h}_{RTF}(k)$ is the room transfer function across the M channels. Since $\mathbf{h}_{room}(\phi, k) \neq \mathbf{h}_{ff}(\phi, k)$ in general, the atom selection step recovers the best free-field approximation rather than the true source direction. Degradation severity depends on reverberation time [36]: modest for reverberation time $T_{60} < 200$ ms where the direct path dominates, potentially severe for $T_{60} > 500$ ms where reflections substantially modify the spatial signature and may confound the frequency-aggregated matching step.

Two mitigation strategies exist: constructing a room-adapted dictionary from room-specific ATF measurements, or applying dereverberation preprocessing before CPSDM estimation. Both are well-studied in the array processing literature [36] and represent viable extensions to reverberant conditions.

Dictionary Coherence and Angular Resolution

The $N = 48$ atom dictionary at 7.5° spacing exhibits increasing mutual coherence as source separation decreases, with the pseudoinverse becoming ill-conditioned when selected atoms approach collinearity. As evidenced by the azimuth configuration results of Section 5.2.3, this sets a practical resolution limit of approximately 30° for the BTE array, which happens to be a structural property of greedy pursuit on a coherent dictionary that cannot be overcome by improving SNR or block size. The BTE geometry's small inter-microphone spacing limits spatial selectivity, meaning adjacent dictionary atoms are highly coherent and the pursuit struggles to discriminate between them.

Two mitigation approaches exist. First, increasing dictionary spacing reduces coherence at the cost of angular resolution; the optimal spacing can be determined from the maximum allowable mutual coherence μ_{max} under greedy recovery guarantees [55]. Second, coherence-aware dictionary pruning, i.e., removing atoms exceeding a coherence threshold with any already-selected atom, can be incorporated into the pursuit loop without modifying the core algorithm. Both are identified as practical directions for future work.

Known Source Sparsity Assumption

The number of active sources d is assumed known and used as the pursuit stopping criterion, which is standard in controlled evaluations but unrealistic in practice, where source count is unknown and time-varying. As analyzed in Section 4.6, underestimating d leaves residual source energy in the noise covariance, while overestimating d absorbs noise into the signal subspace; in deployment, neither the true value nor the direction of error is known a priori. Principled alternatives exist: information-theoretic criteria such as MDL [60] and AIC operate on sample covariance eigenvalues and are well-established for source enumeration, while residual energy thresholds and eigenvalue gap criteria offer simpler alternatives. Evaluating these within the proposed framework is a high-priority direction toward a fully blind binaural processing system requiring no prior scene knowledge.

Chapter Summary

This chapter evaluated the proposed framework through two controlled simulation experiments assessing HRTF estimation accuracy and binaural cue preservation against a sequential GEVD and VAD-based baseline across SNR levels, block sizes, and azimuth configurations, using a $M = 6$ channel BTE array, $N = 48$ atom dictionary at 7.5° spacing, and 30 Monte Carlo trials per condition.

Experiment 1 demonstrated consistently lower Hermitian angle than the baseline across all conditions, with the advantage most pronounced at low SNR where covariance domain processing and frequency-aggregated atom selection provide robustness unavailable to coefficient-domain methods. Iterative deflation produced equitable recovery of both sources regardless of frontal dominance, eliminating the systematic asymmetry observed for the baseline. The block size analysis confirmed near-optimal accuracy at short block lengths, validating frequency aggregation as an effective substitute for temporal averaging under low-latency constraints. Source azimuth configuration results exposed geometric sensitivity near collinear and antipodal source directions, consistent with the mutual coherence analysis of Section 4.4.

Experiment 2 revealed nuanced propagation of estimation quality into binaural cues. Target ITD was well-preserved across the full SNR range, with internal consistency between jointly produced estimates identified as the key enabler of beamformer stability. Target ILD preservation was comparable for both methods but interferer ILD highly benefited from the proposed method’s dictionary-based iterative approach. The four

limitations analyzed in Section 5.5 collectively define the boundary conditions of the current framework and the extensions needed for real-world deployment.

Conclusion

This thesis presented a covariance-domain sparse pursuit framework for joint estimation of acoustic transfer functions, source signal energies, and noise covariance from binaural microphone observations. The framework was motivated by the limitations of existing approaches that treat these estimation problems sequentially, exposing them to error propagation and requiring separate noise reference intervals or prior knowledge of the acoustic scene. By reformulating the source selection problem as a greedy pursuit over a predetermined HRTF dictionary operating on the cross-power spectral density matrix, the proposed algorithm couples all three estimation problems within a single iterative procedure, producing internally consistent outputs that serve as direct inputs to a binaural beamformer without any additional preprocessing.

6.1 Summary of Contributions

Contribution 1: Covariance-domain extension of S-OMP. A novel extension of Simultaneous Orthogonal Matching Pursuit to the covariance domain was derived, replacing the standard inner product matching step with a quadratic form evaluated against the residual CPSDM. This formulation inherits the joint sparsity and iterative deflation properties of S-OMP while operating on second-order spatial statistics, providing noise robustness through temporal averaging unavailable to coefficient-domain pursuit. Its structural connection to the MUSIC pseudospectrum scan was established, situating the approach within the broader array processing literature.

Contribution 2: Joint estimation of HRTF, source energy, and noise covariance. The framework simultaneously recovers active source directions, source signal energies, and a residual noise covariance in a single iterative pass over the observed CPSDM. The closed-form energy update via the Moore-Penrose pseudoinverse and the projection-based residual update were derived in full, with positive semi-definiteness of the residual verified by construction, eliminating the need for separate VAD-gated noise estimation or sequential HRTF estimation pipelines.

Contribution 3: Frequency-aggregated atom selection for broadband consistency. Aggregating quadratic form scores across all frequency bins before maximizing promotes source selections consistent across the full broadband spectrum. This was shown to provide noise robustness at short block sizes, enabling near-optimal HRTF estimation at block lengths as short as 160 ms, near the perceptual latency boundary for hearing aid applications.

Contribution 4: Empirical validation against sequential baselines. The framework was evaluated against a GEVD-based HRTF estimator and VAD-based noise covariance estimator across SNR levels, block sizes, and azimuth configurations. The proposed algorithm achieved consistently lower Hermitian angle than the baseline across

all conditions, with advantages most pronounced at low SNR and short block sizes. When driving a JBLCMV beamformer, the framework produced near-zero and stable target ITD errors as well as interferer ILD errors across the full SNR range, confirming that dictionary-constrained directional accuracy is sufficient to support phase-based binaural cue preservation.

6.2 Conclusions

The experimental results support three overarching conclusions.

First, joint estimation of source directions, energies, and noise covariance within a single iterative procedure confers consistent advantages over sequential estimation across the full range of SNR, block size, and source configuration tested. The breadth of this advantage suggests that coupling the three estimation problems is the primary driver, supporting the general principle that jointly consistent estimates produce better spatial filter inputs than independently optimized ones.

Second, covariance domain processing provides practically meaningful noise robustness unavailable to coefficient-domain methods. At -10 dB SNR, where instantaneous observations are dominated by noise, the framework maintains HRTF estimation accuracy below the sequential baseline’s high-SNR floor, which can be seen as a consequence of two-level averaging: temporal averaging in the CPSDM estimation step and spectral averaging in the atom selection step. This is directly relevant to hearing aid applications, where spatial processing is most critical when noise is severe.

Third, the internal consistency of jointly produced estimates propagates into beamformer output quality beyond what individual estimate accuracy predicts. The proposed method’s stable target ITD preservation across the full SNR range reflects the mutual compatibility of the steering vectors and noise covariance supplied to the beamformer, a structural property of the joint framework that cannot be replicated by combining separately optimized estimators, and one whose implications extend beyond the specific conditions evaluated here.

6.3 Future Work

The limitations identified in Section 5.5 define four concrete priorities for extending the framework toward real-world deployment.

Off-grid source recovery. Assuming source directions coincide with dictionary entries is rarely satisfied in practice and is the single most consequential open problem. Continuous basis pursuit and atomic norm minimization are promising extensions that can in principle be incorporated into the covariance domain pursuit without modifying the core iterative structure; their relative merits in the binaural HRTF estimation setting warrant systematic evaluation.

Reverberant acoustic conditions. The free-field dictionary does not account for room impulse responses present in real environments. Mitigation strategies include room-adapted dictionaries or a dereverberation preprocessing front-end procedure, the latter being particularly attractive as it leaves the core algorithm unchanged. Eval-

uating the framework across a range of reverberation times would characterize the degradation gradient and identify conditions under which the free-field approximation remains acceptable.

Increased number of interferers. All experiments used $d = 2$ sources; real scenes routinely contain more simultaneous talkers. Systematic evaluation across $d \in \{2, 3, 4\}$ would characterize how HRTF estimation accuracy and target ILD preservation scale as the signal subspace becomes more crowded and dictionary coherence effects compound across more selected atoms.

Data-driven stopping criterion. The known sparsity assumption is a significant barrier to deployment where source count is unknown and time-varying. The modular stopping criterion design means that data-driven alternatives (e.g. MDL, AIC, or eigenvalue gap criteria) can be substituted without modifying the atom selection or residual update steps, making this a contained extension with potentially high practical impact.

References

- [1] A. Aich and P. Palanisamy. “On-grid DOA estimation method using orthogonal matching pursuit”. en. In: *2017 International Conference on Signal Processing and Communication (ICSPC)*. Coimbatore: IEEE, July 2017, pp. 483–487. ISBN: 978-1-5090-6730-5. DOI: [10.1109/CSPC.2017.8305896](https://doi.org/10.1109/CSPC.2017.8305896). URL: <http://ieeexplore.ieee.org/document/8305896/>.
- [2] Y. Avargel and I. Cohen. “On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain”. en. In: *IEEE Signal Processing Letters* 14.5 (May 2007), pp. 337–340. ISSN: 1070-9908. DOI: [10.1109/LSP.2006.888292](https://doi.org/10.1109/LSP.2006.888292).
- [3] D. Barchiesi. “Sparse Approximation and Dictionary Learning with Applications to Audio Signals”. en. PhD thesis. Queen Mary University of London, 2013.
- [4] H. Beit-On, T. Shlomo, and B. Rafaely. “Weighted Frequency Smoothing for Enhanced Speaker Localization”. en. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2090–2099. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2023.3277270](https://doi.org/10.1109/TASLP.2023.3277270).
- [5] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. en. Vol. 1. Springer Topics in Signal Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. ISBN: 978-3-540-78611-5. DOI: [10.1007/978-3-540-78612-2](https://doi.org/10.1007/978-3-540-78612-2). URL: <http://link.springer.com/10.1007/978-3-540-78612-2>.
- [6] S. P. Chepuri and G. Leus. “Sparse Sensing for Statistical Inference”. en. In: *Foundations and Trends® in Signal Processing* 9.3–4 (2016), pp. 233–368. ISSN: 1932-8346, 1932-8354. DOI: [10.1561/20000000069](https://doi.org/10.1561/20000000069).
- [7] I. Cohen. “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging”. en. In: *IEEE Transactions on Speech and Audio Processing* 11.5 (Sept. 2003), pp. 466–475. ISSN: 1063-6676. DOI: [10.1109/TSA.2003.811544](https://doi.org/10.1109/TSA.2003.811544).
- [8] I. Cohen. “Relative Transfer Function Identification Using Speech Signals”. en. In: *IEEE Transactions on Speech and Audio Processing* 12.5 (Sept. 2004), pp. 451–459. ISSN: 1063-6676. DOI: [10.1109/TSA.2004.832975](https://doi.org/10.1109/TSA.2004.832975).
- [9] B. Cornelis et al. “Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques”. en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.2 (Feb. 2010), pp. 342–355. ISSN: 1558-7916. DOI: [10.1109/TASL.2009.2028374](https://doi.org/10.1109/TASL.2009.2028374).
- [10] H. Cox. “Resolving power and sensitivity to mismatch of optimum array processors”. en. In: *The Journal of the Acoustical Society of America* 54.3 (Sept. 1973), pp. 771–785. ISSN: 0001-4966, 1520-8524. DOI: [10.1121/1.1913659](https://doi.org/10.1121/1.1913659).
- [11] F. Denk et al. “Adapting Hearing Devices to the Individual Ear Acoustics: Database and Target Response Correction Functions for Various Device Styles”. en. In: *Trends in Hearing* 22 (Jan. 2018), pp. 1–19. ISSN: 2331-2165, 2331-2165. DOI: [10.1177/2331216518779313](https://doi.org/10.1177/2331216518779313).
- [12] J.-F. Determe et al. “On The Exact Recovery Condition of Simultaneous Orthogonal Matching Pursuit”. en. In: *IEEE Signal Processing Letters* 23.1 (Jan. 2016), pp. 164–168. ISSN: 1070-9908, 1558-2361. DOI: [10.1109/LSP.2015.2506989](https://doi.org/10.1109/LSP.2015.2506989).
- [13] J.-F. Determe et al. “On the Noise Robustness of Simultaneous Orthogonal Matching Pursuit”. en. In: *IEEE Transactions on Signal Processing* 65.4 (Feb. 2017), pp. 864–875. ISSN: 1053-587X, 1941-0476. DOI: [10.1109/TSP.2016.2626244](https://doi.org/10.1109/TSP.2016.2626244).

- [14] C. Ekanadham, D. Tranchina, and E. P. Simoncelli. “Recovery of Sparse Translation-Invariant Signals With Continuous Basis Pursuit”. en. In: *IEEE Transactions on Signal Processing* 59.10 (Oct. 2011), pp. 4735–4744. ISSN: 1053-587X, 1941-0476. DOI: [10.1109/TSP.2011.2160058](https://doi.org/10.1109/TSP.2011.2160058).
- [15] D. Fejgin and S. Doclo. “Comparison of Binaural RTF-Vector-Based Direction of Arrival Estimation Methods Exploiting an External Microphone”. en. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. Dublin, Ireland: IEEE, Aug. 2021, pp. 241–245. ISBN: 978-90-827970-6-0. DOI: [10.23919/EUSIPCO54536.2021.9616327](https://doi.org/10.23919/EUSIPCO54536.2021.9616327). URL: <https://ieeexplore.ieee.org/document/9616327/>.
- [16] S. Gannot, D. Burshtein, and E. Weinstein. “Signal enhancement using beamforming and nonstationarity with applications to speech”. en. In: *IEEE Transactions on Signal Processing* 49.8 (Aug. 2001), pp. 1614–1626. ISSN: 1053587X. DOI: [10.1109/78.934132](https://doi.org/10.1109/78.934132).
- [17] S. Gannot et al. “A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation”. en. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4 (Apr. 2017), pp. 692–730. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2016.2647702](https://doi.org/10.1109/TASLP.2016.2647702).
- [18] S. Gao et al. “A Modified Frequency Weighted MUSIC Algorithm for Multiple Sound Sources Localization”. en. In: *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*. Shanghai, China: IEEE, Nov. 2018, pp. 1–4. ISBN: 978-1-5386-6811-5. DOI: [10.1109/ICDSP.2018.8631636](https://doi.org/10.1109/ICDSP.2018.8631636). URL: <https://ieeexplore.ieee.org/document/8631636/>.
- [19] T. Goehring et al. “Tolerable delay for speech production and perception: effects of hearing ability and experience with hearing aids”. en. In: *International Journal of Audiology* 57.1 (Jan. 2018), pp. 61–68. ISSN: 1499-2027, 1708-8186. DOI: [10.1080/14992027.2017.1367848](https://doi.org/10.1080/14992027.2017.1367848).
- [20] I. M. Griffith, R. P. Hess, and J. H. McDermott. “Optimized feature gains explain and predict successes and failures of human selective listening”. en. In: *Nature Human Behaviour* (Mar. 2026). ISSN: 2397-3374. DOI: [10.1038/s41562-026-02414-7](https://doi.org/10.1038/s41562-026-02414-7). URL: <https://www.nature.com/articles/s41562-026-02414-7>.
- [21] E. Hadad, S. Doclo, and S. Gannot. “A generalized binaural MVDR beamformer with interferer relative transfer function preservation”. en. In: *2016 24th European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: IEEE, Aug. 2016, pp. 1643–1647. ISBN: 978-0-9928626-5-7. DOI: [10.1109/EUSIPCO.2016.7760527](https://doi.org/10.1109/EUSIPCO.2016.7760527). URL: <http://ieeexplore.ieee.org/document/7760527/>.
- [22] P. Hoang et al. “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming”. en. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, June 2021, pp. 6119–6123. ISBN: 978-1-7281-7605-5. DOI: [10.1109/ICASSP39728.2021.9414252](https://doi.org/10.1109/ICASSP39728.2021.9414252). URL: <https://ieeexplore.ieee.org/document/9414252/>.
- [23] A.H. Kamkar-Parsi and M. Bouchard. “Improved Noise Power Spectrum Density Estimation for Binaural Hearing Aids Operating in a Diffuse Noise Field Environment”. en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4 (May 2009), pp. 521–533. ISSN: 1558-7916. DOI: [10.1109/TASL.2008.2009017](https://doi.org/10.1109/TASL.2008.2009017).
- [24] C. A. Kokke, R. C. Hendriks, and A. I. Koutrouvelis. “Binaural Beamforming Based on Automatic Interferer Selection”. en. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United

- Kingdom: IEEE, May 2019, pp. 6850–6854. ISBN: 978-1-4799-8131-1. DOI: [10.1109/ICASSP.2019.8683417](https://doi.org/10.1109/ICASSP.2019.8683417). URL: <https://ieeexplore.ieee.org/document/8683417/>.
- [25] A. I. Koutrouvelis et al. “Binaural beamforming using pre-determined relative acoustic transfer functions”. en. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. Kos, Greece: IEEE, Aug. 2017, pp. 1–5. ISBN: 978-0-9928626-7-1. DOI: [10.23919/EUSIPCO.2017.8081157](https://doi.org/10.23919/EUSIPCO.2017.8081157). URL: <http://ieeexplore.ieee.org/document/8081157/>.
 - [26] A. I. Koutrouvelis et al. “Improved multi-microphone noise reduction preserving binaural cues”. en. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 460–464. ISBN: 978-1-4799-9988-0. DOI: [10.1109/ICASSP.2016.7471717](https://doi.org/10.1109/ICASSP.2016.7471717). URL: <http://ieeexplore.ieee.org/document/7471717/>.
 - [27] A. I. Koutrouvelis et al. “Relaxed Binaural LCMV Beamforming”. en. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (Jan. 2017), pp. 137–152. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2016.2628642](https://doi.org/10.1109/TASLP.2016.2628642).
 - [28] A. I. Koutrouvelis et al. “Robust Joint Estimation of Multimicrophone Signal Model Parameters”. en. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (July 2019), pp. 1136–1150. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2019.2911167](https://doi.org/10.1109/TASLP.2019.2911167).
 - [29] A. Kulkarni and H. S. Colburn. “Role of spectral detail in sound-source localization”. en. In: *Nature* 396.6713 (Dec. 1998), pp. 747–749. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/25526](https://doi.org/10.1038/25526).
 - [30] C. Li and R. C. Hendriks. “Multimicrophone Signal Parameter Estimation in a Multi-Source Noisy Reverberant Scenario”. en. In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 678–692. ISSN: 2998-4173. DOI: [10.1109/TASLPRO.2025.3533374](https://doi.org/10.1109/TASLPRO.2025.3533374).
 - [31] C. Li, J. Martinez, and R. C. Hendriks. “Low Complex Accurate Multi-Source RTF Estimation”. en. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 4953–4957. ISBN: 978-1-6654-0540-9. DOI: [10.1109/ICASSP43922.2022.9747170](https://doi.org/10.1109/ICASSP43922.2022.9747170). URL: <https://ieeexplore.ieee.org/document/9747170/>.
 - [32] X. Li et al. “Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction”. en. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 320–324. ISBN: 978-1-4673-6997-8. DOI: [10.1109/ICASSP.2015.7177983](https://doi.org/10.1109/ICASSP.2015.7177983). URL: <http://ieeexplore.ieee.org/document/7177983/>.
 - [33] M. Maazaoui, K. Abed-Meraim, and Y. Grenier. “Blind source separation for robot audition using fixed HRTF beamforming”. en. In: *EURASIP Journal on Advances in Signal Processing* 2012.1 (Dec. 2012), p. 58. ISSN: 1687-6180. DOI: [10.1186/1687-6180-2012-58](https://doi.org/10.1186/1687-6180-2012-58).
 - [34] S. Markovich, S. Gannot, and I. Cohen. “Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals”. en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6 (Aug. 2009), pp. 1071–1086. ISSN: 1558-7916. DOI: [10.1109/TASL.2009.2016395](https://doi.org/10.1109/TASL.2009.2016395).
 - [35] S. Markovich-Golan and S. Gannot. “Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance

- whitening method”. en. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 544–548. ISBN: 978-1-4673-6997-8. DOI: [10.1109/ICASSP.2015.7178028](https://doi.org/10.1109/ICASSP.2015.7178028). URL: <http://ieeexplore.ieee.org/document/7178028/>.
- [36] P. A. Naylor and N. D. Gaubitch. *Speech Dereverberation*. en. Signals and Communication Technology. London: Springer London, 2010. ISBN: 978-1-84996-055-7. DOI: [10.1007/978-1-84996-056-4](https://doi.org/10.1007/978-1-84996-056-4). URL: <http://link.springer.com/10.1007/978-1-84996-056-4>.
- [37] E. Ollila. “Matching Pursuit Covariance Learning”. en. In: *2024 32nd European Signal Processing Conference (EUSIPCO)*. Lyon, France: IEEE, Aug. 2024, pp. 2447–2451. ISBN: 978-94-645936-1-7. DOI: [10.23919/EUSIPCO63174.2024.10715086](https://doi.org/10.23919/EUSIPCO63174.2024.10715086). URL: <https://ieeexplore.ieee.org/document/10715086/>.
- [38] E. Ollila. “Sparse signal recovery and source localization via covariance learning”. en. In: arXiv:2401.13975 (Aug. 2024). arXiv:2401.13975 [stat]. DOI: [10.48550/arXiv.2401.13975](https://doi.org/10.48550/arXiv.2401.13975). URL: <http://arxiv.org/abs/2401.13975>.
- [39] World Health Organization, ed. *World Report on Hearing*. en. 1st ed. Geneva: World Health Organization, 2021. ISBN: 978-92-4-002048-1. URL: <https://www.who.int/publications/i/item/9789240020481>.
- [40] J. Pauwels and L. Picinali. “On the Relevance of the Differences Between HRTF Measurement Setups for Machine Learning”. en. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, June 2023, pp. 1–5. ISBN: 978-1-7281-6327-7. DOI: [10.1109/ICASSP49357.2023.10096689](https://doi.org/10.1109/ICASSP49357.2023.10096689). URL: <https://ieeexplore.ieee.org/document/10096689/>.
- [41] D. Pavlidis et al. “Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array”. en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (Oct. 2013), pp. 2193–2206. ISSN: 1558-7916, 1558-7924. DOI: [10.1109/TASL.2013.2272524](https://doi.org/10.1109/TASL.2013.2272524).
- [42] K. B. Petersen, M. S. Pedersen, et al. “The matrix cookbook”. In: *Technical University of Denmark* 7.15 (2008), p. 510.
- [43] J. Richter et al. “EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation”. en. In: arXiv:2406.06185 (June 2024). arXiv:2406.06185 [eess]. DOI: [10.48550/arXiv.2406.06185](https://doi.org/10.48550/arXiv.2406.06185). URL: <http://arxiv.org/abs/2406.06185>.
- [44] R. Roy and T. Kailath. “ESPRIT-estimation of signal parameters via rotational invariance techniques”. en. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.7 (July 1989), pp. 984–995. ISSN: 00963518. DOI: [10.1109/29.32276](https://doi.org/10.1109/29.32276).
- [45] R. Schmidt. “Multiple emitter location and signal parameter estimation”. en. In: *IEEE Transactions on Antennas and Propagation* 34.3 (Mar. 1986), pp. 276–280. ISSN: 0018-926X, 1558-2221. DOI: [10.1109/TAP.1986.1143830](https://doi.org/10.1109/TAP.1986.1143830).
- [46] R. Serizel et al. “Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants”. en. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (Apr. 2014), pp. 785–799. ISSN: 2329-9290, 2329-9304. DOI: [10.1109/TASLP.2014.2304240](https://doi.org/10.1109/TASLP.2014.2304240).
- [47] M. Souden, J. Benesty, and S. Affes. “A Study of the LCMV and MVDR Noise Reduction Filters”. en. In: *IEEE Transactions on Signal Processing* 58.9 (Sept. 2010), pp. 4925–4935. ISSN: 1053-587X, 1941-0476. DOI: [10.1109/TSP.2010.2051803](https://doi.org/10.1109/TSP.2010.2051803).

- [48] C. Stockle et al. “1-bit direction of arrival estimation based on Compressed Sensing”. en. In: *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. Stockholm, Sweden: IEEE, June 2015, pp. 246–250. ISBN: 978-1-4799-1931-4. DOI: [10.1109/SPAWC.2015.7227037](https://doi.org/10.1109/SPAWC.2015.7227037). URL: <http://ieeexplore.ieee.org/document/7227037/>.
- [49] D. S. Talagala et al. “Binaural sound source localization using the frequency diversity of the head-related transfer function”. en. In: *The Journal of the Acoustical Society of America* 135.3 (Mar. 2014), pp. 1207–1217. ISSN: 0001-4966, 1520-8524. DOI: [10.1121/1.4864304](https://doi.org/10.1121/1.4864304).
- [50] M. Tammen, S. Doclo, and I. Kodrasi. “Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares”. en. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 795–799. ISBN: 978-1-4799-8131-1. DOI: [10.1109/ICASSP.2019.8683321](https://doi.org/10.1109/ICASSP.2019.8683321). URL: <https://ieeexplore.ieee.org/document/8683321/>.
- [51] G. Tang et al. “Compressed Sensing Off the Grid”. en. In: *IEEE Transactions on Information Theory* 59.11 (Nov. 2013), pp. 7465–7490. ISSN: 0018-9448, 1557-9654. DOI: [10.1109/TIT.2013.2277451](https://doi.org/10.1109/TIT.2013.2277451).
- [52] M. Taseska and E. A. P. Habets. “Relative transfer function estimation exploiting instantaneous signals and the signal subspace”. en. In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. Nice: IEEE, Aug. 2015, pp. 404–408. ISBN: 978-0-9928626-3-3. DOI: [10.1109/EUSIPCO.2015.7362414](https://doi.org/10.1109/EUSIPCO.2015.7362414). URL: <http://ieeexplore.ieee.org/document/7362414/>.
- [53] J. Thiemann et al. “Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene”. en. In: *EURASIP Journal on Advances in Signal Processing* 2016.1 (Dec. 2016), p. 12. ISSN: 1687-6180. DOI: [10.1186/s13634-016-0314-6](https://doi.org/10.1186/s13634-016-0314-6).
- [54] J. A. Tropp. “Algorithms for simultaneous sparse approximation. Part II: Convex relaxation”. en. In: *Signal Processing* 86.3 (Mar. 2006), pp. 589–602. ISSN: 01651684. DOI: [10.1016/j.sigpro.2005.05.031](https://doi.org/10.1016/j.sigpro.2005.05.031).
- [55] J. A. Tropp and A. C. Gilbert. “Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit”. en. In: *IEEE Transactions on Information Theory* 53.12 (Dec. 2007), pp. 4655–4666. ISSN: 0018-9448. DOI: [10.1109/TIT.2007.909108](https://doi.org/10.1109/TIT.2007.909108).
- [56] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. “Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit”. en. In: *Signal Processing* 86.3 (Mar. 2006), pp. 572–588. ISSN: 01651684. DOI: [10.1016/j.sigpro.2005.05.030](https://doi.org/10.1016/j.sigpro.2005.05.030).
- [57] T. Van den Bogaert et al. “The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids”. en. In: *The Journal of the Acoustical Society of America* 124.1 (July 2008), pp. 484–497. ISSN: 0001-4966, 1520-8524. DOI: [10.1121/1.2931962](https://doi.org/10.1121/1.2931962).
- [58] R. Varzandeh, M. Taseska, and E. A. P. Habets. “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”. en. In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. San Francisco, CA, USA: IEEE, 2017, pp. 11–15. ISBN: 978-1-5090-5925-6. DOI: [10.1109/HSCMA.2017.7895552](https://doi.org/10.1109/HSCMA.2017.7895552). URL: <http://ieeexplore.ieee.org/document/7895552/>.
- [59] E. Warsitz and R. Haeb-Umbach. “Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition”. en. In: *IEEE Transactions on Audio, Speech and Language*

- Processing* 15.5 (July 2007), pp. 1529–1539. ISSN: 1558-7916. DOI: [10.1109/TASL.2007.898454](https://doi.org/10.1109/TASL.2007.898454).
- [60] M. Wax and T. Kailath. “Detection of signals by information theoretic criteria”. en. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33.2 (Apr. 1985), pp. 387–392. ISSN: 0096-3518. DOI: [10.1109/TASSP.1985.1164557](https://doi.org/10.1109/TASSP.1985.1164557).
 - [61] P. Welch. “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms”. en. In: *IEEE Transactions on Audio and Electroacoustics* 15.2 (June 1967), pp. 70–73. ISSN: 0018-9278. DOI: [10.1109/TAU.1967.1161901](https://doi.org/10.1109/TAU.1967.1161901).
 - [62] F. L. Wightman and D. J. Kistler. “The dominant role of low-frequency interaural time differences in sound localization”. en. In: *The Journal of the Acoustical Society of America* 91.3 (Mar. 1992), pp. 1648–1661. ISSN: 0001-4966, 1520-8524. DOI: [10.1121/1.402445](https://doi.org/10.1121/1.402445).
 - [63] K. J. P. Woods and J. H. McDermott. “Attentive Tracking of Sound Sources”. en. In: *Current Biology* 25.17 (Aug. 2015), pp. 2238–2246. ISSN: 09609822. DOI: [10.1016/j.cub.2015.07.043](https://doi.org/10.1016/j.cub.2015.07.043).
 - [64] H. Yang et al. “Speech enhancement using orthogonal matching pursuit algorithm”. en. In: *2014 International Conference on Orange Technologies*. Xi’an, China: IEEE, Sept. 2014, pp. 101–104. ISBN: 978-1-4799-6284-6. DOI: [10.1109/ICOT.2014.6956609](https://doi.org/10.1109/ICOT.2014.6956609). URL: <http://ieeexplore.ieee.org/document/6956609/>.
 - [65] W. A. Yost and X. Zhong. “Sound source localization identification accuracy: Bandwidth dependencies”. en. In: *The Journal of the Acoustical Society of America* 136.5 (Nov. 2014), pp. 2737–2746. ISSN: 0001-4966, 1520-8524. DOI: [10.1121/1.4898045](https://doi.org/10.1121/1.4898045).
 - [66] W. Zhang, R. A. Kennedy, and T. D. Abhayapala. “Efficient Continuous HRTF Model Using Data Independent Basis Functions: Experimentally Guided Approach”. en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4 (May 2009), pp. 819–829. ISSN: 1558-7916. DOI: [10.1109/TASL.2009.2014265](https://doi.org/10.1109/TASL.2009.2014265).

Derivations



A.1 The Signal Energy Update Closed-Form Solution

To solve (4.9), we exploit the structure of the objective. The Frobenius norm in (4.9) can be written as

$$\left\| \mathbf{R}_x(k) - \tilde{\mathcal{A}}_\Lambda(k) \mathbf{R}_s(k) \tilde{\mathcal{A}}_\Lambda^H(k) \right\|_F^2 = \text{tr} \left[\left(\mathbf{R}_x(k) - \tilde{\mathcal{A}}_\Lambda(k) \mathbf{R}_s(k) \tilde{\mathcal{A}}_\Lambda^H(k) \right)^H \left(\mathbf{R}_x(k) - \tilde{\mathcal{A}}_\Lambda(k) \mathbf{R}_s(k) \tilde{\mathcal{A}}_\Lambda^H(k) \right) \right]. \quad (\text{A.1})$$

Taking the derivative with respect to $\mathbf{R}_s^*(k)$ using Wirtinger calculus [42] and setting to zero gives the stationarity condition

$$\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \mathbf{R}_s(k) \tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) = \tilde{\mathcal{A}}_\Lambda^H(k) \mathbf{R}_x(k) \tilde{\mathcal{A}}_\Lambda(k). \quad (\text{A.2})$$

Provided that $\tilde{\mathcal{A}}_\Lambda(k)$ has full column rank, the matrix $\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \in \mathbb{C}^{p \times p}$ is invertible, and (A.2) can be solved by left- and right-multiplying by $\left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1}$, yielding

$$\hat{\mathbf{R}}_s^{(p)}(k) = \tilde{\mathcal{A}}_\Lambda^\dagger(k) \mathbf{R}_x(k) \tilde{\mathcal{A}}_\Lambda^H(k), \quad (\text{A.3})$$

where $\tilde{\mathcal{A}}_\Lambda^\dagger(k) = \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k) \in \mathbb{C}^{p \times M}$ is the Moore-Penrose pseudoinverse of $\tilde{\mathcal{A}}_\Lambda(k)$.

A.2 The Projection Matrix and Its Properties

Given the sub-dictionary $\tilde{\mathcal{A}}_\Lambda(k) \in \mathbb{C}^{M \times p}$ of selected atoms, we define the orthogonal projector onto the column space of $\tilde{\mathcal{A}}_\Lambda(k)$ as

$$\mathbf{P}_\Lambda(k) = \tilde{\mathcal{A}}_\Lambda(k) \tilde{\mathcal{A}}_\Lambda^\dagger(k) = \tilde{\mathcal{A}}_\Lambda(k) \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k). \quad (\text{A.4})$$

This matrix satisfies two properties that are central to the correctness of the residual update. First, it is idempotent, i.e.

$$\mathbf{P}_\Lambda^2(k) = \tilde{\mathcal{A}}_\Lambda(k) \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k) = \mathbf{P}_\Lambda(k), \quad (\text{A.5})$$

confirming that repeated application of the projector leaves the result unchanged, as expected of a true projection. Second, it is Hermitian, i.e.

$$\mathbf{P}_\Lambda^H(k) = \left(\tilde{\mathcal{A}}_\Lambda(k) \left(\tilde{\mathcal{A}}_\Lambda^H(k) \tilde{\mathcal{A}}_\Lambda(k) \right)^{-1} \tilde{\mathcal{A}}_\Lambda^H(k) \right)^H = \mathbf{P}_\Lambda(k), \quad (\text{A.6})$$

since $\left(\tilde{\mathcal{A}}_{\mathbf{\Lambda}}^{\text{H}}(k)\tilde{\mathcal{A}}_{\mathbf{\Lambda}}(k)\right)^{-1}$ is itself Hermitian when $\tilde{\mathcal{A}}_{\mathbf{\Lambda}}(k)$ has full column rank. Together, both properties confirm that $\mathbf{P}_{\mathbf{\Lambda}}(k)$ is an orthogonal projector in the standard sense, and that its complement $\mathbf{I} - \mathbf{P}_{\mathbf{\Lambda}}(k)$ is the orthogonal projector onto the null space of $\tilde{\mathcal{A}}_{\mathbf{\Lambda}}^{\text{H}}(k)$, i.e. the subspace orthogonal to all selected atoms.

Preliminary Experiment

Before taking its current form, the problem was seen purely from a DOA estimation perspective, it being determined by evaluating which dictionary atom or “virtual source” was selected by the proposed method in a single run across time blocks. To do so, several two-source configurations (with sources named S_1 and S_2) were tested across varying angular separations and SNRs. In each configuration, the algorithm selects two virtual sources from the HRTF dictionary, with the corresponding azimuth entries compared against the true source directions over 75 time blocks (time block size fixed at $L = 50$). One azimuth spacing of the HRTF dictionary on either side of the true azimuth ($\pm 7.5^\circ$ in this case) was defined as an acceptance band, and a detection was considered valid if either selected virtual source fell within either band at a given time block, irrespective of source index assignment. Fig. B.1 shows visualizations of the six configurations considered.

At 10 dB SNR with sources at 0° and 30° (Fig. B.1a), tracking was poor for both directions. The second virtual source exhibited large erratic excursions across the full $\pm 180^\circ$ azimuth range, rendering one of the two selections effectively uninformative for much of the evaluation period. The 20 dB increase in SNR (Fig. B.1b, same angular configuration) produced a visible improvement, with the first virtual source showing a stronger tendency to remain within the 0° acceptance band. However, the tight 30° angular separation continued to challenge the algorithm (explained by the high coherence between the two spatial positions), with the second source frequently failing to settle near the S_2 azimuth.

Increasing the angular separation to 60° while maintaining 30 dB SNR (Fig. B.1c, sources at 0° and 60°) yielded a further improvement. S_1 was tracked reliably for extended stretches, and S_2 received more consistent coverage, though transient escapes remained. A more substantial separation of 120° (Fig. B.1d, sources at 0° and 120°) produced notably stable tracking of S_2 , with the selected virtual source remaining within the 120° band across the majority of time blocks. Notably, however, tracking of S_1 was comparatively less stable in this configuration, which can reflect a directional bias in the estimator toward more lateral azimuths when a strong spatial cue is present at a wide angle.

A consistent observation across the four plots was a gravitational tendency of PROP towards S_1 , wherein one virtual source frequently anchored itself near 0° with relatively high persistence. While this benefited detection of S_1 , it appeared to reduce the algorithm’s responsiveness to a nearby S_2 , particularly in close separations. This broadside bias is consistent with the known behavior of beamforming-based spatial estimators, which typically exhibit higher spatial resolution near broadside due to the geometry of the microphone array.

The symmetric configuration at 30 dB SNR with sources at -30° and 30° (Fig. B.1e)

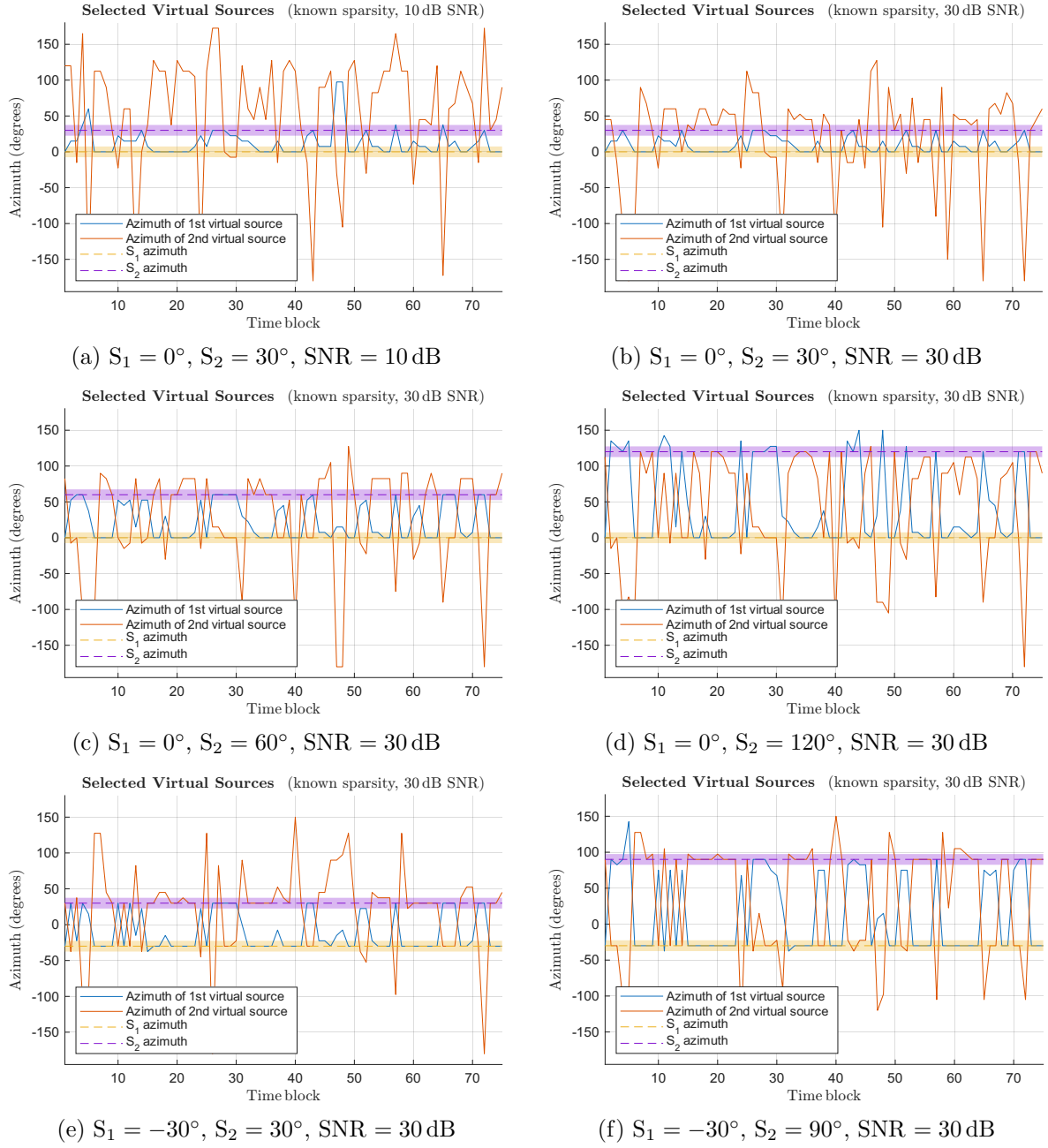


Figure B.1: Virtual source tracking across time blocks.

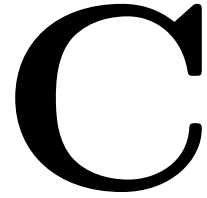
produced the weakest performance among the 30 dB cases, despite sharing the same 60° separation as Fig. B.1c. Both virtual sources alternated extensively between the two bands without stabilizing on either, and there were prolonged periods in which the -30° azimuth received no coverage. This suggests that symmetric placement closer to the frontal direction introduces a geometric ambiguity that the proposed algorithm cannot resolve.

The asymmetric configuration with sources at -30° and 90° (Fig. B.1f) yielded the

most balanced and consistent tracking overall. Both virtual sources spent the majority of time blocks within their respective acceptance bands, with only brief transient departures. The absence of a source at broadside combined with the asymmetric geometry appears to allow the estimator to distribute its selections more evenly across the two true directions without the anchoring effect observed in the broadside configurations.

Taken together, these results suggest that tracking performance is primarily governed by the SNR, which affects the reliability of individual direction estimates, and source geometry (i.e., angular separation and symmetry), which affects the proposed method’s ability to resolve and maintain distinct selections for two simultaneous sources. Wide separations at sufficient SNR produced the most robust detection outcomes under the adopted criterion.

However, upon reflection, these results try to simplify a complex broadband signal like speech to a narrowband signal like radar. The problem can be viewed as one of virtual source selection, but for audio signal processing pipelines, it is standard to instead compare the ATFs/HRTFs estimated, which give a better sense of placement of said “virtual source”. At the same time, analysis drawn from six isolated configurations alone warranted a more thorough investigation into how SNR, time block size and source azimuth configuration affect the performance of the proposed algorithm. All the factors mentioned above led to the switch to HRTF estimation covered in Section 5.2.



Code Snippets

C.1 Covariance-Based Simultaneous Orthogonal Matching Pursuit

```
1 function [selected_az,est_signal_cpsdm_blocks,residual_cpsdm_blocks] =  
    simul_omp_for_spatially_sparse_audio(input_stft,hrtf_dictionary,  
        az_grid)  
%SIMUL_OMP_FOR_SPATIALLY_SPARSE_AUDIO Simultaneous OMP for speech  
%enhancement and virtual source selection  
% This function is an implementation of Simultaneous Orthogonal  
    Matching  
% Pursuit for speech and audio signals (already transformed to  
6 % STFT domain). The aim is to select azimuths from the dictionary of  
    HRTFs  
% which match best to the spatial locations of sources in the real  
    auditory  
% scene recorded in the signals. The azimuths are represented by HRTFs  
% which can be used as "virtual sources", which means real-time  
    estimation  
% of RTFs may no longer be required during beamforming.  
11 % Output:  
% selected_az - indices of azimuths selected (virtual sources selected)  
% output_stft - reconstructed audio (in STFT domain) after S-OMP is  
    performed over  
% input_stft (frequency bins x time frames x channels)  
% residual_stft - residual left in performing S-OMP (in STFT domain) on  
16 % input_stft (frequency bins x time frames x channels)  
% Input:  
% input_stft - measured input in STFT domain (frequency bins x time  
    frames x channels)  
% hrtf_dictionary - tensor with HRTFs stored for each azimuth (  
    frequency bins x channels x azimuths)  
% az_grid - azimuths of virtual sources (HRTFs)  
21 % initializing variables  
[numFreqBins, numTimeFrames, numChannels] = size(input_stft);  
[~,~, numAzimuths] = size(hrtf_dictionary);  
% brute force variable - for algorithm verification  
26 maxIterations = 10; % max number of iterations for algorithm to run in  
    each time block  
numSources = 2; % fixing number of sources based on prior knowledge of  
    mixed audio  
  
% based on practicalities, it is better to only use positive frequencies
```

```

% and/or DC component (ignore negative frequencies)
31 % from FFT/STFT, understood that frequency ordering is in ascending order
    ,
    % which means the first half of frequencies can be ignored
    usableFreqBins_stft = (numFreqBins/2):numFreqBins; % should be 1024:2048
    for 2048-point STFT
    usableFreqBins_fft = (numFreqBins/2):numFreqBins; % should be 1024:2048
    for 2048-point centred (and modified) FFT
    numFreqBins_usable = length(usableFreqBins_stft);
36 eff_input_stft = input_stft(usableFreqBins_stft, :, :);
    eff_hrtf_dictionary = hrtf_dictionary(usableFreqBins_fft, :, :);

    % creating overlapping blocks of 50 time frames (with 50% overlap) to
    track
    % sources
41 % real-time implelemtation: collect time frames, process for that block,
    % then forget half of it and collect more time frames (not in the scope
    of
    % this thesis)
    idxsTimeBlockStart = 1:25:numTimeFrames;
    numTimeBlocks = length(idxsTimeBlockStart);
46 timeBlock_end = 50;
    eff_input_stft = permute(eff_input_stft, [3,1,2]);
    input_cpsdm_blocks = zeros(numChannels, numChannels, numFreqBins_usable,
        numTimeBlocks);
    time_frames_for_cpsdm = timeBlock_end;
    for t = 1:numTimeBlocks
51     if timeBlock_end > numTimeFrames
        timeBlock_end = numTimeFrames;
    end
    for k = 1:numFreqBins_usable
        temp = squeeze(eff_input_stft(:,k,idxsTimeBlockStart(t):
            timeBlock_end));
56     temp2 = temp * temp' ./ (timeBlock_end - idxsTimeBlockStart(t) +
        1);
        input_cpsdm_blocks(:, :, k, t) = temp2 + 1e-3 * trace(temp2) /
            numChannels * eye(size(temp2)); % diagonal loading
    end
    timeBlock_end = timeBlock_end + time_frames_for_cpsdm/2;
end
61 % est_signal_cpsdm_blocks = zeros(numChannels, numChannels,
    numFreqBins_usable, numTimeBlocks); % when sparsity is unknown
    est_signal_cpsdm_blocks = zeros(numSources, numSources,
        numFreqBins_usable, numTimeBlocks); % when sparsity is fixed
    residual_cpsdm_blocks = input_cpsdm_blocks;
    eff_hrtf_dictionary = permute(eff_hrtf_dictionary, [2,3,1]);
    selected_az = [];
66 % normalizing HRTF dictionary along channel dimension
    eff_hrtf_dictionary = eff_hrtf_dictionary ./ vecnorm(eff_hrtf_dictionary,
        2, 1);

    % updated algorithm: perform S-OMP over time blocks, return a matrix

```

```

    array
71 % showing position estimates in all time blocks processed
    for t = 1:numTimeBlocks
        corr_matrix = zeros(numAzimuths, numAzimuths, numFreqBins_usable);
        az_idx_current = [];
        for i = 1:numSources % update iterations HERE!
76         for k = 1:numFreqBins_usable
            corr_matrix(:, :, k) = real(eff_hrtf_dictionary(:, :, k)' *
                residual_cpsdm_blocks(:, :, k, t) * eff_hrtf_dictionary(:, :, k
                ));
            % potentially update stopping criteria logic HERE!
        end
        corr_matrix_total_l1 = vecnorm(corr_matrix, 1, 3);
81        [~, az_index] = max(diag(corr_matrix_total_l1));
        az_idx_current = [az_idx_current; az_index];
        numSourcesIdentified = size(az_idx_current, 1);
        % az_idx_current = sort(az_idx_current);
        for k = 1:numFreqBins_usable
86            active_hrtfs = squeeze(eff_hrtf_dictionary(:, az_idx_current, k
                ));
            pseudo_inverse = (active_hrtfs' * active_hrtfs) \
                active_hrtfs';
            projection = active_hrtfs * pseudo_inverse;
            est_signal_cpsdm_blocks(1:numSourcesIdentified, 1:
                numSourcesIdentified, k, t) = pseudo_inverse *
                input_cpsdm_blocks(:, :, k, t) * pseudo_inverse';
            est_measured_cpsdm = projection * input_cpsdm_blocks(:, :, k, t)
                * projection';
91            residual_cpsdm_blocks(:, :, k, t) = input_cpsdm_blocks(:, :, k, t)
                - est_measured_cpsdm;
        end
        % if norm(residual_cpsdm_blocks(:, :, :, t), 'fro') <= 1e-3
        %     break
        % end
96    end
    selected_az = [selected_az; az_grid(az_idx_current)];
end
end

```

C.2 Joint Binaural Linearly Constrained Minimum Variance (JBLCMV) Beamformer

```

1 function [leftWeights, rightWeights] = jblcmv_util(noiseCovariance,
    targetHRTF, interfererHRTF)
    %JBLCMV_UTIL Joint Binaural Linearly Constrained Minimum Variance
    Beamformer utility
    % Minimizes noise covariance given ATFs of target and interferer(s);
    outputs beamformer weights for left and right hearing-aid units
    [numChannels, ~, numFreqBins] = size(noiseCovariance);
    noiseCovariance_binaural = zeros(2*numChannels, 2*numChannels, numFreqBins)
    ;
6 for k = 1:numFreqBins

```

```

        noiseCovariance_binaural(:, :, k) = kron(eye(2), noiseCovariance(:, :, k))
    ;
end

targetHRTF = permute(targetHRTF, [2, 1]); % M x K
11 targetHRTF_left = targetHRTF ./ targetHRTF(1, :);
    targetHRTF_right = targetHRTF ./ targetHRTF(numChannels, :);
    interfererHRTF = permute(interfererHRTF, [2, 1]); % M x K
    interfererHRTF_left = interfererHRTF ./ interfererHRTF(1, :);
    interfererHRTF_right = interfererHRTF ./ interfererHRTF(numChannels, :);
16
    Lambda_target = zeros(2*numChannels, 2, numFreqBins);
    Lambda_interferer = zeros(2*numChannels, numFreqBins);
    for k = 1:numFreqBins
        Lambda_target(:, :, k) = [kron([1 0], targetHRTF_left(:, k)); ...
21         kron([0 1], targetHRTF_right(:, k))];
        Lambda_interferer(:, k) = [interfererHRTF_left(:, k); ...
            -interfererHRTF_right(:, k)];
    end
    Lambda_interferer_resaped = reshape(Lambda_interferer, [2*numChannels, 1,
        numFreqBins]);
26 Lambda = cat(2, Lambda_target, Lambda_interferer_resaped); % HRTF
    constraints
    f = [1; 1; 0] .* ones(1, numFreqBins); % constraint values

    weights = zeros(2*numChannels, numFreqBins);

31 for k = 1:numFreqBins
    R_tilde = squeeze(noiseCovariance_binaural(:, :, k)); % binaural noise
        CPSDM for kth frequency bin
    Lambda_k = squeeze(Lambda(:, :, k));
    f_k = f(:, k);

36    w_num = R_tilde \ Lambda_k;
    w_den = Lambda_k' * w_num;
    weights(:, k) = w_num * (w_den \ f_k);
end

41 leftWeights = weights(1:numChannels, :);
    rightWeights = weights(numChannels+1:end, :);
end

```

Submitted Paper

D

Parts of this work was submitted as a paper in the **19th International Workshop on Acoustic Signal Enhancement (IWAENC 2026)** on April 24, 2026 after six months of work. Notification of acceptance is on July 3, 2026.

A COVARIANCE-BASED SIMULTANEOUS ORTHOGONAL MATCHING PURSUIT FRAMEWORK FOR SPATIAL SOURCE LOCALIZATION AND BEAMFORMING

Kunal Kashyap, Johannes W. de Vries, Giovanni Bologni and Richard C. Hendriks

Signal Processing Systems (SPS) Group, Delft University of Technology, Delft, The Netherlands

ABSTRACT

Binaural beamforming techniques such as the binaural linearly constrained minimum variance (BLCMV) beamformer require two key inputs: the acoustic transfer functions (ATFs) of the target and interfering sources, and the noise cross power spectral density matrix (CPSDM). These quantities are typically estimated independently, adding system complexity and compounded estimation error. We propose a greedy sparse pursuit framework operating in the covariance domain that estimates the head-related transfer function (HRTF) subspace from a predetermined dictionary while yielding a residual CPSDM as a byproduct, supplying both beamforming ingredients within a unified framework. The proposed method achieves consistently lower HRTF estimation error than a GEVD-based baseline across SNR, time block size, and source configuration. When used to drive a joint BLCMV beamformer, it yields near-zero and stable target interaural phase difference errors across a broad SNR range, while target interaural level difference preservation remains an open challenge shared by both methods.

Index Terms— Binaural beamforming, sparse recovery, orthogonal matching pursuit, covariance matrix estimation, interaural cues

1. INTRODUCTION

Hearing loss affects over 1.5 billion people worldwide, and hearing-aid devices represent a primary intervention for restoring speech intelligibility in complex acoustic environments [1]. Binaural beamforming is a central technique in hearing-aid signal processing, enabling spatial filtering to suppress interference while preserving target speech and aiming to preserve their spatial information. Linearly constrained minimum variance (LCMV) and related beamformers [2, 3, 4] require accurate knowledge of two types of quantities: the acoustic transfer functions (ATFs), encoding the acoustic path from each source to the microphone array and used in the beamformer constraint set, and the noise cross power spectral density matrix (CPSDM), used in the objective function. Traditionally, these are estimated separately: ATF estimation from the observed mixture (e.g., GEVD-based methods) [5, 6] and noise CPSDM estimation often relying

on voice activity detection (VAD) or noise-only periods [7, 8], introducing computational complexity and compounded estimation error. Joint estimation approaches have also been explored [9], but do not exploit the structural prior offered by predetermined head-related transfer function (HRTF) measurements in sparse multi-source scenarios. Dictionaries of such measurements have been proposed [10] as an alternative to real-time estimation in multi-source scenarios [11, 12], but existing approaches constrain the beamformer to all dictionary entries simultaneously, limiting degrees of freedom for noise reduction.

In this paper, we propose a covariance-domain sparse pursuit method that selectively identifies the active HRTF subspace from a predetermined dictionary. The algorithm jointly provides an estimated HRTF matrix and a residual CPSDM interpretable as a noise covariance estimate, supplying both beamforming ingredients within a unified framework. It achieves consistently lower HRTF estimation error than a GEVD-based baseline, and when used to drive a joint binaural LCMV (JBLCMV) beamformer [4], yields near-zero and stable target interaural phase difference errors across a broad SNR range. The remainder of this paper is organized as follows: Section 2 covers background, Section 3 the proposed algorithm, Section 4 simulations and discussion, and Section 5 conclusions.

2. BACKGROUND

Consider a hearing-aid system with M microphones and d simultaneously active point sources, where source 1 is the target and sources $2, \dots, d$ are interferers. In the STFT domain, the $M \times 1$ observation vector at frequency bin k and time frame l is

$$\mathbf{x}(k, l) = \underbrace{\mathbf{A}(k, l)\mathbf{s}(k, l)}_{\mathbf{y}(k, l)} + \mathbf{n}(k, l), \quad (1)$$

where $\mathbf{A}(k, l) = [\mathbf{a}_1(k, l) \cdots \mathbf{a}_d(k, l)]$ is the ATF matrix, $\mathbf{s}(k, l) = [s_1(k, l) \cdots s_d(k, l)]^T$ is the source vector, and $\mathbf{n}(k, l)$ captures late reverberation, ambient noise, and microphone self-noise. Assuming mutually independent sources,

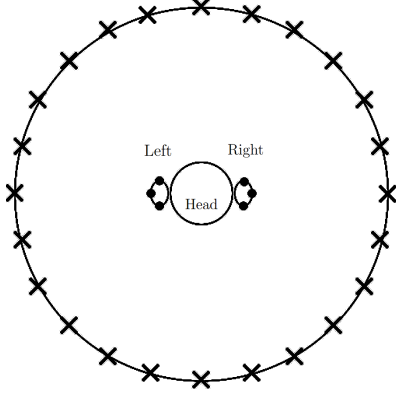


Fig. 1: Virtual source dictionary with $N = 24$ azimuths.

the cross power spectral density matrix (CPSDM) of $\mathbf{x}$ is:

$$\begin{aligned} \mathbf{R}_{\mathbf{x}}(k, l) &= \mathbb{E} [\mathbf{x}(k, l) \mathbf{x}^H(k, l)] \\ &= \underbrace{\mathbf{A}(k, l) \mathbf{R}_{\mathbf{s}}(k, l) \mathbf{A}^H(k, l)}_{\mathbf{R}_{\mathbf{y}}(k, l)} + \mathbf{R}_{\mathbf{n}}(k, l), \end{aligned} \quad (2)$$

where $\mathbb{E}$ denotes the expectation operator and $\{\cdot\}^H$ the Hermitian transpose. $\mathbf{R}_{\mathbf{s}}(k, l)$ is diagonal by the independence assumption.

Accurate knowledge of $\mathbf{A}(k, l)$ and $\mathbf{R}_{\mathbf{n}}(k, l)$ is required for minimum variance-based beamforming, yet both are generally unknown in practice. Rather than estimating these in real time, we follow Koutrovelis et al. [10] and use HRTFs measured at predetermined directions as a dictionary. They take the head of the hearing-aid user into account which motivates their use as an approximation to ATFs from the user's perspective. Unlike [10], however, which constrains the beamformer to all dictionary entries simultaneously (limiting degrees of freedom for noise reduction), we propose to select only the dictionary entries closest to the active sources in terms of ATF/HRTF similarity. HRTFs measured at N locations form the dictionary $\mathcal{A} \in \mathbb{C}^{M \times N \times K}$, with the k^{th} frequency-bin slice being $\mathcal{A}(k) = [\mathbf{h}_1(k) \ \mathbf{h}_2(k) \ \cdots \ \mathbf{h}_N(k)]$, where $\mathbf{h}_n(k)$ is one such measurement for the n^{th} azimuthal location. Fig. 1 shows the virtual source locations of an example dictionary with $N = 24$ azimuths.

Although a general dictionary could include HRTF measurements from non-zero elevations and different distances, we limit ourselves to zero elevation and a distance satisfying the far-field assumption for the sake of convenience. Moreover, the dictionary is used under the following assumptions: (i) the number of real sources $d \ll N$, and $d \leq M$, so that $N \geq M$; (ii) sources are static or slowly moving, so $\mathbf{A}(k, l) \approx \mathbf{A}(k)$ over a block of L frames; and (iii) when the directions of source i and dictionary align, $\mathbf{a}_i(k) \subset \mathcal{A}(k)$, and when they do not, the dictionary provides the approximation.

3. SPATIAL COVARIANCE-BASED SIMULTANEOUS ORTHOGONAL MATCHING PURSUIT

Orthogonal Matching Pursuit (OMP) [13] and its multi-measurement extension Simultaneous OMP (S-OMP) [14] recover sparse representations from overcomplete dictionaries by iteratively selecting the atom most correlated with the residual, refining via orthogonal projection. However, classical OMP variants operate on signal coefficients and do not account for the spatial covariance structure central to beamforming. We therefore propose a covariance-domain extension of S-OMP that operates on the CPSDM computed over a short collection of consecutive time frames (i.e. time block), exploiting second-order spatial statistics for robust virtual source selection.

3.1. Proposed Algorithm

The proposed method takes as input the measured CPSDM $\mathbf{R}_{\mathbf{x}}(k) \in \mathbb{C}^{M \times M}$, the dictionary of HRTFs $\mathcal{A}(k) \in \mathbb{C}^{M \times N}$ normalized along the channel dimension and a stopping criterion, and returns the estimated signal energy tensor $\hat{\mathbf{R}}_{\mathbf{s}}(k)$, residual $\mathbf{R}_{res}(k)$, support set Λ , and iteration count p . Conceptually, Λ identifies which dictionary entries (i.e. which HRTFs) are associated with active sources in the scene. The diagonal elements of $\hat{\mathbf{R}}_{\mathbf{s}}(k)$ hold the scaling factors of the signal energy contributed by a source and $\mathbf{R}_{res}(k)$ represents the portion of the observed CPSDM unexplained by the selected sources, interpretable as a noise covariance estimate. Together, these outputs directly supply the two quantities required by a beamformer as discussed in Section 2. Initialized with $\mathbf{R}_{res}^{(0)}(k) = \mathbf{R}_{\mathbf{x}}(k)$ and $\Lambda^{(0)} = \emptyset$, the algorithm iterates as follows.

Atom selection. At iteration p , the atom maximizing aggregated projection energy across frequency bins is selected by

$$\lambda^{(p)} = \arg \max_{1 \leq n \leq N} \left| \sum_{k=1}^K \mathbf{h}_n^H(k) \mathbf{R}_{res}^{(p-1)}(k) \mathbf{h}_n(k) \right|, \quad (3)$$

promoting consistent support across frequencies. Rather than evaluating (3) independently for each atom, all N correlations can be obtained simultaneously by extracting the diagonal of the full dictionary projection matrix, which is then accumulated across frequency bins before maximizing via $\mathcal{A}^H(k) \mathbf{R}_{res}^{(p-1)}(k) \mathcal{A}(k)$ across k through

$$\mathbf{c}^{(p)}(k) = \text{diag} \left(\mathcal{A}^H(k) \mathbf{R}_{res}^{(p-1)}(k) \mathcal{A}(k) \right), \quad (4)$$

$$\mathbf{c}_{agg}^{(p)} = \sum_{k=1}^K \mathbf{c}^{(p)}(k), \quad \lambda^{(p)} = \arg \max_n \left| (c_{agg}^{(p)})_n \right|. \quad (5)$$

The support is augmented as $\Lambda^{(p)} = \Lambda^{(p-1)} \cup \{\lambda^{(p)}\}$, and $\mathcal{A}_{\Lambda}(k) \in \mathbb{C}^{M \times p}$ denotes the sub-dictionary of selected atoms from $\mathcal{A}(k)$.

Algorithm 1 Covariance-Based S-OMP (PROP)

- 1: **Input:** Measured CPSD tensor $\mathbf{R}_\mathbf{x} \in \mathbb{C}^{M \times M \times K}$, (normalized) dictionary $\mathbf{A} \in \mathbb{C}^{M \times N \times K}$, stopping criterion
 - 2: **Output:** Estimated signal CPSD tensor $\hat{\mathbf{R}}_\mathbf{s}$, residual CPSD tensor $\mathbf{R}_{res}$, support set Λ , iteration count p
 - 3: $p \leftarrow 0$
 - 4: $\mathbf{R}_{res} \leftarrow \mathbf{R}_\mathbf{x}$
 - 5: $\Lambda \leftarrow \emptyset$
 - 6: **while** stopping criterion not met **do**
 - 7: $p \leftarrow p + 1$
 - 8: **Matching I:**
 - 9: **for** $k = 1$ to K **do**
 - 10: Calculate correlations as per (4)
 - 11: **end for**
 - 12: **Matching II:** Summing up and identifying support as per (5)
 - 13: **Updating signal CPSD tensor:**
 - 14: **for** $k = 1$ to K **do**
 - 15: Calculate as per (6)
 - 16: **end for**
 - 17: **Updating residual CPSD tensor:**
 - 18: **for** $k = 1$ to K **do**
 - 19: Calculate as per (7), (8)
 - 20: **end for**
 - 21: **end while**
-

Signal energy update. Substituting $\mathcal{A}_\Lambda(k)$ for $\mathbf{A}(k)$ in (2) and minimizing the mean-square error with respect to $\mathbf{R}_\mathbf{s}(k)$ yields the closed-form estimate

$$\hat{\mathbf{R}}_\mathbf{s}^{(p)}(k) = \mathcal{A}_\Lambda^\dagger(k) \mathbf{R}_\mathbf{x}(k) \mathcal{A}_\Lambda^{\dagger H}(k), \quad (6)$$

where $\{\cdot\}^\dagger$ denotes the Moore-Penrose pseudoinverse.

Residual update. The projection matrix $\mathbf{P}_\Lambda(k) = \mathcal{A}_\Lambda(k) \mathcal{A}_\Lambda^\dagger(k)$ gives the residual

$$\hat{\mathbf{R}}_\mathbf{y}^{(p)}(k) = \mathbf{P}_\Lambda(k) \mathbf{R}_\mathbf{x}(k) \mathbf{P}_\Lambda^H(k), \quad (7)$$

$$\mathbf{R}_{res}^{(p)}(k) = \mathbf{R}_\mathbf{x}(k) - \hat{\mathbf{R}}_\mathbf{y}^{(p)}(k). \quad (8)$$

Iteration stops when the sparsity target (i.e. known number of sources in the scene) is reached. In general, any stopping criterion could be designed suited to a specific beamforming goal or acoustic scenario. Note that based on the above stopping criterion, $\hat{\mathbf{R}}_\mathbf{s}(k) \in \mathbb{C}^{d \times d}$. Algorithm 1 summarizes the full procedure.

4. SIMULATIONS AND DISCUSSION

The proposed method (PROP) is evaluated for HRTF estimation and binaural cue preservation against GEVD-based HRTF estimation [5] and noise CPSDM estimation with VAD [7] (referred to as BASE), using a behind-the-ear (BTE) setup of $M = 6$ microphones and $d = 2$ sources (Fig. 2). Each

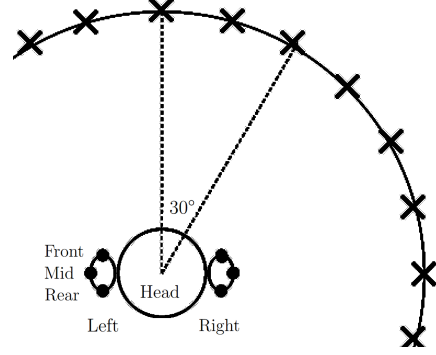


Fig. 2: Simulation setup: BTE microphone placement with two sources at 0° and 30° .

source consists of a 30 s sample formed by concatenating 3 s snippets from 10 randomly selected speakers from the EARS dataset [15], with each snippet drawn from a different speaker, convolved with HRIRs from a Brüel & Kjær 4128C dummy head from the OlHeadHRTF dataset [16] ($f_s = 48$ kHz). The HRTF dictionary covers -180° to 172.5° at 7.5° spacing and zero elevation. Microphone self-noise and diffuse noise at each channel are simulated through a zero-mean white Gaussian process. STFT and FFT processing use a 32 ms Hann window with 50% overlap over 2048 frequency bins; consecutive time blocks (i.e. collection of time frames) overlap by 50%. Both sources are placed on-grid with a source-sparsity-based stopping criterion.

HRTF estimation error is measured by the Hermitian angle (HA) [17]

$$\frac{\sum_t \sum_{k=1}^{K/2+1} \arccos \left(\frac{|\mathbf{a}_i^H(t,k) \mathbf{h}_i(t,k)|}{\|\mathbf{a}_i(t,k)\| \|\mathbf{h}_i(t,k)\|} \right)}{T(K/2+1)}, \quad (9)$$

where $\mathbf{a}_i(t,k)$ and $\mathbf{h}_i(t,k)$ are the true HRTF and estimated HRTF in the t^{th} time block and k^{th} frequency bin for the i^{th} source. The estimated HRTF and noise CPSDM outputs are then used as inputs to a JBLCMV beamformer [4], and binaural cue preservation is assessed via interaural level difference (ILD) and interaural phase difference (IPD) errors on the interaural transfer function (ITF) [18] computed over 3-8 kHz and 0-1.5 kHz respectively. Three parameters are varied across 30 trials each for HRTF estimation: SNR (-10 to 30 dB), time block size (TBS, 10-100 frames ~ 320 ms-3.2 s), and source azimuth configuration (fixing one source at 0° while varying the other's position). Binaural cue preservation is evaluated across the same SNR range.

HRTF estimation. Sources S_1 and S_2 are fixed at 0° and 30° respectively for Fig. 3. PROP demonstrates a consistent advantage over BASE across the full SNR range (Fig. 3a): even at -10 dB, PROP's HA for S_2 remains below BASE's high-SNR floor, underscoring the noise robustness afforded by the dictionary-based model. A persistent asymmetry be-

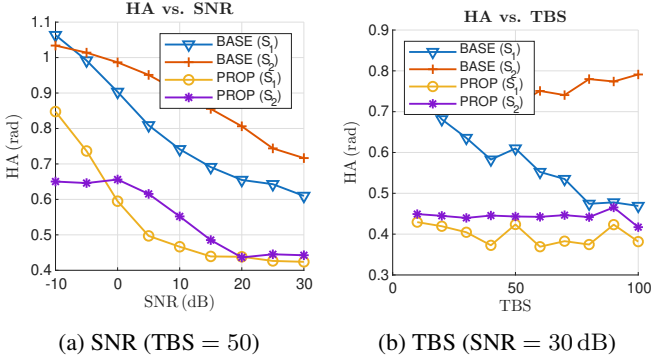


Fig. 3: HRTF estimation error under varying SNR and TBS.

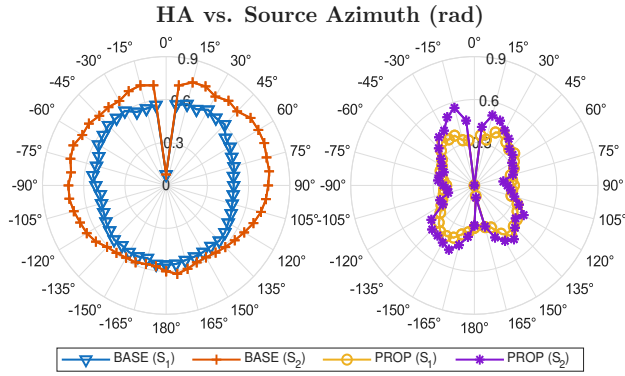


Fig. 4: HRTF estimation error vs. source azimuth (S_1 fixed at 0°, SNR = 30 dB, TBS = 50).

tween S_1 and S_2 reflects their geometric relationship: BASE struggles considerably more with the off-frontal S_2 , while PROP maintains comparably low HA for both sources. Increasing TBS (Fig. 3b) benefits BASE for S_1 but yields no meaningful improvement for S_2 , suggesting BASE’s error on the off-frontal source is model-limited rather than data-limited. PROP, by contrast, operates at consistently lower HA across all block sizes with only mild variation, implying it requires far fewer frames to achieve better accuracy, advantageous in low-latency or non-stationary settings. The polar plots (Fig. 4) confirm these trends spatially: BASE (left panel) produces near-uniform, high-HA contours for both sources regardless of azimuth, while PROP (right panel) yields a compact, low-HA contour for S_1 across most directions. PROP’s S_2 contour is notably irregular, with somewhat elevated errors near frontal and rear azimuths, revealing a geometric sensitivity when S_2 approaches collinearity or antipodality with S_1 .

Binaural cue preservation. S_1 (0°) serves as the target and S_2 (30°) as the interferer. For ILD errors (Fig. 5a), both methods perform poorly for S_1 : BASE shows modest improvement with SNR while PROP remains flat and performs worse at high SNR, indicating that HRTF estimation gains do not translate to target ILD preservation for either

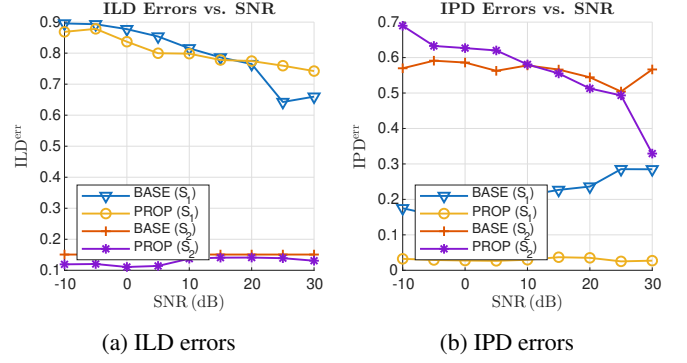


Fig. 5: Binaural cue preservation errors for S_1^t and S_1^i under varying SNR.

method. For S_2 , both methods achieve near-identical, low and stable errors across SNR: expected behavior from a well-designed JBLCMV beamformer, driven primarily by beamformer geometry rather than estimation quality. For IPD errors (Fig. 5b), PROP achieves remarkably low and stable errors for S_1 across the full SNR range, while BASE degrades counterintuitively at higher SNR, possibly reflecting overfitting to noise structure that partially regularized phase at low SNR. For S_2 , BASE maintains moderate but stable IPD errors, whereas PROP starts high and only improves substantially at high SNR, suggesting that tighter spatial constraints from the dictionary-based model introduce greater phase distortion on the interferer. From a perceptual standpoint, IPD preservation for S_1 is the most critical outcome, governing the listener’s ability to localize and segregate the intended speech source [19]; PROP’s near-zero target IPD error is therefore its most practically significant result, directly supporting spatial release from masking. The corresponding interferer IPD degradation is less consequential, as hearing-aid systems must suppress rather than localize interferers. The universal failure to preserve target ILD warrants further investigation, as ILD is the dominant high-frequency cue and its degradation may impair elevation perception and externalization.

5. CONCLUSIONS

We proposed a covariance-domain sparse pursuit framework that exploits a predetermined HRTF dictionary to jointly estimate the HRTF subspace and a residual noise CPSDM in a single pass. It achieves consistently lower HRTF estimation error than a GEVD-based baseline across SNR, time block size, and source configuration, with particular advantage in low-SNR and short-block regimes. When used to drive a JBLCMV beamformer, the proposed algorithm yields near-zero and stable target IPD errors across the full SNR range. Future work will address off-grid source scenarios, increased number of interferers and generalization to reverberant room acoustic conditions.

6. REFERENCES

- [1] World Health Organization, Ed., *World Report on Hearing*, World Health Organization, Geneva, 1st ed edition, 2021.
- [2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1 of *Springer Topics in Signal Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [3] E. Hadad, S. Doclo, and S. Gannot, “A generalized binaural MVDR beamformer with interferer relative transfer function preservation,” in *2016 24th European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, Aug. 2016, p. 1643–1647, IEEE.
- [4] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, “Improved multi-microphone noise reduction preserving binaural cues,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, Mar. 2016, p. 460–464, IEEE.
- [5] M. Taseska and E. A. P. Habets, “Relative transfer function estimation exploiting instantaneous signals and the signal subspace,” in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Nice, Aug. 2015, p. 404–408, IEEE.
- [6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
- [7] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, “Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.
- [8] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.
- [9] P. Hoang, Z.-H. Tan, J. M. De Haan, and J. Jensen, “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, June 2021, p. 6119–6123, IEEE.
- [10] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, J. Jensen, and M. Guo, “Binaural beamforming using pre-determined relative acoustic transfer functions,” in *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017, p. 1–5, IEEE.
- [11] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust Joint Estimation of Multimicrophone Signal Model Parameters,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1136–1150, July 2019.
- [12] C. Li and R. C. Hendriks, “Multimicrophone Signal Parameter Estimation in a Multi-Source Noisy Reverberant Scenario,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 678–692, 2025.
- [13] J. A. Tropp and A. C. Gilbert, “Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit,” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
- [14] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit,” *Signal Processing*, vol. 86, no. 3, pp. 572–588, Mar. 2006.
- [15] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, “EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation,” , no. arXiv:2406.06185, June 2024, arXiv:2406.06185 [eess].
- [16] F. Denk, S. M. A. Ernst, S. D. Ewert, and B. Kollmeier, “Adapting Hearing Devices to the Individual Ear Acoustics: Database and Target Response Correction Functions for Various Device Styles,” *Trends in Hearing*, vol. 22, pp. 1–19, Jan. 2018.
- [17] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation,” in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, 2017, p. 11–15, IEEE.
- [18] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Relaxed Binaural LCMV Beamforming,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 137–152, Jan. 2017.
- [19] F. L. Wightman and D. J. Kistler, “The dominant role of low-frequency interaural time differences in sound localization,” *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.