Microphone arrays have been used in many applications, such as: teleconferencing, speech recognition, talker characterization, speech enhancement, source localization and separation, etc. Despite the fast-paced development in microphone-array hardware, software and algorithms, there still exist numerous challenges for microphone-array processing. A few of these are: Real-time issues, efficient and effective algorithms to deal with a variety of environments and noisy conditions. This thesis is concerned with effective and efficient algorithms for the tasks of sound-source localization and separation, using a large-aperture microphone array in an adverse environment.

The phase transform (PHAT) has been shown experimentally to cope effectively with the reverberation noise, which is the main challenge in most environments. In this thesis, first, an analytic solution of the steered response power using the phase transform (SRP-PHAT) is derived to explain the robustness of the PHAT against reverberation. Next, three computationally efficient algorithms to locate a single sound source, which reduce the computational cost by more than three orders of magnitude compared to a brute-force grid search are presented. Third, an enhancement, pre-alignment, to any cross-correlation-based functional, such as the SRP-PHAT, in the task of locating multiple sources is proposed. Fourth, a new functional, minimum-variance distortionless reponse (MVDR) using the phase transform (MVDR-PHAT) augmented by the prealignment enhancement, is introduced, which will be shown to give improved performance than both the nominal SRP-PHAT and enhanced SRP-PHAT functionals in detecting and locating multiple sources. Fifth, two multiple-source localization algorithms using just a frame of data, Gaussian mixture models-
based (GMM) and region zeroing (RZ), are introduced. And finally, an application of the newly proposed enhanced MVDR-PHAT functional in the task of separating multiple sound sources is presented. All the work, except the validation of the analytic SRP-PHAT, was verified using real recordings by the Huge Microphone Array (HMA) system in a real, noisy room with a reverberation time $T_{60}$ of 450 ms and a lot of background noise generated from the computer fans, ventilation, and the HMA console itself.
Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array

by

Hoang T. Do

Master of Science, Brown University, 2007

Thesis
Submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
in the School of Engineering at Brown University,
Providence, Rhode Island

May 2011
This dissertation by Hoang T. Do is accepted in its present form by
the School of Engineering as satisfying the
dissertation requirement for the degree of
Doctor of Philosophy

Harvey F. Silverman, Advisor

Michael S. Brandstein, Reader

Ivan J. Tashev, Reader

Gabriel Taubin, Reader

Peter M. Weber
Dean of the Graduate School
The Vita of Hoang Tran Huy Do

Hoang Tran Huy Do was born on February 10, 1982 in Hanoi, Vietnam. Despite growing up in a family of architects and winning the first prize of the Hanoi art contest for elementary school students, his interest in science prevailed. He attended College of Staten Island - The City University of New York (CSI-CUNY) from 2000 to 2004, graduating with the degree of Bachelor of Science in Engineering with honors, Magna Cum Laude.

In the fall of 2005, he attended Brown University to pursue a Ph.D. degree in electrical engineering. He received a Master’s degree, with a thesis on real-time single sound-source localization in May 2007. His research interests include microphone arrays, audio signal processing and speech processing. In his free time, he enjoys playing the “real” football, i.e., soccer, watching American football or practicing his Iron-Chef cooking. He is a member of the IEEE and ISCA.
Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Harvey F. Silverman, for his continuous support and patience, invaluable advice and guidance throughout my studies and research at Brown University. My sincere thanks to the readers of my Ph.D. dissertation: Dr. Michael Brandstein, Dr. Ivan Tashev and Professor Gabriel Taubin, not only for their time spent on my work but also for the continuous encouragement, generous help and constructive comments.

I would like to thank Kenneth Silverman for his help and creation of softwares that are very useful to my work. My thanks also go to Arpie Kaloustian for his assistance on hardware and maintenance. My former and current colleagues, Ying Yu, Matthew Gillette, Avram Levi, Brian Reggiannini, and Xiaoxue Li are much appreciated for their help, sharing and interesting talks.

Last but not least, from the bottom of my heart, I thank my Grandma, Mom and Dad, my sister and my wife, Trang, for their unconditional love, endless support and encouragement. This work is dedicated to all of you.
# Contents

1 Introduction .................................................. 1
   1.1 Prologue .................................................. 1
   1.2 Review of current work .................................. 3
   1.3 Experimental philosophy and conditions ................. 5
   1.4 Contributions of the thesis ................................ 6

2 The steered response power using the phase transform (SRP-PHAT) 7
   2.1 A simplified acoustic model ............................... 8
   2.2 Analytic solution of SRP-PHAT based on the Acoustic Model .......... 12
   2.3 Why SRP-PHAT is robust? .................................. 18
   2.4 On PHAT robustness validation ................................ 24

3 Computationally efficient SRP-PHAT methods for locating a single source 29
   3.1 Problem statement ......................................... 29
   3.2 Stochastic region contraction (SRC) ........................ 30
      3.2.1 Overview .............................................. 30
      3.2.2 SRC algorithm ......................................... 31
   3.3 Coarse-to-fine region contraction (CFRC) .................. 35
   3.4 Stochastic particle filtering (SPF) ........................ 37
3.5 Comparison of computational cost and accuracy for single-source, SRP-PHAT localization methods ............................................. 41
3.6 Conclusion .................................................................................. 42

4 Determining locations of multiple sources and proposed enhancements

4.1 Increased complexity when multiple sources are considered .......... 44
4.2 Enhancement for SRP-PHAT functional: Prealignment ................. 45
   4.2.1 Preliminary experiment to justify the use of prealignment ... 47
   4.2.2 Prealignment ................................................................. 53
      4.2.2.1 Brute-force method ................................................. 53
      4.2.2.2 A more efficient prealignment method .................... 56
   4.2.3 Computational cost of prealignment methods ..................... 58
      4.2.3.1 Cost of brute-force method .................................... 59
      4.2.3.2 Cost of proposed, more efficient method ................. 60
      4.2.3.3 Comparison ............................................................ 60

5 New functional for multiple-source localization: The enhanced minimum variance distortionless response using the phase transform (MVDR-PHAT) 63

5.1 The traditional MVDR ............................................................ 64
   5.1.1 Formulation ................................................................... 64
5.2 The steered enhanced MVDR-PHAT response power ................. 68

6 Methods for multiple-source localization using the proposed functionals 71

6.1 Gaussian Mixture Model (GMM) method ................................. 73
6.2 Region Zeroing (RZ) method ................................................... 75
9 Conclusions and Future work

9.1 Conclusions .................................................. 108
9.2 Future work ................................................... 110
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Performance of analytic SRP-PHAT for different values of $T_{60}$</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Probabilistic relation to determine number of fe’s required</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>Performance and cost of SRP-PHAT using a full grid search over all speech frames; SPF, SRC, and CFRC over good frames for 5 different locations.</td>
<td>42</td>
</tr>
<tr>
<td>4.1</td>
<td>Functions and their computational costs</td>
<td>58</td>
</tr>
<tr>
<td>7.1</td>
<td>Multiple source localization performance of the three functionals</td>
<td>88</td>
</tr>
<tr>
<td>8.1</td>
<td>Objective measures of the performance of source separation algorithms for source 1. Highest value in each column in bold type</td>
<td>104</td>
</tr>
<tr>
<td>8.2</td>
<td>Objective measures of the performance of source separation algorithms for source 2. Highest value in each column in bold type</td>
<td>105</td>
</tr>
</tbody>
</table>
## List of Figures

2.1  The source (red star) and 24 microphones (blue dots) in a simulated room. The arrow indicates orientation of the source. .................................................. 25
2.2  Side-views of analytic SRP-PHAT surface for $T_{60}=25$ ms .......................... 26
2.3  Side-views of analytic SRP-PHAT surface for $T_{60}=50$ ms .......................... 26
2.4  Side-views of analytic SRP-PHAT surface for $T_{60}=200$ ms ......................... 27
2.5  Side-views of analytic SRP-PHAT surface for $T_{60}=540$ ms ....................... 27
2.6  Side-views of analytic SRP-PHAT surface for $T_{60}=990$ ms ....................... 28

3.1  2D example of SRC ................................................................. 31
3.2  Performance of SRC-I as a function of parameter $N$ for four different source locations ................................................................. 34
3.3  2D example of CFRC ............................................................... 35
3.4  Performance of CFRC as a function of parameter $N$ for four different source locations ................................................................. 37
3.5  *Top view of the array, showing source locations and panels. This experiment used 24 microphones of the 128 on panels H, I, J, K. The arrows indicate the orientation of the talker and the SNR’s are for background noise only.* ................................................................. 42

4.1  (a) the functional surface from a 2-stage LEMSAlg and (b) SRP-PHAT surface .................................................................................................................. 46
4.2 Mismatch of temporal events between microphone signals $z_u$ and $z_v$.

4.3 3-D view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame.

4.4 Z-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 51.2 ms.

4.5 X-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 51.2 ms.

4.6 X-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 25.6 ms.

4.7 Z-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 25.6 ms.

4.8 X-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 102.4 ms.

4.9 Z-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 102.4 ms.

4.10 X-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 204.8 ms.

4.11 Z-dim. side view of a slice of the nominal SRP-PHAT using a fixed-time frame and the enhanced SRP-PHAT using a prealigned frame of length 204.8 ms.
4.12 Percent cost saving vs. the number of microphones, M ................. 61
4.13 Percent cost saving vs. the number of microphones, M, for a small-
aperture array ............................................................... 62
6.1 SRP-PHAT 3D illustration for 5 talkers ................................. 73
6.2 5 talkers: (a) 2-D SRP-PHAT energy map and (b) GMM peaks. .... 75
7.1 Top view of the room with 10 talkers and 182 HMA microphones ... 79
7.2 A snapshot of hand-labeled speech activity of talker 1 ............... 80
7.3 Example of labeling speech activity for transitional frames ........ 82
7.4 Top views of (a) the ground-truth talkers and 182 microphones and the
surfaces of (b) the nominal SRP-PHAT functional, (c) the enhanced
SRP-PHAT functional, and (d) the enhanced MVDR-PHAT functional 84
7.5 3-D views of three different functional surfaces for a frame of 6 simul-
taneous, active talkers ...................................................... 86
7.6 Percent correct estimates of the three functionals for \( n \)-talker frames,
\( n = 2, \ldots, 6 \) ................................................................. 90
7.7 Percent missed estimates of the three functionals for \( n \)-talker frames,
\( n = 2, \ldots, 6 \) ................................................................. 90
7.8 Percent extra estimates of the three functionals for \( n \)-talker frames,
\( n = 2, \ldots, 6 \) ................................................................. 91
8.1 Flow-chart of the proposed source-separation algorithm .......... 97
8.2 The two sources (red dots) and 16 microphones (blue dots): (a) 3-D
view and (b) top view ......................................................... 98
8.3 Spectrograms of (a) mixed signal, (b) clean-speech signal of source 1,
(c) baseline 1 (DSBF), and (d) baseline 2 (MVDR) for source 1 .... 100
8.4 Spectrograms of (a) mixed signal, (b) clean-speech signal of source 1, (c) baseline 3 (CASA using SRP-PHAT), and (d) the proposed algorithm for source 1 ................................................................. 101

8.5 Spectrograms of (a) mixed signal, (b) clean-speech signal of source 2, (c) baseline 1 (DSBF), and (d) baseline 2 (MVDR) for source 2 . . . 102

8.6 Spectrograms of (a) mixed signal, (b) clean-speech signal of source 2, (c) baseline 3 (CASA using SRP-PHAT), and (d) the proposed algorithm for source 2 ................................................................. 103

8.7 SNRloss scores of the three baselines and proposed algorithm as a function of the number of microphone used for source 1. Note: Higher score means better intelligibility/separation ......................... 106

8.8 SNRloss scores of the three baselines and proposed algorithm as a function of the number of microphone used for source 2 ............... 107
Chapter 1

Introduction

1.1 Prologue

Microphone arrays have been used in many applications, such as teleconferencing [19, 39, 51, 52, 109], speech recognition [1, 18, 41, 48, 91, 89], talker characterization [100], source separation [83, 17, 56], voice capture in reverberant environments [38, 49, 114], sound-source localization [9, 55, 26, 14, 29], speaker recognition [59, 65, 77, 31], etc. Many array designs have been implemented and studied, including spherical arrays [35], superdirectivity arrays [20], linear arrays [93], etc. On these arrays, work has been done on the construction and designs of hardware as well as the implementation of real-time applications.

As a result of the rapid development of inexpensive DSP’s and general-purpose microprocessors, microphone arrays have been commercialized and implemented as an everyday product, such as the unit by Acoustic Magic [63] using the technology in [11, 10]. They have also been incorporated using only a few microphones in some everyday products such as: cellular phones and bluetooth headsets for hand-free operations, automobile speech enhancement and noise cancelation for audio communication, etc. In addition, large companies such as Polycom and Microsoft have applied microphone-
array technology in their systems for videoconferencing systems [109, 115], in-car infotainment systems [102], and the recently released Kinect device [111].

While many products utilize small-aperture microphone arrays due to the cost of hardware, simplicity of software, and portability of the devices, such arrays are often ineffective in dealing with remote sources in adverse environments, where background and reverberation noise are high. Also, smaller arrays can achieve only a certain level of performance due to the lack of aperture, for example, circular-arrays [105], triangular-arrays [4] or round-table microphone-arrays [115] often only yield directions-of-arrival (DOA’s) from the sources instead of pin-pointing their 3-D locations. Hence, the enhancement of speech signals by means of beamforming from such arrays is also limited. On the other hand, as the cost of hardware rapidly decreases with processing power per dollar increasing everyday, the choice of building and using large-aperture microphone-arrays becomes more practical. Such arrays would be able to yield robust and effective performance in adverse environments for a variety of tasks: 3-D sound-source localization, separation of sources, speech enhancement, noise reduction and dereverberation, etc.

Despite the fast-paced development in microphone-array hardware, software and algorithms, there still exist numerous challenges for microphone-array processing. A few of these are: Real-time issues, optimal microphone-array configuration or microphone placements, efficient and effective algorithms, compatibility to a variety of environments, etc. This thesis is concerned with effective and efficient algorithms for the tasks of sound-source localization and separation, using a large-aperture microphone array in an adverse environment.
1.2 Review of current work

In most of speech-based communication and human/machine interaction systems, it is very useful, if not essential, to have reliable estimates of sound-source location(s). In general, one classifies source-localization algorithms into two types: 1) two-stage, time-difference of arrival (TDOA)-based approaches and 2) one-stage, steered beamformer-based (i.e., functional steered response) approaches.

A two-stage algorithm processes in two steps. In the first step, or the time-delay estimation (TDE) step, the goal is to produce pairwise time-difference of arrivals (TDOA’s) of speech sounds between pairs of microphones. This is usually done using some type of cross-correlation-based technique, such as the generalized cross-correlation with the phase-transform (GCC-PHAT) [53, 101], or from a blind-source separation procedure [14, 70]. In the second step, the source locations are estimated from the TDOA’s using some optimization techniques. Many methods to solve the second step have been proposed, such as maximum likelihood estimation, the least-square error method, a linear intersection method, spherical interpolation, etc. [22, 99, 86, 85, 97]. However, the pairwise technique often suffers considerable degradation from acoustic reverberation or needs some smoothing over multiple frames to get a reliable estimate. DiBiase [23] and Birchfield [8] indicated these two-stage methods usually have biases. Moreover, for multiple-source localization, the TDOA-based, two-stage algorithms usually have to go through the blind-source separation (BSS) procedures [70, 14, 84], which either require an accumulation of speech data for a few seconds [70], or an extra processing to solve the permutation problem inherent to BSS. This procedure becomes more complex as the number of sources increases.

A one-stage, steered-beamformer-based algorithm reverses the problem of beamforming. It hypothesizes a spatial, point-source location in a predefined region and computes a beamforming functional at that location, i.e., a beamformer acoustically steers to different spatial locations, and the one(s) yielding the maximum values of
the functional are the estimated source locations [26, 115, 29]. The one-stage method exploits the multitude of microphones in order to overcome the complication of severe reverberation and background noise. The method avoids the process of making early, and often incorrect, decisions of selecting “good” TDOA’s as done in the two-stage methods. Statistically, it follows the principle of least commitment. However, the robust performance of this approach and its bias-free property [23, 8] comes at the price of high computational cost. The values of the beamformer’s functional have a lot of local maxima. Hence, traditionally brute-force methods are required to find any “global” maxima, i.e, the true source locations in three dimensions.

In general, both approaches have their own pros and cons. Depending upon the application, conditions and the amount of resources, one should be able to choose a suitable solution.

In this thesis, two different source localization problems are addressed: 1) computationally efficient single source localization and 2) locating multiple sources under adverse conditions using robust functionals.

Another problem that the thesis addresses is to separate multiple sources, i.e., extracting speech signals of individual sources from a mixture, in an adverse environment. This turns out to be closely related to the multiple-source localization problem, since it is most likely that an effective mechanism which is capable of localizing the sound sources could also be applied in the separation of signals coming from the sources. Traditionally, there are three major approaches for source separation tasks:

- Blind-source separation (BSS) [64]
- Beamforming (BF)[78]
- Computational Auditory Scene Analysis (CASA)[108]

The three approaches will be discussed in more detail in chapter 8. There are also hybrid algorithms using more than of the three above, such as: blind-source separation
combined with beamforming [82, 107], blind-source separation combined with CASA [81, 87], and CASA combined with beamforming [34, 56]. Recently, in [56], the authors proposed to use the steered response power (SRP) with the phase transform (SRP-PHAT) functional, or essentially, the delay-and-sum beamformer using the phase-transform, incorporated with the CASA procedure to appropriately extract the time-frequency (TF) bins of the desired source. Essentially, this method does the separation in the temporal-frequency domain using the SRP-PHAT functional as the measure. In this thesis, a more spatially separable beamformer, i.e., a null-steering beamformer, is combined with the CASA procedure in the same manner as in [56], to provide an extra dimension of separation in the spatial domain. The proposed algorithm will be compared to standard beamforming techniques and the recently proposed algorithm in [56].

1.3 Experimental philosophy and conditions

The work throughout this thesis has always been based on the following principles:

- Tackle new and challenging problems
- Use real data and real room conditions for evaluation
- Aim at practical solutions for real systems

Many algorithms, which perform nicely in simulation, fail when tested with real data in a real environment. This is because, in reality, no model is perfect. Hence, the use of real data from real environments is essential in testing the true practicality and value of an algorithm and/or a system.

In the field of sound and acoustics, noise and reverberation are the real challenges, and they often behave wildly in real rooms. This is clearly evident in our experimental room with a variety of furniture; reflective surfaces (vinyl tiled floor, painted plaster
walls, glass-door cabinets, etc.); a standing wave prone, vaulted ceiling; loud fan noises from the computers, ventilation, and the microphone array’s console. This environment is very challenging for any microphone system to work in efficiently. Also, recordings made with real human sources often exhibit extra difficulties, such as: uncertainty in the actual source locations as a result of mouth, head, and body movements, uncontrolled loudness and softness, comfort distances among different talkers for the multiple talker case, etc.

In this thesis, all experiments were carried out in this acoustically challenging environment and real human talkers were used in the important experiments.

1.4 Contributions of the thesis

There are six major contributions presented in the thesis. First, an analytic solution of SRP-PHAT is derived from a simple acoustic model, which is useful in explaining the robustness of the PHAT against reverberation. Second, three new techniques, which reduce the computational cost of SRP-PHAT methods by more than three orders of magnitude compared to using the brute-force search in locating a single sound-source are presented. Third, a novel enhancement technique, prealignment, to augment any cross-correlation-based functional in the task of detecting and locating multiple sources is proposed. Fourth, a new robust functional, enhanced MVDR-PHAT, is introduced. Fifth is the proposal of two algorithms, Gaussian mixture models-based (GMM) and region zeroing (RZ), using the newly developed functionals, enhanced SRP-PHAT and MVDR-PHAT, to detect and locate multiple sound-sources using just a single frame (100 ms) of the large-aperture microphone-array data. Finally, an application of the proposed MVDR-PHAT functional for sound-source separation is presented.
Chapter 2

The steered response power using the phase transform (SRP-PHAT)

The phase transform (PHAT), discussed early by Knapp and Carter [53], has been widely used in signal processing tasks, such as: source localization and tracking [101, 97, 30, 26, 104], speech enhancement and spectral estimation [113, 27], voice-activity detection [117], calibration of wireless sensor networks [76], and source separation [56], etc. Its popularity is largely due to its simple implementation and robustness against reverberation. The superior performance of the PHAT against reverberation has been shown experimentally in [23, 8, 10]. However, very little has been done to formally explain this robustness. In [32], the authors studied the performance of the PHAT and its variant, $\beta$-PHAT, for different bandwidths of the signals. The evaluations were done experimentally without an analytic explanation. The most recent work on explaining the performance of PHAT was done in [116]. The authors developed the relationship between the PHAT and the Maximum Likelihood (ML) algorithm, and showed that the PHAT remained optimal in the ML sense under heavy reverberation. This explanation, however, depends upon the validity of the ML’s optimality.

In this thesis, a different approach is taken. I start from the signal model and
derive an analytic solution of the PHAT in the framework of the steered-response power (SRP-PHAT) under some environmental assumptions. This derivation has twofold benefits. First, it gives us a new, analytic formulation for the SRP-PHAT. Second, it helps us explain the robustness of the PHAT against reverberation.

2.1 A simplified acoustic model

Most microphone-array applications are done in acoustic enclosures, such as rooms, hallways, and vehicles. Those enclosures often include sound-reflecting surfaces, such as walls, furniture, and windows. In the presence of such surfaces, the sound waves produced by a single sound source propagate along multiple acoustic paths, the non-direct paths having been reflected one or more times (reverberation). Under heavy reverberation, the performance of speech-based applications often degrade drastically [12, 89]. Acoustic enclosures have been modeled extensively as linear systems [118, 54]. In such systems, the multi-path (one direct and multiple reflection paths) propagation over time \( t \) of the sound wave originating from the source location \( \vec{l}_s \) to a particular location \( \vec{l} \) in the enclosure is characterized by the room impulse response (RIR) \( h(\vec{l}, \vec{l}_s, t) \). Note that for my purposes, the RIR of microphone channel \( u \), \( h(\vec{l}_u, \vec{l}_s, t) \), implicitly includes the response \( \gamma_u(\vec{l}_u, t) \), characterizing the electrical, mechanical and acoustical properties of the microphone itself, the response \( \alpha_u(\vec{l}_u, \vec{l}_s, t) \) of the propagation channel from the source to microphone \( u \), as well as the source’s response \( \gamma_s(\phi, \theta, t) \), where \( \phi, \theta \) are angles that parameterizing the spatial directivity pattern of the source and particularly depend on \( \vec{l}_u, \vec{l}_s \):

\[
h(\vec{l}_u, \vec{l}_s, t) = \gamma_u(\vec{l}_s) \ast \alpha_u(\vec{l}_u, \vec{l}_s, t) \ast \gamma_s(\phi, \theta, t) \quad (2.1)
\]

For this development the following conditions are obeyed:

- Ignore any shaping due to the source and microphone models as well as from
frequency dependent reflection in the room. These critical assumptions imply an RIR consisting of $K + 1$ impulses, one for the direct path and the other $K$ impulses characterizing the significant, perfect reflections for a source in a room.

- Assume all reflections have no frequency dependence
- Assume reflections after some finite number $K$ are insignificant
- There is no background noise in the room

This last assumption is made as I wish to analyze the effects of reverberation noise only. The signal received at microphone $u$, ($u = 1, \ldots, M$), in the time-domain is thus:

$$x_u(t) = A_u^{(D)} s(t - \frac{d_u}{C}) + A_u^{(R_1)} s(t - \frac{r_{u1}}{C}) + A_u^{(R_2)} s(t - \frac{r_{u2}}{C}) + \ldots + A_u^{(R_K)} s(t - \frac{r_{uK}}{C})$$

(2.2)

where:

- $d_u$ and $r_{uk}$ denote the propagation distances of the direct-path and the $k^{th}$ reverberant path, respectively.
- $C$ is the speed of sound (assumed constant in the room).
- $A_u^{(D)}$ and $A_u^{(R_k)}$ model the inverse square law attenuation of the source signal $s(t)$ at microphone $u$ due only to the propagation length along the direct-path and the reverberant paths respectively.

These attenuation factors are proportional to the inverse distances of the direct path and reverberant paths:

$$A_u^{(D)} = \frac{G_u}{d_u}$$

$$A_u^{(R_k)} = \frac{G_u}{r_{uk}}$$

(2.3)
where $G_u$ is a constant depending upon the specific microphone processing channel.

For simplicity and no loss in generality, for this development, I assume the microphone channels are similar and let $G_1 = ... = G_u = ... = G_M = D \equiv \min_{1 \leq m \leq M} d_m$ (which makes all values of $A_u^{(D)}, A_u^{(R_k)}$ in the range $[0,1]$), to make the direct microphone signal always attenuated.

Therefore, the time-signal at microphone $u$ can be modeled as,

$$x_u(t) = D d_u s(t - \frac{d_u}{C}) + \frac{D}{r_{u_1}} s(t - \frac{r_{u_1}}{C}) + \ldots + \frac{D}{r_{u_K}} s(t - \frac{r_{u_K}}{C})$$

$$= \frac{D}{d_u} s(t - \frac{d_u}{C}) + \sum_{k=1}^{K} \frac{D}{r_{u_k}} s(t - \frac{r_{u_k}}{C}) \quad (2.4)$$

In the time signals, I sample for no aliasing where the highest frequency containing important speech information is $f_h$, and the sampling frequency, $f_s$, is $f_s > 2f_h$.

Evaluating $x_u$ at discrete points $t = nT$,

$$x_u(nT) = \frac{D}{d_u} s(nT - \frac{d_u}{C}) + \sum_{k=1}^{K} \frac{D}{r_{u_k}} s(nT - \frac{r_{u_k}}{C})$$

$$= \frac{D}{d_u} s \left( (n - \frac{d_u}{CT})T \right) + \sum_{k=1}^{K} \frac{D}{r_{u_k}} s \left( (n - \frac{r_{u_k}}{CT})T \right) \quad (2.5)$$

As sampling is done properly, $\frac{d_u}{CT}$ is a fairly large positive number, so one can assume $s(t)$ is sufficiently smooth to allow,

$$n - \frac{d_u}{CT} \approx n - \left\lfloor \frac{d_u}{CT} \right\rfloor$$

$$n - \frac{r_{u_k}}{CT} \approx n - \left\lfloor \frac{r_{u_k}}{CT} \right\rfloor \quad (2.6)$$

where $\lfloor . \rfloor$ is the rounding operator. Thus I closely approximate $x_u(nT)$ as,

$$x_u(nT) \approx \frac{D}{d_u} s((n - \left\lfloor \frac{d_u}{CT} \right\rfloor)T) + \sum_{k=1}^{K} \frac{D}{r_{u_k}} s((n - \left\lfloor \frac{r_{u_k}}{CT} \right\rfloor)T) \quad (2.7)$$
Defining my sampling by,

\[ x_u[n] \equiv Tx_u(nT), \]
\[ s_u[n] \equiv Ts_u(nT), \] (2.8)

I obtain the discrete time-signal \( x_u[n] \):

\[ x_u[n] = \frac{D}{d_u} s[n - \lfloor \frac{d_u}{CT} \rfloor] + \sum_{k=1}^{K} \frac{D}{r_{uk}} s[n - \lfloor \frac{r_{uk}}{CT} \rfloor] \] (2.9)

From Eq. 2.6, 2.9, and using the DTFT’s shift-in-time property, I have:

\[ X_u(e^{j\omega T}) \approx \sum_{n=-\infty}^{\infty} \frac{D}{d_u} s[n - \lfloor \frac{d_u}{CT} \rfloor] e^{-j\omega nT} + \sum_{n=-\infty}^{\infty} \sum_{k=1}^{K} \frac{D}{r_{uk}} s[n - \lfloor \frac{r_{uk}}{CT} \rfloor] e^{-j\omega nT} \]

\[ \approx \frac{D}{d_u} S(e^{j\omega T}) e^{-j\omega \lfloor \frac{d_u}{CT} \rfloor T} + \sum_{k=1}^{K} \frac{D}{r_{uk}} S(e^{j\omega T}) e^{-j\omega \lfloor \frac{r_{uk}}{CT} \rfloor T} \]

\[ \approx S(e^{j\omega T}) \left( \frac{D}{d_u} e^{-j\omega \lfloor \frac{d_u}{CT} \rfloor T} + \sum_{k=1}^{K} \frac{D}{r_{uk}} e^{-j\omega \lfloor \frac{r_{uk}}{CT} \rfloor T} \right) \] (2.10)

To simplify notation, I define the travel time (in samples) for the direct path and the \( k^{th} \) \( (k = 1, 2, \ldots, K) \) reverberant path of microphone \( u \) as,

\[ q_{u0} \equiv \lfloor \frac{d_u}{CT} \rfloor \]
\[ q_{uk} \equiv \lfloor \frac{r_{uk}}{CT} \rfloor \] (2.11)

which yields,

\[ X_u(e^{j\omega T}) = S(e^{j\omega T}) \left( \frac{D}{d_u} e^{-j\omega q_{u0} T} + \sum_{k=1}^{K} \frac{D}{r_{uk}} e^{-j\omega q_{uk} T} \right) \] (2.12)
As the frequency-domain expressions for the signals received at the microphones have been established, a generalized cross-correlation using the phase transform (GCC-PHAT) between a pair of microphone signals can be derived, and thus, the steered response power using the phase transform (SRP-PHAT) for all pairs of microphones.

2.2 Analytic solution of SRP-PHAT based on the Acoustic Model

To calculate the GCC-PHAT, first I compute the cross-power spectrum of microphone $u$ (K significant reflections) and microphone $v$ (L significant reflections):

$$C_{uv}(e^{j\omega T}) = X_u(e^{j\omega T})X_v^*(e^{j\omega T})$$

(2.13)

where $*$ denotes the complex conjugate operator. Substituting Eq. 2.12,

$$C_{uv}(e^{j\omega T}) = |S(e^{j\omega T})|^2 \left( \frac{D^2}{d_u d_v} e^{j\omega (q_{v0} - q_{u0}) T} + \sum_{l=1}^L \frac{D^2}{d_u r_{vl}} e^{j\omega (q_{vl} - q_{u0}) T} + \sum_{k=1}^K \frac{D^2}{r_{uk} d_v} e^{j\omega (q_{v0} - q_{uk}) T} + \sum_{k=1}^K \sum_{l=1}^L \frac{D^2}{r_{uk} r_{vl}} e^{j\omega (q_{vl} - q_{uk}) T} \right)$$

$$\equiv |S(e^{j\omega T})|^2 M_{uv}(e^{j\omega T})$$

(2.14)
where $M_{uv}(e^{j\omega T})$ is,

$$M_{uv}(e^{j\omega T}) \equiv \frac{D^2}{d_ud_v} e^{j\omega(q_v0 - q_u0)T} + \left( \sum_{l=1}^{L} \frac{D^2}{d_u r_{vl}} e^{j\omega(q_vl - q_u0)T} \right) + \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{D^2}{r_{uk}r_{vl}} e^{j\omega(q_vl - q_{uk})T}$$

$$\equiv \frac{D^2}{d_ud_v} e^{j\omega(q_v0 - q_u0)T} + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk}b_{vl}} e^{j\omega(q_vl - q_{uk})T}$$

$$= \text{Direct-term} + \text{Cross-terms} \quad (2.15)$$

Note: for Eq. 2.15, $k$ and $l$ are not equal to 0 at the same time, $a_{u0} = d_u$, $b_{v0} = d_v$; $a_{uk} = r_{uk}$, $k = 1, ..., K$, and $b_{vl} = r_{vl}$, $l = 1, ..., L$. I have separated the direct-path term from all the reverberant cross-terms.

The phase transform (PHAT) in the frequency domain for microphone pair $(u, v)$ and frequency $\omega$ is achieved by severely “whitening” the cross-power spectrum, i.e., removing its magnitude:

$$P_{uv}(e^{j\omega T}) \equiv \frac{C_{uv}(e^{j\omega T})}{|C_{uv}(e^{j\omega T})|}$$

$$= \frac{|S(e^{j\omega T})|^{2}M_{uv}(e^{j\omega T})}{|S(e^{j\omega T})|^{2}|M_{uv}(e^{j\omega T})|}$$

$$= \frac{M_{uv}(e^{j\omega T})}{|M_{uv}(e^{j\omega T})|} \quad (2.16)$$

$M_{uv}(e^{j\omega T})$ can be expressed as a complex number in rectangular form as:

$$M_{uv}(e^{j\omega T}) \equiv H_{uv}(\omega) + jE_{uv}(\omega), \quad (2.17)$$

where using Euler’s formula on each exponential in Eq. 2.15, $H_{uv}(\omega)$ includes only
the cosine terms, and \( E_{uv}(\omega) \) includes only the sine terms:

\[
H_{uv}(\omega) = \frac{D^2}{d_u d_v} \cos(\omega(q_{v0} - q_{u0})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \cos(\omega(q_{vl} - q_{uk})T) \tag{2.18}
\]

\[
E_{uv}(\omega) = \frac{D^2}{d_u d_v} \sin(\omega(q_{v0} - q_{u0})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \sin(\omega(q_{vl} - q_{uk})T) \tag{2.19}
\]

Hence, the PHAT value computed in Eq. 2.16 is,

\[
P_{uv}(e^{j\omega T}) = \frac{H_{uv}(\omega) + jE_{uv}(\omega)}{|H_{uv}(\omega) + jE_{uv}(\omega)|} \arctan\left(\frac{E_{uv}(\omega)}{H_{uv}(\omega)}\right)\]

\[
e^{j\arctan\left(\frac{E_{uv}(\omega)}{H_{uv}(\omega)}\right)} \tag{2.20}
\]

The PHAT value in the discrete-time domain at time-sample \( q \) can be derived by taking the inverse discrete-time Fourier Transform (IDTFT) of the band-limited PHAT value in the continuous frequency-domain computed in Eq. 2.20 (here over the period \([0, \frac{2\pi}{T}]\)):

\[
R_{uv}[q] = \lim_{N \to \infty} \frac{T}{2\pi} \sum_{i=0}^{N-1} e^{j\arctan\left(\frac{E_{uv}(\omega_i)}{H_{uv}(\omega_i)}\right)} e^{j\omega_i q T} d\omega \tag{2.21}
\]

Define \( \omega_i \equiv \frac{2\pi i}{NT} \) where \( N \) is a constant, \( i \) is an integer in \([0, ..., N - 1]\). I can evaluate the integral by using a Riemann sum over the range \([0, \frac{2\pi}{T}]\):

\[
R_{uv}[q] = \lim_{N \to \infty} \frac{T}{2\pi} \sum_{i=0}^{N-1} \left( e^{j\arctan\left(\frac{E_{uv}(\omega_i)}{H_{uv}(\omega_i)}\right)} e^{j\omega_i q T} \right) \left( \frac{2\pi}{NT} \right) \tag{2.22}
\]
For a sufficiently large value of $N$, one can approximate $R_{uv}[q]$ as,

$$R_{uv}[q] \approx \frac{1}{N} \sum_{i=0}^{N-1} e^{j \left( \arctan \left( \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \right) + \omega_i q T \right)}$$

$$= \frac{1}{N} \sum_{i=0}^{N-1} e^{j \Psi_{uv}[q, \omega_i]} \quad (2.23)$$

where I define,

$$\Psi_{uv}[q, \omega_i] \equiv \arctan \left( \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \right) + \omega_i q T \quad (2.24)$$

Note that there is a trigonometric identity:

$$\arctan(x) + \arctan(y) = \arctan \left( \frac{x + y}{1 - xy} \right) \quad (2.25)$$

Using Eq. 2.25 I have:

$$\Psi_{uv}[q, \omega_i] = \arctan \left( \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \right) + \omega_i q T$$

$$= \arctan \left( \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \right) + \arctan \left( \tan(\omega_i q T) \right)$$

$$= \arctan \left( \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \tan(\omega_i q T) \right) \frac{1 - \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \tan(\omega_i q T)}{1 - \frac{E_{uv}[\omega_i]}{H_{uv}[\omega_i]} \tan(\omega_i q T)}$$

$$= \arctan \left( \cos(\omega_i q T)E_{uv}[\omega_i] + \sin(\omega_i q T)H_{uv}[\omega_i] \right) \cos(\omega_i q T)H_{uv}[\omega_i] - \sin(\omega_i q T)E_{uv}[\omega_i]$$

$$\equiv \arctan (Q_{uv}[q, \omega_i]) \quad (2.26)$$

In practice, due to starting with real data, the PHAT values are real. Hence, from
Eq. 2.23, I have:

\[ G_{uv}[q] \equiv \text{Re}(R_{uv}[q]) = \text{Re} \left( \frac{1}{N} \sum_{i=0}^{N-1} e^{j \Psi_{uv}[q, \omega_i]} \right) \]

\[ = \frac{1}{N} \sum_{i=0}^{N-1} \cos(\Psi_{uv}[q, \omega_i]) \]  

(2.27)

Using the trigonometric identity:

\[ \cos(\arctan(x)) = \frac{1}{\sqrt{1 + x^2}}, \]  

(2.28)

with Eq. 2.26, Eq. 2.27, I have:

\[ G_{uv}[q] = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{\sqrt{1 + Q_{uv}^2[q, \omega_i]}} \]  

(2.29)

\( Q_{uv}[q, \omega_i] \) of Eq. 2.26 into Eq. 2.29 yields the expression for the real PHAT value of microphone pair \((u, v)\) at time sample \(q\) as follows:

\[ G_{uv}[q] \]

\[ = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{\sqrt{1 + \left( \frac{\cos(\omega_i q T) E_{uv}[\omega_i] + \sin(\omega_i q T) H_{uv}[\omega_i]}{\cos(\omega_i q T) H_{uv}[\omega_i] - \sin(\omega_i q T) E_{uv}[\omega_i]} \right)^2}} \]

\[ = \frac{1}{N} \sum_{i=0}^{N-1} \left( \frac{\cos(\omega_i q T) H_{uv}[\omega_i] - \sin(\omega_i q T) E_{uv}[\omega_i]}{\sqrt{H_{uv}^2[\omega_i] + E_{uv}^2[\omega_i]}} \right)^{-1} \]

\[ = \frac{1}{N} \sum_{i=0}^{N-1} \frac{Z_i[\omega_i, q]}{\sqrt{H_{uv}^2[\omega_i] + E_{uv}^2[\omega_i]}} \]  

(2.30)

This is the analytic solution of the Generalized cross-correlation using the Phase Transform (GCC-PHAT) for microphone pair \((u, v)\).

Using the expressions of \(H_{uv}[\omega_i]\) and \(E_{uv}[\omega_i]\) from Eq. 2.18 and Eq. 2.19 to obtain
an expression for $Z_i[\omega_i, q]$:

$$Z_i[\omega_i, q] = \left(\cos(\omega_i q T) \left[ \frac{D^2}{d_ud_v} \cos(\omega_i(q_{e0} - q_{a0})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \cos(\omega_i(q_{e_l} - q_{a_k})T) \right] - \sin(\omega_i q T) \left[ \frac{D^2}{d_ud_v} \sin(\omega_i(q_{e0} - q_{a0})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \sin(\omega_i(q_{e_l} - q_{a_k})T) \right] \right)$$

(2.31)

I next group the cosine and sine terms of the direct-path and reverberant-paths into two separated groups:

$$Z_i[\omega_i, q] = \left(\frac{D^2}{d_ud_v} \left[ \cos(\omega_i q T) \cos(\omega_i(q_{e0} - q_{a0})T) - \sin(\omega_i q T) \sin(\omega_i(q_{e0} - q_{a0})T) \right] + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \left[ \cos(\omega_i q T) \cos(\omega_i(q_{e_l} - q_{a_k})T) - \sin(\omega_i q T) \sin(\omega_i(q_{e_l} - q_{a_k})T) \right] \right)$$

$$= \left(\frac{D^2}{d_ud_v} \cos(\omega_i(q_{e0} - q_{a0})T + \omega_i q T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \cos(\omega_i(q_{e_l} - q_{a_k})T + \omega_i q T) \right)$$

$$= \frac{1}{N} \sum_{i=0}^{N-1} \left( \frac{D^2}{d_ud_v} \cos(\omega_i(q_{e0} - q_{a0})T + \omega_i q T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \cos(\omega_i(q_{e_l} - q_{a_k})T + \omega_i q T) \right)$$

(2.32)

where $q_{uv} \equiv q_{e0} - q_{a0}$ and $q_{ukvl} \equiv q_{ak} - q_{e_l}$. Note that, $q_{uv}$ is the true time-difference of arrival (TDOA) in samples between two signals at microphones $u$ and $v$. Suppose I hypothesizes a source location at a 3D-point in the room, $\vec{x}$. Given $\vec{x}$ and the two microphone locations $\vec{l}_u$ and $\vec{l}_v$, I can compute the TDOA to the nearest sample as,

$$\Delta_{uv}(\vec{x}) = \left[ \frac{||\vec{l}_u - \vec{x}|| - ||\vec{l}_v - \vec{x}||}{CT} \right]$$

(2.33)
For this specific sample value, I get the analytic GCC-PHAT value as,

\[
G_{uv}[\vec{x}] = \frac{1}{N} \sum_{i=0}^{N-1} \left( \frac{D^2_d\cos(\omega_i(\Delta_{uv}(\vec{x}) - q_{uv})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2_{a_k b_{vl}}\cos(\omega_i(\Delta_{uv}(\vec{x}) - q_{uv})T)}{\sqrt{H^2_{uv}[\omega_i] + E^2_{uv}[\omega_i]}}}{\sqrt{H^2_{uv}[\omega_i] + E^2_{uv}[\omega_i]}} \right)
\]

(2.34)

Note that here I have assumed that the important parts of the room’s impulse response (RIR) for each microphone is known for every point in the room to be evaluated. Hence, all the quantities \(q_{uv}, q_{ukvl}, d_u, d_v, a_u, b_v\) need to be known. Therefore, the GCC-PHAT value is only a function of the hypothesized TDOA \(\Delta_{uv}(\vec{x})\). Also note that only the numerator is a function of \(\vec{x}\), whereas the denominator serves as a normalization.

The analytic SRP-PHAT functional value for a hypothesized location \(\vec{x}\) can be computed as the sum of GCC-PHAT values over all unique combinations of microphone pairs \((u, v)\) [24, 101] as follows,

\[
P(\vec{x}) =
\frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} \sum_{i=0}^{N-1} \left( \frac{D^2_d\cos(\omega_i(\Delta_{uv}(\vec{x}) - q_{uv})T) + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2_{a_k b_{vl}}\cos(\omega_i(\Delta_{uv}(\vec{x}) - q_{uv})T)}{\sqrt{H^2_{uv}[\omega_i] + E^2_{uv}[\omega_i]}}}{\sqrt{H^2_{uv}[\omega_i] + E^2_{uv}[\omega_i]}} \right)
\]

(2.35)

### 2.3 Why SRP-PHAT is robust?

**Lemma 1:** A sum of a large number of cosine values, where the angle \(\theta\) of the cosine is a uniformly distributed random variable in \([0, 2k\pi]\) where \(k\) is an integer, approaches
zero. Mathematically, one has: Given $\theta_n \sim U[0, 2k\pi]$, then $\sum_{n=1}^{N} \cos(\theta_n) \to 0$, when $N \to \infty$.

This follows directly from the Law of large numbers (LLN) [42]: Given a random variable with a finite expected value, if its values are repeatedly sampled, as the number of these observations increases, the sample mean will tend to approach and stay close to the expected value. Let us compute the expected value of a cosine, $E\{\cos(x)\}$, for a probability density function $f(x)$:

$$E\{\cos(x)\} = \int_{0}^{2k\pi} \cos(x) f(x) \, dx$$

$x \sim U[0, 2k\pi] \rightarrow f(x) = \frac{1}{2k\pi}$, hence:

$$E\{\cos(x)\} = \frac{1}{2k\pi} \int_{0}^{2k\pi} \cos(x) \, dx$$

$$= \frac{1}{2k\pi} \sin(x)|_{0}^{2k\pi}$$

$$= 0 \quad (2.37)$$

Therefore, the mean, and thus, the sum of a large number of cosines, where the angle is a uniformly distributed random variable, approaches zero.

Recall the analytic SRP-PHAT solution from Eq. 2.35:

$$P(\vec{x}) = \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \left( \frac{D^2 \cos(\omega_i (q(\vec{x}) - q_{uv})^T)}{d_\omega d_\nu D_{EH}[\omega_i]} + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2 \cos(\omega_i (q(\vec{x}) - q_{ukvl})^T)}{a_{uk} b_{vl} D_{EH}[\omega_i]} \right) \quad (2.38)$$

19
where $D_{EH}[\omega_i]$ denotes the denominator of Eq. 2.35 or,

$$D_{EH}[\omega_i] = (H_{uv}^2[\omega_i] + E_{uv}^2[\omega_i])^{\frac{1}{2}}$$

$$= \left( \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \cos (\omega_i (q_{el} - q_{uk})T) \right)^2 + \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2}{a_{uk} b_{vl}} \sin (\omega_i (q_{el} - q_{uk})T) \right)^2 \right)^{\frac{1}{2}}$$

$$= \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{uk} b_{vl})^2} + \sum_{k_1} \sum_{l_1} \sum_{k_2} \sum_{l_2} \frac{2D^4}{(a_{uk_1} b_{vl_1})(a_{uk_2} b_{vl_2})} \cos (\omega_i (q_{el_1} - q_{uk_1} - q_{el_2} + q_{uk_2})T) \right)^{\frac{1}{2}}$$

$$= \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{uk} b_{vl})^2} + \sum_{k_1} \sum_{l_1} \sum_{k_2} \sum_{l_2} B_1(u, v, k, l) \cos (\omega_i (q_{el_1} - q_{uk_1} - q_{el_2} + q_{uk_2})T) \right)^{\frac{1}{2}}$$

$$= \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{uk} b_{vl})^2} + \sum_{k_1} \sum_{l_1} \sum_{k_2} \sum_{l_2} Q_1(u, v, k_1, l) \right)^{\frac{1}{2}},$$

(2.39)

where $k_1, k_2 \in [0, K]$, $l_1, l_2 \in [0, L]$, and $k_1, l_1$ are not equal to $k_2, l_2$ respectively at the same time. Note that $D_{EH}[\omega_i] > 0$. One has:

$$B_1(u, v, k, l) = \frac{2D^4}{(a_{uk_1} b_{vl_1})(a_{uk_2} b_{vl_2})}$$

(2.40)

And from Eq. 2.39, I have:

$$Q_1(u, v, k, l) \equiv B_1(u, v, k, l) \cos (\omega_i (q_{el_1} - q_{uk_1} - q_{el_2} + q_{uk_2})T)$$

$$= B_1(u, v, k, l) O_1(u, v, k, l)$$

(2.41)

Earlier, I defined: $D \equiv \min_{1 \leq m \leq M} d_m$. Hence, $D^4 \leq (d_u d_v)^2 \leq (a_{uk_1} b_{vl_1})(a_{uk_2} b_{vl_2})$. Thus, $\forall k_1, k_2, l_1, l_2, a, b, B_1(u, v, k, l)$ in Eq. 2.40 may be considered a random value bounded in [0,2], and the cosine value $O_1(u, v, k, l)$ varies in [-1,1]. Thus, the two
variables $O_1(u, v, k, l, i)$ and $B_1(u, v, k, l)$ are approximately uncorrelated or $\text{cov}\{B_1, O_1\} = 0$. Therefore,

$$E\{Q_1\} = E\{B_1, O_1\}$$

$$= E\{B_2\}E\{O_1\} + \text{cov}\{B_1, O_1\}$$

$$= E\{B_1\}E\{O_1\}$$

$$= \frac{1}{(1 - 0)^0}$$

$$= 0$$  \hspace{1cm} (2.42)

Eq. 2.39 and Eq. 2.42 yield,

$$D_{EH}[\omega_i] \rightarrow \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{uk} b_{vl})^2}\right)^{\frac{1}{2}}$$  \hspace{1cm} (2.43)

When steering to the true source location $\vec{x}_s$, for microphone pair $(u, v)$: $\Delta_{uv}(\vec{x}) = \Delta_{uv}(\vec{x}_s) = q_{uv}$, hence from Eq. 2.38 I have:

$$P(\vec{x}_s) = \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \left( \frac{D^2}{d_{uv} D_{EH}[\omega_i]} + \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{\cos(\omega_i (q_{uv} - q_{ukvl}) T)}{a_{uk} b_{vl} D_{EH}[\omega_i]} \right)$$

$$\equiv \lambda_1(\vec{x}_s) + \lambda_2(\vec{x}_s)$$  \hspace{1cm} (2.44)

The first term of Eq. 2.44, $\lambda_1(\vec{x}_s) = \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} \sum_{i=0}^{N-1} \frac{D^2}{d_{uv} D_{EH}[\omega_i]}$, is a positive constant, which corresponds to the direct paths only. Let us consider the second term of Eq. 2.44, which includes all the cross-terms between direct paths and reverberant
paths, and among reverberant paths themselves:

\[
\lambda_2(\vec{x}_s) = \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2 \cos (\omega_i (q_{uv} - q_{ukvl}) T)}{a_{uk} b_{vl} D_{EH}[\omega_i]}
\]

\[
= \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \sum_{k=0}^{K} \sum_{l=0}^{L} B_2(u, v, k, l, i) \cos (\omega_i (q_{uv} - q_{ukvl}) T) \quad (2.45)
\]

Considering \( B_2(u, v, k, l, i) \) of Eq. 2.45 and from Eq. 2.43, I has:

\[
B_2(u, v, k, l, i) = \frac{D^2}{a_{uk} b_{vl} D_{EH}[\omega_i]} \frac{D^2}{a_{uk} b_{vl}} \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{uk} b_{vl})^2} \right)^{\frac{1}{2}}
\]

\[
= \frac{1}{ \left( 1 + \sum_{k'=0}^{K} \sum_{l'=0}^{L} \frac{(a_{uk'} b_{vl'})^2}{(a_{uk} b_{vl})^2} \right)^{\frac{1}{2}}} \quad (2.46)
\]

Note that \( a_{uk}, b_{vl} > 0 \) and \( B_2(u, v, k, l, i) < 1 \). Consider the term in Eq. 2.45:

\[
Q_2 = \frac{D^2 \cos (\omega_i (q_{uv} - q_{ukvl}) T)}{a_{uk} b_{vl} D_{EH}[\omega_i]}
\]

\[
= B_2(u, v, k, l, i) \cos (\omega_i (q_{uv} - q_{ukvl}) T)
\]

\[
= B_2(u, v, k, l, i) O_2(u, v, k, l, i) \quad (2.47)
\]

While the cosine value \( O_2(u, v, k, l, i) \) varies in \([-1,1]\), \( B_2(u, v, k, l, i) \) may be considered a random value bounded in \([0,1]\). Thus, the two variables \( B_2(u, v, k, l, i) \) and \( O_2(u, v, k, l, i) \) are approximately uncorrelated or \( \text{cov}\{B_2, O_2\} = 0 \). Using a similar argument to Eq. 2.42 and from Eq. 2.45 yield,

\[
\lambda_2(\vec{x}_s) \to 0 \quad (2.48)
\]
Therefore, from Eq. 2.48, 2.44, and Eq. 2.43 I have:

\[
\begin{align*}
P(\vec{x}_s) &\to \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} \sum_{i=0}^{N-1} d_u d_v D^2 \frac{d}{d u} D_{EH}[\omega_i] \\
&\approx \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} \sum_{i=0}^{N-1} d_u d_v \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^4}{(a_{u_k} b_{v_l})^2} \right)^{\frac{1}{2}} \\
&= \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} \sum_{i=0}^{N-1} d_u d_v \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{1}{(a_{u_k} b_{v_l})^2} \right)^{\frac{1}{2}}.
\end{align*}
\]

(2.49)

As the reverberation noise increases, i.e., \( K, L \) increase, Eq. 2.49 yields a smaller value of \( P(\vec{x}_s) \). This means the SRP-PHAT value decreases as the reverberation noise increases. However, as \( K \) and \( L \) get bigger, the amplitudes of the reflection impulses get smaller, i.e., \( \frac{1}{a_{u_k}} \) and \( \frac{1}{b_{v_l}} \) attenuate. At some point \( (k \geq K_1 \text{ and } l \geq L_1) \), these amplitudes would only contribute insignificant values to the sum of Eq. 2.43. Therefore, I would expect that \( D_{EH}[\omega_i] \) reaches a relatively constant value as the reverberation time reaches some high value. In other words, the SRP-PHAT value stays the same regardless of the length of the reverberation noise when the length of the noise reaches a certain high value. This phenomenon will be verified experimentally in Sec.2.4.

When steering to a location \( \vec{x}_n \) other than \( \vec{x}_s \), the first term of Eq. 2.38, \( \cos (\omega (q(\vec{x}_n) - q_{uv})) d_u d_v D_{EH}[\omega_i] \), is no longer a positive constant. It now is a cosine value that can be either negative or positive, and is similar to one of the cosines of the cross-terms in Eq. 2.38. Therefore, in this case one can generalize the expression of the analytic SRP-PHAT in Eq. 2.38.
as follows,

\[
P(\vec{x}_n) = \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \left( \sum_{k=0}^{K} \sum_{l=0}^{L} \frac{D^2 \cos(\omega_i (q(\vec{x}_n) - q_{ukvl}) T)}{a_{uk} b_{vl} D_{EH}[\omega_i]} \right)
\]

\[
= \frac{1}{N} \sum_{u=1}^{M-1} \sum_{v=(u+1)}^{M} \sum_{i=0}^{N-1} \left( \sum_{k=0}^{K} \sum_{l=0}^{L} Q_3(\vec{x}_n, u, v, k, l, i) \right)
\]

(2.50)

The expression in Eq. 2.50 is similar to the expression of \(\lambda_2(\vec{x}_s)\) in Eq. 2.45. Therefore, using similar arguments that led to Eq. 2.48, I have,

\[
P(\vec{x}_n) \to 0
\]

(2.51)

From Eq. 2.49 & 2.51, it can be seen that the analytic SRP-PHAT value is always enhanced when steering to the true source location, and is small when steering elsewhere, regardless of the amount of reverberation (under my assumption of no background noise and perfect reflections in the signal model). Next, I will verify this experimentally.

2.4 On PHAT robustness validation

Although normally real data in a real room is used for validations, in this instance, simulated data was used because the reflective properties of the walls, the number of significant reflections, and the background noise level could be controlled. The Brown Acoustic Simulator (BAS) [57] was used to generate perfect data for a rectangular room with perfectly reflecting walls, floor, and ceiling of size \((x,y,z)\): 4.8m x 3.3m x 6.8m, the size of the room in which we normally take real data. I consider a focal area of the microphone array (using 24 microphones) to be 4m x 1m x 6m, restricting closeness to walls and to the range of human heights. A sampling frequency of 20KHz \((T=50\mu s)\) and frames of 51.2 ms (or \(N=1024\) samples) were used. No background
noise was added. The room, microphone array, and source location and orientation are shown in Fig. 2.1. The source was assumed a perfect spherical radiation from a point source. The source signal is a white Gaussian noise signal. From the microphone array’s impulse responses, I can obtain all the quantities $d_u, d_v, a_{uk}, b_{v1}$, and $q_{uv}, q_{ukvl}$ to construct the analytic SRP-PHAT solution as in Eq. 2.38.

I implemented a 1cm-resolution grid-search in the focal area using the analytic SRP-PHAT as the functional for 5 different room reverberation times (in milliseconds). Table 2.1 shows the average number of reflections per microphone channel and the SNR level (the analytic SRP-PHAT global peak w.r.t the background noise level) for each value of $T_{60}$. Fig. 2.2, 2.3, 2.4, 2.5, and 2.6 show the side-views (from $z$ and $x$-directions) of the analytic SRP-PHAT surfaces for different $T_{60}$ times.

Clearly, the analytic SRP-PHAT grid-search surface exhibits a global peak at the true source location for all reverberation levels. Moreover, under my assumption of no background noise and perfect reflection, the SRP-PHAT performs in accordance
<table>
<thead>
<tr>
<th>$T_{60}$ (ms)</th>
<th>Avg. No. reflections</th>
<th>SNR (dB)</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>5</td>
<td>7</td>
<td>Fig. 2.2</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>6</td>
<td>Fig. 2.3</td>
</tr>
<tr>
<td>200</td>
<td>88</td>
<td>5.5</td>
<td>Fig. 2.4</td>
</tr>
<tr>
<td>540</td>
<td>139</td>
<td>5</td>
<td>Fig. 2.5</td>
</tr>
<tr>
<td>990</td>
<td>141</td>
<td>5</td>
<td>Fig. 2.6</td>
</tr>
</tbody>
</table>

Table 2.1: Performance of analytic SRP-PHAT for different values of $T_{60}$

Figure 2.2: Side-views of analytic SRP-PHAT surface for $T_{60}=25$ ms

... to my previous analysis. As the reverberation time gets longer ($T_{60} = 25\text{ms} \rightarrow 50\text{ms} \rightarrow 200\text{ms} \rightarrow 500\text{ms}$), the SRP-PHAT value of the global peak slightly decreases (relatively to the noise level). However, even as the reverberation time is quite long ($T_{60} \geq 500\text{ms}$), the SRP-PHAT value of the peak stays the same. In other words, under my assumptions, the performance of SRP-PHAT is consistently robust regard-

Figure 2.3: Side-views of analytic SRP-PHAT surface for $T_{60}=50$ ms
Figure 2.4: Side-views of analytic SRP-PHAT surface for $T_{60}=200$ ms

Figure 2.5: Side-views of analytic SRP-PHAT surface for $T_{60}=540$ ms

less of the severity of reverberation. One should remember that this conditions are in one sense “worst case”, as I have assumed all 100% reflectivity of the surfaces of the room. This observation agrees with previous work, see [116]. In all cases of $T_{60}$, the SNR enhancement given by the SRP-PHAT is about 5 - 7 dB. The general rule for perfect (non-reverberant situation with no background noise and perfect point sources and locations) is 6dB/factor of 4 in the number of microphones [95]. This one has the theoretical limit of $20\log_{10}(N^{\frac{1}{2}}) = 10\log_{10}(\sqrt{24}) = 13.8$dB.
Figure 2.6: Side-views of analytic SRP-PHAT surface for $T_{60} = 990$ ms
Chapter 3

Computationally efficient SRP-PHAT methods for locating a single source

3.1 Problem statement

In this thesis, I address two localization problems:

- Algorithms for making single-source localization computationally feasible
- Robust methods for locating multiple sources in adverse environments

Note that in both problems, only a single frame of microphone-array data is used to do locationing without tracking. If location estimates are available for multiple frames, smoothing techniques or some statistical averaging can be applied to get more accurate estimates of the source locations.

In noisy conditions – those with high background and reverberation noise, the performance of a single-source locator using TDOA-based, two-stage methods [101, 11, 97] often degrades quickly [23, 8]. Steered-beamformer-based methods using the
SRP-PHAT functional are more robust in this scenario [23, 101]. As briefly stated in Chapter 1, to find the source location, I steer the beamformer over “all” candidate locations in a focal volume containing the source. The point that gives the maximum weighted output power (SRP-PHAT) of the beamformer will indicate the point-source location, i.e.,

\[
\vec{x}_s = \arg\max_{\vec{x}} P(\vec{x}) \tag{3.1}
\]

where \( P(\vec{x}) \) is the SRP-PHAT at point \( \vec{x} \). For notation, the calculation of any particular point of \( P(\vec{x}) \) will be called a \textbf{functional evaluation (fe)}.

The hypothesis is that the SRP-PHAT will peak at the actual source location even under very noisy and highly reverberant conditions. However, the problem with SRP-PHAT is its very high computational cost because the search space has many local maxima, and thus computationally-intensive grid-search methods have often been used to find the global maximum. This chapter introduces three different methods to cut down the computational cost by more than three orders of magnitude compared to that of the full grid-search, thus making the SRP-PHAT single-source locator more practical in real-time. The three methods are: stochastic region contraction (SRC), coarse-to-fine region contraction (CFRC), and stochastic particle filtering (SPF).

### 3.2 Stochastic region contraction (SRC)

#### 3.2.1 Overview

Stochastic region contraction (SRC) was first introduced by Berger and Silverman [7] as a global-optimization technique to optimize a microphone array’s configuration. An optimization problem for placements and gains of microphones in an array was solved by minimizing the power spectral dispersion function (PSD) at its worst case over some given noise field. However, the PSD surface exhibited many local maxima...
and minima (hundreds of thousands). Hence, common optimization techniques like gradient descent or simplex search, which rely on the functional field be smooth, cannot be used. The introduced optimization technique, SRC, was shown to be robust in finding the global optimum in this case. Observing a similarity between this and the global maximum finding problem existed in SRP-PHAT, I proposed using SRC to optimize the searching of the focal area in SRP-PHAT, thus making SRP-PHAT more efficient and practical for real-time use.

### 3.2.2 SRC algorithm

The basic idea of the SRC algorithm is, given an initial rectangular search volume containing the desired global optimum and perhaps many local maxima or minima, gradually, in an iterative process, contract the original volume until a sufficiently small subvolume is reached in which the global optimum is trapped (the uncertainty voxel \( V_u \)). The contraction operation on iteration \( i \) is based on a stochastic exploration of the SRP-PHAT, \( P(\vec{x}) \) functional in the current subvolume.

![Figure 3.1: 2D example of SRC: The surface is \( P(\vec{x}) \). \( j \) is the iteration index. The rectangular regions show the contracting search regions](image)

31
The first step is to determine a number of random points, or the number of functional evaluations (fe’s), $J_0$, that need to be evaluated to ensure that one or more is likely to be in the volume, $V_{peak}$, of higher values (than the rest of the focal volume) surrounding the global maximum of $P(\bar{x})$. see, e.g. Figure 3.1. Unfortunately, $V_{peak}$ is not easy to determine and in my data changes substantially as the source is farther from the microphones. However, if $V_{room}$ is the original search volume or focal volume, for example, I can estimate the number of fe’s needed to ensure that the probability of missing $V_{peak}$ altogether is less than a given percent.

The probability of a random point hitting $V_{peak}$ in the initial search volume $V_{room}$ is,

$$P(\text{hit } V_{peak}) = \frac{V_{peak}}{V_{room}} \quad (3.2)$$

Hence, the probability of a random point missing $V_{peak}$ is,

$$P(\text{miss } V_{peak}) = 1 - \frac{V_{peak}}{V_{room}} \quad (3.3)$$

The event of throwing a random point is independent from one to another. Therefore, the probability of throwing $n$ random points missing $V_{peak}$ is,

$$P(\text{n-misses}) = (1 - \frac{V_{peak}}{V_{room}})^n \quad (3.4)$$

Taking the logarithm of both sides, separating $n$ to one side, I have,

$$n = \frac{\log P(\text{n-misses})}{\log(1 - \frac{V_{peak}}{V_{room}})} \quad (3.5)$$

From this relationship between the probability of throwing $n$ random points missing $V_{peak}$ and the ratio between $V_{peak}$ and $V_{room}$, I can determine how many random points, $n$, are needed to be thrown to ensure that the missing probability, $P(\text{n-misses})$ is negligible (substantially small in a realistic sense). I calculate $n$ for different values
of missing probability and ratio \( \frac{V_{\text{peak}}}{V_{\text{room}}} \) as shown in follows,

\[
\begin{array}{|c|cccc|}
\hline
\frac{V_{\text{peak}}}{V_{\text{room}}} & P(\text{miss, } V_{\text{peak}}) & 0.1 & 0.01 & 0.001 & 0.0001 \\
\hline
1\% & 44 & 459 & 4603 & 46,050 \\
0.1\% & 66 & 688 & 6905 & 69,075 \\
0.01\% & 88 & 917 & 9206 & 92,099 \\
\hline
\end{array}
\]

Table 3.1: Number of fe’s required for three probabilities of missing \( V_{\text{peak}} \) and four values of the ratio \( \frac{V_{\text{peak}}}{V_{\text{room}}} \).

In my case, \( V_{\text{room}} = 400cm \times 100cm \times 600cm = 24 \times 10^6\,cm^3 \), and from preliminary experimental results, i.e., a grid search, \( V_{\text{peak}} = 12 \times 10^4\,cm^3 \) for a low SNR situation. This makes \( \frac{V_{\text{peak}}}{V_{\text{room}}} \approx 0.005 \), from Table 3.1 implying that a value of \( J_0 = 3000 \) will err by missing the peak volume less than 0.1% of the time. Note that, \( V_{\text{peak}} \) is dependent on the room, microphones, and source conditions, thus, as these conditions change, a preliminary experiment should be carried out to determine roughly the worst-case value of \( V_{\text{peak}} \).

The following notations are defined:

- \( J_i \) as the number of random points evaluated for iteration \( i \)
- \( N_i \), the number of points used to define the new source volume
- \( V_{i+1} \), having a rectangular-boundary column vector \( \vec{B}_{i+1} \equiv [x_{\text{max}}(i+1), x_{\text{min}}(i+1), y_{\text{max}}(i+1), y_{\text{min}}(i+1), z_{\text{max}}(i+1), z_{\text{min}}(i+1)]' \)
- \( I \), the number of iterations
- \( FE_i \), the total number of fe’s evaluated as of iteration \( i \)
- \( \Phi \), the maximum number of fe’s allowed to be computed.

I have found it very effective for my problem to set a fixed value for \( N_i \) based on experimentation as shown in Fig. 3.2. I computed SRC using various fixed values of \( N \), using four real talkers and many frames of data from each. Correct location
performance is plotted in Fig. 3.2 as a function of the fixed N value for each of the sources. Here I see that a value of \( N = 100 \) is about optimal. That is a lower value of N implies lower computational cost, so a value with full performance which is as small as possible should be used. Thus, I let \( N_i \equiv N = 100 \).

Figure 3.2: Performance of SRC-I as a function of parameter N for four different source locations

Given \( J_0 \) and \( N_i \) defined as above, the SRC algorithm for finding the global maximum is,

1. **Initialize**: \( i = 0, J_0 = 3000, N_i = N = 100 \) and \( V_0 = V_{room} \).
2. **Evaluate**: \( P(\vec{x}) \) for \( J_i \) random points in \( V_0 \).
3. **Sort**: the best \( N \ll J_i \) points
4. **Contract**: the search region to the smaller region \( V_{i+1}(\vec{B}_{i+1}) \) that contains these \( N_i \) points.
5. **Test**: IF: \( V_{i+1} < V_u \), or \( FE_i > \Phi \) and \( V_{i+1} < T_1 V_u \), where \( T_1 \) is a parameter (about 10); determine \( \bar{x}_u(i^*) \), \( I = i \), STOP, KEEP RESULT.

   ELSE IF \( FE_i > \Phi \), STOP, DISCARD RESULT.

   ELSE: Among the \( N_i \) points, keep a subset \( G_i \) points that have values greater than the mean, \( \mu_i \) of the \( N_i \) points.
6. **Evaluate**: \( J_{i+1} \) new random points in \( V_{i+1} \).
7. **Form**: the set of the \( N_{i+1} \) as the union of \( G_i \) and the best \( N_{i+1} - G_i \) points from the \( J_{i+1} \) just evaluated. This gives \( N_{i+1} \) high points for iteration \( i + 1 \).
8. **Iterate**: \( i = i + 1 \). GO TO STEP 4.
Depending on how $J_i$ with $i \geq 1$ is defined, I have three variants of the SRC algorithm as follows,

- **SRC-I.** Let $J_i$ be that number of random fe’s needed find $N_i - G_i$ points greater than $\mu_i$. Guarantees monotone increasing $\mu_i$. Use finite value of $\Phi$.
- **SRC-II.** Let $J_i$ be that number of random fe’s needed to find $N_i - G_i$ points higher than the minimum of the full set $N_i$. $\mu_i$ increases for almost all iterations. Use finite value of $\Phi$.
- **SRC-III.** Fix $J_i = J$. Keep the highest $N_i - G_i$ points for each iteration. Does not guarantee monotone increasing $\mu_i$. Set $\Phi \to \infty$.

### 3.3 Coarse-to-fine region contraction (CFRC)

Similar to stochastic region contraction (SRC), coarse-to-fine region contraction (CFRC) method also uses the idea of region contraction, or contracting the search volume smaller and smaller until the global maximum is captured. However, in CFRC, the contraction operation of iteration $i$ is based on a sub-grid search of the $P(\vec{x})$ functional in the current sub-volume.

![Figure 3.3: 2D example of CFRC: The surface $P(\vec{x})$ has many local maxima. $j$ is the iteration index. The rectangular regions show the contracting search regions.](image)

SRP surface has many local maxima and minima (SNR ~ 1.9dB)
Analogous to SRC, the first step of CFRC is to determine the number of initial grid points, \( J_0 \), that need to be evaluated to guarantee that at least a single grid-point lies in \( V_{\text{peak}} \). Again, from my preliminary experimental data for a low SNR case, \( V_{\text{peak}} \) is determined to be \( 28\text{cm} \times 10\text{cm} \times 30\text{cm} \). Recall that my \( V_{\text{room}} \) is \( 400\text{cm} \times 100\text{cm} \times 600\text{cm} \). Hence, to have at least a grid-point in \( V_{\text{peak}} \), I have to evaluate \( \frac{400}{28} \approx 15 \) grid-points in \( x \), \( \frac{100}{10} = 10 \) grid-points in \( y \), and \( \frac{600}{30} = 20 \) grid-points in \( z \), implying \( J_0 = 15 \times 10 \times 20 = 3000 \) equally spaced grid points in 3D.

I define \( J_i \) as the number of grid points evaluated for iteration \( i \). \( N_i, V_{i+1}, \vec{B}_{i+1}, FE_i \) and \( \Phi \) are defined the same way as in SRC. CFRC can be implemented in many ways, the difference is usually determined by the methods used to update \( J_i \) and \( V_i \) for each iteration. The general algorithm is,

1. **Initialize iteration:** \( i = 0, J_0, N_0, V_0 = V_{\text{room}} \)
2. **Evaluate:** \( P(\vec{x}) \) for \( J_0 \) points.
3. **Sort:** the best \( N_1 \ll J_0 \) points.
4. **Contract:** the search region to the smaller region \( V_{i+1} \), \( \vec{B}_{i+1} \) that contains these \( N_i \) points.
5. **Test:** IF: \( V_{i+1} < V_u \), and \( FE_i < \Phi \); determine \( \vec{x}_s \), STOP, KEEP RESULT.
   ELSE IF \( FE_i > \Phi \), STOP, DISCARD RESULT.
   ELSE: From the \( N_i \) points, keep a subset \( G_i \) points that have values \( \geq \) the mean, \( \mu_i \) of the \( N_i \) points.
6. **Evaluate:** \( J_{i+1} \) new grid points in \( V_{i+1} \).
7. **Sort:** to obtain the best \( N_{i+1} \) points from the union of \( G_i \) and \( J_{i+1} \) points just evaluated.
8. **Iterate:** \( i = i + 1 \). GO TO STEP 4.

For my data, I chose the simplest algorithm to select \( J_i \) and \( N_i \). I made each of these constant, although for iterations \( i \geq 1 \), \( J_i \equiv J \neq J_0 \). \( J \) was selected to give “best”\(^1\)

\(^1\)For the current data, “best” implied perfect – all locations determined by a full grid search were obtained by the CFRC method.
performance with lowest cost, \( J = 10 \times 5 \times 15 = 750 \) points. As I did before in Fig. 3.2 for SRC, I ran the CFRC algorithm for various values of \( N \), as shown in Figure 3.4. An \( N = 100 \) turned out to preserve perfect performance at a low cost.

![Figure 3.4: Performance of CFRC as a function of parameter \( N \) for four different source locations](image)

### 3.4 Stochastic particle filtering (SPF)

Particle filtering is a sequential Monte Carlo methodology where the basic idea is the recursive computation of relevant probability distributions using the concepts of importance sampling and approximation of probability distributions with discrete random measures [25]. The earliest applications of sequential Monte Carlo methods were in the area of growing polymers [44, 80]. Sequential Monte Carlo methods found limited use in the past, primarily due to their very high computational complexity and the lack of adequate computing resources of the time. The fast advances of computers in the last several decades and the outstanding potential of particle filters have made them recently a very active area of research. Their potential for parallel implementation represents additional impetus for their development. The current interest in particle filtering for signal processing applications can be found in much work [33, 21, 61].
In contrast to many particle filtering algorithms\cite{110, 106}, where the initial set of particles (or 3-D points) is chosen randomly or uniformly without considering the probabilistic relation between the global maximum and the initial search volume, SPF uses the probabilistic approach of SRC to select the initial set of highly probable points (importance sampling step) that guarantees to capture the global peak with a negligible missing probability at a much faster convergence rate. In addition, I introduce a new resampling algorithm in SPF to avoid the degeneracy (also referred as impoverishment) problem of particle filters. Another new feature of the algorithm is that the SRP-PHAT value is used as the particle weight.

Many particle filtering algorithms \cite{110, 106}, choose the initial set of particles (or points) randomly or uniformly without regarding a probabilistic measure of converging to the global maximum. On the other hand, SPF selects the initial set of particles, $J_0$, in such a way that it guarantees a particle in the volume surrounding the global maximum, $V_{\text{peak}}$, is captured with a missing probability of less than 0.1%. This is the same initialization step used to get $n$ particles in SRC, see Eq. 3.5. In addition, a new, robust resampling technique is introduced to avoid the degeneracy problem of particle filters. The SPF searching algorithm for a frame $f$ consists of two steps: Importance sampling and Resampling as follows,

(i) **Importance Sampling:**

1) Initialize: $k = 0$. For $p = 1, \ldots, J_k$, randomly sample $\vec{x}(p) \sim U[\vec{B}_k]$, where $U$ denotes the uniform distribution, and $\vec{B}_k$ defines the 3-D boundary of the search volume $V_k$.

2) Evaluate to get the set $W$ of the $J_k$ importance weights (SRP-PHAT values):

$$ W = P(\vec{x}(p)) $$

3) Select a set $N_k$ containing the best $N \ll J_k$ out of the $J_k$ particles, $\vec{x}(n)$, $n = 1, \ldots, N$:

$$ W = \text{SORT}^{-}(W) $$

$$ W(\vec{x}(n)) = \bar{W}(n), n = 1, \ldots, N, $$

where $\text{SORT}^{-} \equiv$ sorting in descending order.
4) Calculate the standard-deviation, $\sigma(k)$ of the $N$ particles $\bar{x}(n)$.

5) Normalize the importance weights:

$$SW_k = \sum_{i=1}^{N} P(\bar{x}(n))$$  \hspace{1cm} (3.8)

$$\bar{P}(\bar{x}(n)) = \frac{P(\bar{x}(n))}{SW_k}, n = 1, ..., N$$  \hspace{1cm} (3.9)

6) Number of functional evaluations (fe’s):

$$\Phi_k = J_k$$  \hspace{1cm} (3.10)

(ii) **Resampling:** IF $\sigma(k) \geq 0.05m$:

1) $SORT^\sim$ $\bar{x}(n), n = 1, ..., N$ in terms of their weights

2) Calculate the number of replications (a binary value, either 0 or 1, since all the weights are normalized from Eq. 3.9), $i(k,n)$ for each particle $\bar{x}(n)$:

For $n = 1 \rightarrow N$:

$$i(k,n) = \lfloor \bar{P}(\bar{x}(n)) \times N \rfloor; n = 1, ..., N.$$  \hspace{1cm} (3.11)

3) Calculate the number of particles that are eliminated, $e(k)$:

$$e(k) = |i(k,n)| \supset i(k,n) = 0$$  \hspace{1cm} (3.12)

where $|.|$ denotes cardinality of the set.

4) Find the set of $R$ replicated particles:

$$\hat{X}_k = \bar{x}(r) = \bar{x}(n) \times \mathbf{1}(i(k,n))$$  \hspace{1cm} (3.13)

where $r = 1, ..., R \leq N$, and $\mathbf{1}$ denotes a vector having all elements equal to 1. This creates $i(k,n)$ copies of $\bar{x}(n)$ for each $n$.

5) Find the set of replicated (de-normalized) weights:

$$W_k = P(\hat{X}_k) = \bar{P}(\bar{x}(n)) \times \mathbf{1}(i(k,n)) \times SW_k$$  \hspace{1cm} (3.14)

6) Adding $e(k)$ new particles, which are randomly selected within 0.1m-deviation of the “top
weighted” $e(k)$ original particles:

$$\hat{X}_k = \bar{x}(a) \pm \gamma \times 0.1$$  \hspace{1cm} (3.15)

where $a = 1, \ldots, e(k)$, and $\gamma$ is a random number in $[0,1]$.

7) Evaluate the importance weights of the $e(k)$ newly added particles:

$$\hat{W}_k = P(\hat{X}_k)$$  \hspace{1cm} (3.16)

8) Create the new set of particles and weights for next iteration:

$$X_{k+1} = [\hat{X}_k; \hat{X}_k]$$  \hspace{1cm} (3.17)

$$W_{k+1} = [\hat{W}_k; \hat{W}_k]$$  \hspace{1cm} (3.18)

9) Update the normalizing factor (weight sum):

$$SW_{k+1} = \sum_{n=1}^{N} (W_{k+1})$$  \hspace{1cm} (3.19)

10) Normalize the importance weights:

$$W_{k+1} = \frac{W_{k+1}}{SW_{k+1}}$$  \hspace{1cm} (3.20)

11) Calculate the new standard-deviation, $\sigma(k + 1)$ of the $N$ particles $\bar{x}(n)$ in the set $X_{k+1}$.

12) Update the number of fe’s:

$$\Phi_{k+1} = \Phi_k + e(k)$$  \hspace{1cm} (3.21)

13) If $\Phi_{k+1} \geq \Phi_T$: Exit and discard the estimate.

14) Iterate: $k = k + 1$.

ELSE: Output the final location estimate, $\bar{x}$ and number of fe’s, $\Phi$:

$$\bar{x} = \sum_{n=1}^{N} (\bar{x}(n) \times \acute{P}(\bar{x}(n)))$$  \hspace{1cm} (3.22)
Next, I will discuss the performances and computational costs of the three proposed methods.

3.5 Comparison of computational cost and accuracy for single-source, SRP-PHAT localization methods

The HMA system [92, 93] is situated in a room with a $T_{60} = 450\text{ms}$ and a focal volume $V_{\text{room}} = 4\text{m} \times 1\text{m} \times 6\text{m}$ was used in my experiments. The source was an Advent AV009 speaker playing an 8-second clean-speech recording (wav file) at 5 different locations facing the 24 microphones in the room as shown in Fig. 3.5. A recording of different native American English talkers was played at each location. Hence, there were 5 recordings of 5 different talkers (1 female and 4 males) played at 5 locations. By varying the source locations, volumes, and talkers, versatility of SRC, CFRC, and SPF could be tested for different SNR situations and talker characteristics.

Frames of 102.4ms, advancing each 25.6ms, and a sampling rate of 20 KHz were the conditions for testing. I selected “speech frames” by hand-labeling frames from the close-talking data that have speech in at least 90% of the framelength. They total to about 62% of all 1500 frames. On the “speech” frames, performance of the full grid-search was compared to the measured locations of the source. On the frames where the full grid-search gave correct estimates, denoted as “good” frames (~ 916 frames), performance of SPF, SRC-I, and CFRC were compared relative to the performance of the grid search, i.e., performance is listed as 100% if SPF, SRC-I, & CFRC achieved the global maximum everywhere the grid-search did. Results are
Figure 3.5: Top view of the array, showing source locations and panels. This experiment used 24 microphones of the 128 on panels H, I, J, K. The arrows indicate the orientation of the talker and the SNR’s are for background noise only.

given in Table 3.2 for accuracy and the average number of fe’s used for SPF, SRC-I, and CFRC on “good” frames, and for grid-search on all “speech” frames. A location estimate was considered an error if it was either off by more than 5cm in x or z or 10cm in y, the vertical dimension².

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>SNR</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.2dB</td>
<td>9.83dB</td>
<td>12.73dB</td>
<td>7.67dB</td>
<td>10.30dB</td>
<td></td>
</tr>
<tr>
<td>% Corr. # fe’s</td>
<td>% Corr. # fe’s</td>
<td>%Corr. # fe’s</td>
<td>% Corr # fe’s</td>
<td>% Corr # fe’s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grid Search</td>
<td>98.54 98.17</td>
<td>97.96 97.95</td>
<td>71.57 71.57</td>
<td>89.96 89.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPF</td>
<td>99 99</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRC - I</td>
<td>99 11.649</td>
<td>100 10.300</td>
<td>100 10.884</td>
<td>100 10.913</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFRC</td>
<td>98 12.171</td>
<td>100 11.031</td>
<td>99 11.380</td>
<td>100 11.510</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Performance and cost of SRP-PHAT using a full grid search over all speech frames; SPF, SRC, and CFRC over good frames for 5 different locations.

3.6 Conclusion

- The SRP-PHAT functional fails more often as distance between the source and the array grows

²The aperture of the array in the vertical dimension is about \( \frac{1}{3} \) that for the other two dimensions
All three proposed methods gave compatible performance nearly as well as the full grid search.

The number of fe’s for all three algorithms is about the same order of magnitude.

More than three orders of magnitude in computation are saved by each of the three proposed methods.

In summary, this chapter has introduced three global-optimization techniques: stochastic region contraction (SRC), coarse-to-fine region contraction (CFRC), and stochastic particle filtering (SPF) to reduce the computational cost of SRP-PHAT by three orders of magnitude, relative to a full grid search, while maintaining the same accuracy of the full search. This makes SRP-PHAT methods more feasible for a real-time, single-source locator. For multiple-source cases, especially in adverse environments, it is essential for the locator to have a robust functional. In fact, we have been able to do a carefully constructed SRC SRP-PHAT locator in real time on a quad-core Intel I7 930 processor for 24 microphones at a rate of 10 frames estimated per second. In the next chapter, the prealignment enhancement, which augment the SRP-PHAT functional in the tasks of multiple-source beamforming and localization, and two multiple-source localization algorithms will be presented.
Chapter 4

Determining locations of multiple sources and proposed enhancements

4.1 Increased complexity when multiple sources are considered

Locating multiple sources, especially in adverse environments, is a much more challenging task than is single-source localization. The effects of interference, cross-talk, background and reverberation noise really complicate the task. Quite often, when multiple sources are active simultaneously, the speech signals coming from soft-talkers are buried by the dominant ones and hence, the soft-talkers are not easily detected. Also, reverberation creates “spurious” sources that can be mistaken for actual sources. This problem is further complicated when only a single short segment of microphone-array data, i.e., a frame, is available. In this chapter, the problem of locating multiple sound-sources using just a single frame of array data, without tracking, in an adverse environment is studied. The approach that I employ in this problem is the steered
beamforming-based (i.e., functional steered response) approach, the same one that I used for single-source localization. In the next section, an enhancement for the classic SRP-PHAT functional and any cross-correlation-based functionals in general is proposed.

4.2 Enhancement for SRP-PHAT functional: Pre-alignment

SRP-PHAT methods, besides robustness against reverberation, exhibit a nice property of showing multiple peaks in the functional space, corresponding to multiple desired sources. Fig. 4.1 shows (a) a slice of the two-stage LEMSAlg’s functional (minimum mean-square error (MMSE) between the hypothesized TDOAs and measured TDOAs) [97] and (b) a slice of SRP-PHAT surface through the average height of the two talkers. Clearly, the SRP-PHAT surface exhibits two distinct peaks for two talkers, while the LEMSAlg’s functional, although is very smooth, shows an inaccurate peak.

Some previous work has been done on improving SRP-PHAT. In [32], the authors introduced a heuristic improved version of PHAT, $\beta$-PHAT, which showed a small improvement in locating a single source. In this chapter and the next one, I shall present two general, robust enhancements: prealignment and microphone pair selection, for SRP-PHAT methods based on observations of the cross-correlation functions.

Consider a set of $M$ microphones with discrete signals $z_u[n]$, where $u = 1, \ldots, M$, $n$ is the time index, and $\vec{x}_u$ the 3-D Cartesian coordinate location for microphone $u$. Also let $\vec{x}_h$ be some hypothetical point-source location in the focal area of the array. An “upper-triangular, average” SRP-PHAT\(^1\) value evaluated at $\vec{x}_h$ maybe computed

\[ \text{Direct computation will include the diagonal terms plus twice the “upper triangular” terms} \]
\[ \text{Averaging is done to make the measure independent of the number of microphone pairs} \]

45
Figure 4.1: (a) the functional surface from a 2-stage LEMSAlg and (b) SRP-PHAT surface.

[23] as the combination of the generalized cross-correlations using the phase transform (GCC-PHAT), of all unique microphone pairs \{u, v\} as follows,

$$\text{SRP}(\vec{x}_h) \equiv \frac{2}{M(M-1)} \sum_{u=1}^{M-1} \sum_{v=u+1}^{M} R_{uv}[d_{uv}],$$  \hspace{1cm} (4.1)

where $d_{uv}$ is the time-difference of arrival (TDOA) in samples between microphones $u$ and $v$ and the hypothesized source, and $R_{uv}[d_{uv}]$ is the quantized GCC-PHAT of microphone pair \{u, v\} and is defined as,

$$R_{uv}[d_{uv}] \equiv \frac{1}{N} \sum_{r=0}^{N-1} \frac{Z_u[r] Z_v^*[r]}{|Z_u[r] Z_v[r]|} e^{\frac{2\pi r d_{uv}}{N}},$$  \hspace{1cm} (4.2)

where $r$ is the frequency index, $N$ is the number of frequency samples (frame-length), and "*" denotes the complex conjugate. $Z_u, Z_v$ are the DFT representations of the time-domain microphone signals $z_u, z_v$ respectively.

For a fixed time-frame of $N$ samples, when the TDOA between two microphone signals is sufficiently large compared to $N$ as depicted in Fig. 4.2, all correlation values between $z_u[n]$ and $z_v[n]$ are poor because there is a mismatch of temporal events. The observed mismatch, if repeated in many microphone pairs \{u, v\}, will
result in an inappropriately low value of the SRP-PHAT from Eq. 4.1. This mismatch

\[ z_u[n] \]

\[ z_v[n] \]

\[ n = 0 \quad n = N-1 \]

Figure 4.2: Mismatch of temporal events between microphone signals \( z_u \) and \( z_v \)

phenomenon is often evident when there is a significant spatial separation between two
microphones \( u \) and \( v \), as happens in larger rooms using a large-aperture microphone
array. Interestingly, large separation of microphones should give better resolution
than small ones! The errors induced by this problem can be eliminated by some
natural prealignments.

### 4.2.1 Preliminary experiment to justify the use of prealignment

A preliminary experiment in a real room to justify the enhancement given the pre-
alignment over using a fixed time-frame to the SRP-PHAT functional was done using
the same recording (of talker T2) used in Sec. 3.5 and the set up shown in Fig. 3.5. A
51.2-ms (1024 samples using \( F_s = 20 \) KHz) “speech” frame of talker 2 recorded using
the 24 microphones on panel H, I, J, K of the HMA system was used. Fig. 4.3(a)
shows a slice through the talker’s height of a 1-cm grid search using the nominal SRP-
PHAT, i.e., SRP-PHAT with a fixed time frame, and Fig. 4.3(b) shows the same slice
but using the enhanced SRP-PHAT, i.e., all microphone signals are prealigned in the
frame used.
Fig. 4.3: 3-D view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame.

Fig. 4.4(a) shows the $Z$-dimension side view of the nominal SRP-PHAT surface, and Fig. 4.4(b) shows the $Z$-dimension side view of the enhanced SRP-PHAT surface. From these figures, the enhanced SRP-PHAT surface (using prealignment) is not only smoother than the nominal SRP-PHAT surface (using a fixed time frame), but also the true peak corresponding to the source is enhanced better relative to the background level. This enhancement is again observed in the $X$-dimension side views in Fig. 4.5(a) and Fig. 4.5(b).
Figure 4.4: Z-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 51.2 ms

Figure 4.5: X-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 51.2 ms
The effects of prealignment for different frame lengths are shown in Fig. 4.6(a), 4.6(b), 4.7(a), 4.7(b) for frame length of 25.6 ms (512 samples), Fig. 4.8(a), 4.8(b), 4.9(a), 4.9(b) for frame length of 102.4 ms (2048 samples), and Fig. 4.10(a), 4.11(b), 4.11(a), 4.11(b) for frame length of 204.8 ms (4096 samples).

Figure 4.6: X-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 25.6 ms
Figure 4.7: Z-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 25.6 ms

Figure 4.8: X-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 102.4 ms
Figure 4.9: Z-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 102.4 ms

Figure 4.10: X-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 204.8 ms
Figure 4.11: Z-dim. side view of a slice of the surface of (a) the nominal SRP-PHAT using a fixed-time frame, and (b) the enhanced SRP-PHAT using a prealigned-time frame of length 204.8 ms

Clearly, using prealignment potentially improves the SRP-PHAT specifically, and any cross-correlation functional in general. More experimental evaluations will be presented in the task of detecting and locating multiple sources in Chapter 7.

The problem is that a prealignment requires different cross-correlation computations for each hypothesized location $\bar{x}_h$. As $Q$, the number of $\bar{x}_h$ evaluated, is often large, the computational cost of the prealignment method is $Q$ times larger than the nominal SRP-PHAT computation at least for the early DSP stage. In this chapter, I propose a reduced-cost method of prealignment.

4.2.2 Prealignment

4.2.2.1 Brute-force method

For a frame $f$ with length of $N$ and $\alpha$-sample advance, I can prealign the microphone signals to the signal at the hypothesized location $\bar{x}_h$ by shifting the microphone sig-
nals by the corresponding time-delays in samples, \( d_u \), from the location \( \vec{x}_h \) to the microphone locations \( \vec{x}_u, u = 1, ..., M \):

\[
\tilde{z}_u[n] = z_u[n + d_u],
\]  

(4.3)

where \( \tilde{z}_u \) is the prealigned signal of microphone \( u \), \( n = f\alpha, ..., f\alpha + N - 1 \), and the time-delay in samples, \( d_u \), is defined as:

\[
d_u \equiv \left| \frac{||\vec{x}_h - \vec{x}_u||}{F_s} \right|, \quad C,
\]  

(4.4)

where \( F_s \) is the sampling frequency, and \( C \) is the speed of sound. Hence, the **brute-force prealignment method** for \( Q \) locations \( \vec{x}_h \) using \( K \) microphone pairs is,

1. **for** \( q = 1 \rightarrow Q \) **do**
2. Calculate \( d_u \) according to Eq. 4.4
3. Align all microphone signals to get \( \tilde{z}_u[n], u = 1, ..., M; n = f\alpha, ..., f\alpha + N - 1 \) as in Eq. 4.3
4. Compute the DFT, \( Z_u[f, r], r = 0, ..., N - 1 \), of all \( \tilde{z}_u[n], u = 1, ..., M \)
5. Do the phase transform (only half of the spectrum is needed):

\[
\bar{Z}_u[f, r] = \frac{Z_u[f, r]}{|Z_u[f, r]|}, \forall r \in \left[ 0, \frac{N}{2} - 1 \right], u \in [1, M],
\]  

(4.5)

6. Compute the phase cross-power spectrum for each pair \( p = \{u, v\}, p = 1, ..., K \):

\[
C_p[f, r] = \bar{Z}_u[f, r]\bar{Z}_v^*[f, r], \forall r \in \left[ 0, \frac{N}{2} - 1 \right]
\]  

(4.6)

7. There might be some mismatches (due to an imprecise speed-of-sound, round-off errors, etc.) in computing the time-delay in samples, \( d_u, u = 1, ..., M \).
Therefore, a “hard” selection of the $0^{th}$-lag for the prealigned GCC-PHAT in the time-domain will be replaced by a “softer” pick in a neighboring range of the $0^{th}$-lag, that is:

$$
\bar{R}_p[0, f] = \max(R_p[q, f]) \text{ s.t. } q \in [-2, 2], \forall p \in [1, K],
$$

(4.7)

where $R_p[q, f]$ is the real-valued $q^{th}$-sample of the IDFT of $\tilde{C}_p[f, r]$ in frame $f$:

$$
R_p[q, f] = \text{Re} \left( \frac{1}{N} \sum_{r=0}^{N-1} \tilde{C}_p[f, r] e^{-j2\pi qr/N} \right), \quad q \in [-2, 2], \quad p \in [1, K]
$$

(4.8)

with $\tilde{C}_p[f, r]$ is defined as,

$$
\tilde{C}_p[f, r] \equiv \begin{cases} 
C_p[f, r] & \forall r \in [0, 1, \ldots, N/2 - 1] \\
C^*_p[f, N - r] & \forall r \in [N/2, N/2 + 1, \ldots, N - 1]
\end{cases}
$$

(4.9)

8: Sum-up the GCC-PHAT of $K$ pairs to get the SRP-PHAT:

$$
\text{SRP}(f, \vec{x}_h^{(q)}) = \frac{1}{K} \sum_{p=1}^{K} \bar{R}_p[0, f]
$$

(4.10)

9: end for

Clearly, this brute-force approach requires a tremendous amount of work ($O(QKN\log(N))$) vs. $O(KN\log(N))$ of the nominal SRP-PHAT computation) since the computation of the entire process is required for each $\vec{x}_h$. To this end, I propose a prealignment method that is based on a look-up table, thus, trading memory for speed, reducing the amount of computation needed.
4.2.2.2 A more efficient prealignment method

For a specific microphone configuration and search volume, there exists a finite, upper-bound $D$ of the time-delays in samples, $d_u$, i.e., $d_u \in [0, ..., D]$. For example, in a typical rectangular room, with microphones on the wall/ceiling surfaces, $D$ would approximately be the main diagonal of the room. Next, I propose creation of a look-up table of the phase values of the prealigned microphone signals for all possible $D + 1$ time-delays.

The DFT of the $d^{th}$-delay, prealigned signal for microphone $u$ in frame $f$ is,

$$Z_u[f, r, d] = \sum_{n=f\alpha+d}^{f\alpha+N-d-1} z_u[n]W_N^{(n-d-f\alpha)r}$$  \hspace{1cm} (4.11)$$

At $d = 0$, the base-DFT is:

$$Z_u[f, r, 0] = \sum_{n=f\alpha}^{f\alpha+N-1} z_u[n]W_N^{(n-f\alpha)r}.$$  \hspace{1cm} (4.12)

where $r \in [0, ..., N - 1]$ is the frequency sample, and $W_N = e^{-j2\pi r/N}$. The 1st-delay DFT ($d = 1$) can be computed from the base-DFT as follows,

$$Z_u[f, r, 1] = \sum_{n=f\alpha+1}^{f\alpha+N} z_u[n]W_N^{(n-f\alpha-1)r}$$

$$= \sum_{n=f\alpha}^{f\alpha+N-1} z_u[n]W_N^{(n-f\alpha-1)r} - z_u[f\alpha]W_N^{-r} + z_u[f\alpha + N]W_N^{(N-1)r}$$

$$= W_N^{-r} \left( \sum_{n=f\alpha}^{f\alpha+N-1} z_u[n]W_N^{(n-f\alpha)r} - z_u[f\alpha]W_N^{0} + z_u[f\alpha + N]W_N^{N_r} \right)$$

$$= W_N^{-r} (Z_u[f, r, 0] - z_u[f\alpha] + z_u[f\alpha + N])$$

$$= W_N^{-r} (Z_u[f, r, 0] + A_1(f, u)),$$  \hspace{1cm} (4.13)

where $A_1(f, u) = -z_u[f\alpha] + z_u[f\alpha + N]$ is a real number for each $f$ and $u$. Similarly,
I can compute the $d^{th}$-delay DFT based on the $(d - 1)^{th}$-DFT:

$$Z_u[f, r, d] = W_N^{-r} \left( Z_u[f, r, d - 1] + A_d(f, u) \right), \quad (4.14)$$

where $A_d(f, u) = -z_u[f \alpha + d - 1] + z_u[f \alpha + N + d - 1]$, $d = 1, ..., D$. Applying the phase transform (PHAT), i.e., removing the magnitude of the DFT to get the unit-magnitude complex spectrum:

$$P_u[f, r, d] = \frac{Z_u[f, r, d]}{|Z_u[f, r, d]|} = e^{j \theta_u[f, r, d]}, \quad (4.15)$$

where $\theta_u[f, r, d]$ is the corresponding phase angle at frequency $r$ and time-delay $d$ at frame $f$. These unit-magnitude “phase” values in rectangular form in the frequency domain are stored in a real array $P$ of size $M \times \frac{N}{2} \times (D + 1) \times 2 = M \times N \times (D + 1)$ (Here I only need to use $r = 0, ..., \frac{N}{2} - 1$, i.e., half of the spectrum). Therefore, for each $x_h$, I can compute the time-delays $d_u, d_v$ for each pair $p = \{u, v\}$, look-up the respective phase values $P_u[f, r, d_u], P_v[f, r, d_v]$ from the stored array $P$. Using these phase values, I can compute the cross-power spectrum of a microphone pair $p = \{u, v\}$ in the frequency-domain as follows,

$$C_p[f, r, d_u, d_v] = P_u[f, r, d_u]P_v^*[f, r, d_v]$$

$$= \left[ \text{Re} \left( P_u[f, r, d_u] \right) \text{Re} \left( P_v[f, r, d_v] \right) + \text{Im} \left( P_u[f, r, d_u] \right) \text{Im} \left( P_v[f, r, d_v] \right) \right] + j \left[ \text{Im} \left( P_u[f, r, d_u] \right) \text{Re} \left( P_v[f, r, d_v] \right) - \text{Re} \left( P_u[f, r, d_u] \right) \text{Im} \left( P_v[f, r, d_v] \right) \right], \quad (4.16)$$

The corresponding GCC-PHAT value in the time-domain at the $0^{th}$-lag, $\tilde{R}_p[0, f, d_u, d_v]$, is computed similarly to Eq. 4.9, 4.8, and 4.7 in the brute-force approach. Hence, my proposed prealignment algorithm for $Q$ points $x_h$ using $K$ pairs in frame $f$
is,

1: Compute the base-DFT ($d = 0$) for $M$ microphones as in Eq. 4.12
2: Compute the $d^{th}$-delay DFT, $d = 1, ..., D$ for $M$ microphones according to Eq. 4.14
3: Compute the unit-magnitude complex spectrum for $M$ microphones as in Eq. 4.15 and store in $P$

4: for $q = 1 \rightarrow Q$ do
5: Calculate $d_u$, $u = 1, ..., M$ according to Eq. 4.4
6: Look-up $P_u[f, r, d_u]$ from $P$ for $u = 1, ..., M$
7: Compute the GCC-PHAT values for $K$ pairs according to Eq. 4.16, 4.9, 4.8, and 4.7
8: Sum-up the GCC-PHAT of $K$ pairs to get the SRP-PHAT:

\[
\text{SRP}(f, \hat{x}_h^{(q)}) = \frac{1}{K} \sum_{p=1}^{K} \hat{R}_p[0, f, d_u, d_v]
\] (4.17)

9: end for

4.2.3 Computational cost of prealignment methods

Table 4.1 shows the approximated computational costs (floating-point operations, flops) of some functions [45] for a standard PC Intel environment used in my algorithms.

<table>
<thead>
<tr>
<th>Functions (real)</th>
<th>Number of flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 addition (ADD)</td>
<td>1</td>
</tr>
<tr>
<td>1 multiplication (MULT)</td>
<td>1</td>
</tr>
<tr>
<td>1 division (DIV)</td>
<td>4</td>
</tr>
<tr>
<td>1 square-root (sqrt)</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.1: Functions and their computational costs
4.2.3.1 Cost of brute-force method

The cost of each step of the brute-force algorithm shown in section 4.2.2.1 for a single frame $f$ is:

- **Step 2**: Calculate the time-delay in samples, $d_u$: $(3\text{MULTs} + 5\text{ADDs} + 1\text{sqrt}) + 1\text{MULT} + 1\text{ round-off} \approx 14\text{ ops/microphone} \rightarrow 14M\text{ ops for all } M\text{ microphones}$

- **Step 3**: Align the microphone signals: 1 op for shifting operation per microphone $\rightarrow M\text{ ops for } M\text{ microphones}$

- **Step 4**: Compute the DFT's of aligned signals $Z_u[f,r]$: $5MN\log_2(N)$ ops for $M$ DFT's of length $N$

- **Step 5**: Do the PHAT: For each $r$ and $u$, calculating the magnitude of $Z_u[f,r]$ costs 7 ops, and a complex division costs 8 ops, hence, $15\frac{N}{2}M = 7.5MN\text{ ops for all values of } r \text{ and } u$

- **Step 6**: Compute the phase cross-power spectrum $C_p[f,r]$: 6 ops for a complex multiplication, hence, $6\frac{N}{2}M = 3MN\text{ ops for all } r \text{ and } u$

- **Step 7**: For each pair $p$, constructing $C_p[f,r]$ in Eq. 4.9 costs only $\frac{N}{2} - 1\text{ ops, hence, } 0.5KN - K\text{ ops for } K\text{ pairs. Computing Eq. 4.8 requires } 4N\text{ ops for each } p \text{ and } q$, hence, $4KN\text{ ops for } K\text{ pairs and 4 values of } q$. Eq. 4.7 costs 4 ops for each $p$, hence, $4K\text{ ops for } K\text{ pairs. Therefore, the total cost of step II.A7 is } 0.5KN - K + 4KN + 4K = 4.5KN + 3K\text{ ops}$

- **Step 8**: Sum up the GCC-PHAT of $K\text{ pairs to get the SRP-PHAT: } (K - 1)\text{ADDs} + 1\text{MULT} \rightarrow K\text{ ops}$

- **Total cost**:

$$\lambda_1 = (14M + M + 5MN\log_2(N) + 7.5MN + 3MN + 4.5KN + 3K + K)Q$$

$$= (5MN\log_2(N) + 4.5KN + 10.5MN + 15M + 4K)Q\text{ ops} \quad (4.18)$$

59
4.2.3.2 Cost of proposed, more efficient method

The cost of each step of the more efficient prealignment algorithm proposed in section 4.2.2.2 for a single frame $f$ is,

- Step 1: Compute the base-DFT for $M$ microphones: $5MN\log_2(N)$ ops

- Step 2: Compute the $d^{th}$-delay DFT's: For each $(u, r, d)$: $1\text{ADD} + 1\text{ complex MULT}$ (or $4\text{MULTs} + 2\text{ADDs}$) $= 7$ ops $\rightarrow 7MND$ ops for $M$ microphones, $N$ frequency samples, and $D$ time-delays

- Step 3: Compute the unit-magnitude complex spectrum of $M$ microphones: For each $(u, r, d)$: $2\text{DIVs} + (2\text{MULTs} + 1\text{ADD} + 1\text{sqrt}) = 15$ ops $\rightarrow 7.5MN(D+1)$ for $M$ microphones, $\frac{N}{2}$ frequency samples, and $(D+1)$ delays

- Step 5: Calculate the time-delay in samples, $d_u$: $14M$ ops

- Step 6: Look up the unit-magnitude complex spectrum $P_u[f, r, d_u]$: For each microphone $u$: 1 op for pulling out the entire row (all frequency samples $r = 0, ..., \frac{N}{2}$) corresponding to time-delay $d_u \rightarrow M$ ops for $M$ microphones

- Step 7: Compute the GCC-PHAT values for $K$ pairs: Computing Eq. 4.16 requires 6 ops for each frequency value $r$, hence, $3N$ ops for all values of $r$. Computing the GCC-PHAT is similar to step 7 in section 4.2.2.1, which requires: $4.5KN + 3K$ ops

- Step 8: Sum up the GCC-PHAT of $K$ pairs to obtain the SRP-PHAT: $K$ ops

**Total cost:**

\[
\lambda_2 = 5MN\log_2(N) + 8MND + 7.5MN(D + 1) + (14M + M + 4.5KN + 3K + K)Q
\]
\[
= 5MN\log_2(N) + 15.5MND + 7.5MN + (4.5KN + 15M + 4K)Q \text{ ops} \quad (4.19)
\]

4.2.3.3 Comparison

In this work, the pairwise SRP-PHAT computation is used. Eq. 4.18 implies $\lambda_1 \approx O(KNQ) + O(MNQ\log_2(N))$, and Eq. 4.19 indicates $\lambda_2 \approx O(KNQ) + O(MND)$. 

60
In this work, the parameters used are \( M = 182, N = 2048, D = 400, K = \frac{M(M-1)}{2} \), and \( Q = 30000 \). Although both \( \lambda_1 \) and \( \lambda_2 \) are dominated by \( O(KNQ) \), the remaining cost of \( \lambda_1 \), \( O(MNQ\log_2(N)) \), is about \( Q \) times larger than that of \( \lambda_2 \). The number of points evaluated in the focal volume, \( Q \), is often large, thus the savings obtained by \( \lambda_2 \) relative to \( \lambda_1 \) can be significant. While the framelength \( (N) \), the number of points evaluated \( (Q) \), and the maximum delay in samples \( (D) \) remain approximately constant for a specific room, the only parameter that varies is the number of microphones used, \( M \). Note that \( K \) is directly related to \( M \). The computational cost saving percentage \( (\frac{\lambda_1 - \lambda_2}{\lambda_1}) \) is plotted versus the number of microphones \( M \) in Fig. 4.12. Fig. 4.12 shows that the computational cost of the proposed prealignment is about 18% lower than that of the brute-force one for \( M = 182 \). This amount of cost saving is not that much due to the fact that the number of pairs increases quickly \( (O(M^2)) \) as the number of microphones, \( M \), increases. The proposed method would reduce the computational cost significantly if just a small number of microphones is used.

A popular microphone-array size is often no more than 16 microphones [105, 4, 115]. Fig. 4.13 shows the computational cost saving percentage versus the number of microphones, where other parameters are kept fixed as follows: \( N = 2048, D = 150 \) (corresponding to the maximum distance between the microphones and the source.

![Figure 4.12: Percent cost saving vs. the number of microphones, M](image-url)
of 2.6m, a typical distance for small-aperture arrays), $Q = 30000$, and $K = \frac{M(M-1)}{2}$.

Clearly, the proposed prealignment method saves at least 58% relative to the brute-force prealignment method when a small-aperture array is used.

In the next chapter, a new, robust functional for multiple-source localization using steered beamforming-based methods is proposed.

Figure 4.13: Percent cost saving vs. the number of microphones, $M$, for a small-aperture array

In the next chapter, a new, robust functional for multiple-source localization using steered beamforming-based methods is proposed.
Chapter 5

New functional for multiple-source localization: The enhanced minimum variance distortionless response using the phase transform (MVDR-PHAT)

In previous chapters, I have discussed the nominal SRP-PHAT functional and the enhanced SRP-PHAT functional. Since SRP-PHAT is essentially a delay-and-sum beamformer weighted by the phase-transform, it does not provide much spatial separation as more sophisticated beamformers, such as the null-steering beamformers do. In this chapter, I propose a new functional using the null-steering beamformer, minimum variance distortionless response (MVDR), weighted by the phase-transform (MVDR-PHAT), for the task of detecting and locating multiple sound sources.
5.1 The traditional MVDR

The minimum variance distortionless response (MVDR), first proposed by Capon [15] for narrow-band spectral estimation is a special case of the Linearly Constrained Minimum Variance (LCMV), also known as Frost beamformer [71]. The basic idea of MVDR beamformer is to retain the signal in the look direction\(^1\) undistorted, while minimizing the power of the signals coming from other directions. In other words, the MVDR is a null-steering beamformer, which puts nulls in every direction except the look (target) direction. This property, from multiple-source localization perspective, is very interesting, since I would like to have a beamforming functional with the capability of spatially suppressing the signals coming from directions (locations) other than the target one. This is important when the target direction (location) is the one of a weak desired source. Therefore, the MVDR can potentially serves as a more robust functional in the task of detecting and locating multiple sources.

5.1.1 Formulation

In this section, I will develop the minimum variance distortionless response (MVDR) beamformer as the optimal beamformer in minimum-mean-square error (MMSE) sense. For some hypothesized location \(\vec{x}_h\), define the steering vector, \(\vec{d}\), which represents the propagation delays and attenuations from the \(M\) microphones to \(\vec{x}_h\) in the frequency domain as:

\[
\vec{d} = [a_0 e^{-j\omega \tau_0}, a_1 e^{-j\omega \tau_1}, \ldots, a_{M-1} e^{-j\omega \tau_{M-1}}]^T,
\]

(5.1)

where \(a_u = l_u^{-1}\), \(l_u\) is the distance from the location of microphone \(u\), \(\vec{x}_u\), to \(\vec{x}_h\), and \(\tau_u\) is the travel time from \(\vec{x}_h\) to \(\vec{x}_u\). For the purpose of presenting the concept, I ignore

\(^1\) The term “direction” implies the far-field assumption usually taken for most beamforming applications.
the effects of reverberation in the formulation. In the frequency domain, the signals received at \( M \) microphones are,

\[
Z(\omega) = \tilde{d}(\omega)S(\omega) + N(\omega), \tag{5.2}
\]

where \( Z(\omega) \) is an \((M \times 1)\) column-vector representing the received microphone signals at \( M \) microphones in the frequency domain:

\[
Z(\omega) = [Z_0(\omega), \ldots, Z_{M-1}(\omega)]^T \tag{5.3}
\]

\( S(\omega) \) is the source signal in the frequency domain, and \( N(\omega) \) is a \((M \times 1)\) vector of additive noise components in \( M \) channels:

\[
N(\omega) = [N_0(\omega), \ldots, N_{M-1}(\omega)]^T \tag{5.4}
\]

Define a weight vector of the beamformer in the frequency domain as,

\[
W \equiv [w_0(\omega), w_1(\omega), \ldots, w_{M-1}(\omega)]^T \tag{5.5}
\]

The beamformed signal can then be obtained as,

\[
Y(\omega) = W^H(\omega)Z(\omega), \tag{5.6}
\]

where “\( H \)” denotes the Hermitian transpose. Eq. 5.2 and Eq. 5.6 yield,

\[
Y(\omega) = W^H(\omega)\left[\tilde{d}(\omega)S(\omega) + N(\omega)\right] = W^H(\omega)\tilde{d}(\omega)S(\omega) + W^H(\omega)N(\omega) \tag{5.7}
\]
The goal of the beamformer is to estimate the original source signal \( S(\omega) \) as closely as possible. In other words, I would like to minimize the error:

\[
\epsilon(\omega) \equiv |Y(\omega) - S(\omega)|
\]  

(5.8)

Hence, from Eq. 5.7, 5.8, the mean square error (MSE) can be obtained as,

\[
\mu(W(\omega)) = E\{\epsilon^2(\omega)\} \\
= E\{|Y(\omega) - S(\omega)|^2\} \\
= E\{\left| \left( W^H(\omega)d(\omega) - 1 \right) S(\omega) + W^H(\omega)N(\omega) \right|^2 \} 
\]  

(5.9)

Define the power spectral density (PSD) of the source signal \( S(\omega) \) as:

\[
\Phi_S(\omega) \equiv E\{S(\omega)S^H(\omega)\} 
\]  

(5.10)

and the cross-power spectral density (CPSD) matrix of the noise signals as,

\[
\Phi_N(\omega) \equiv E\{N(\omega)N^H(\omega)\} \\
= \begin{bmatrix}
N_0(\omega)N_0^H(\omega) & \ldots & N_0(\omega)N_{M-1}^H(\omega) \\
\vdots & \ddots & \vdots \\
N_{M-1}(\omega)N_0^H(\omega) & \ldots & N_{M-1}(\omega)N_{M-1}^H(\omega)
\end{bmatrix}, 
\]  

(5.11)

Assuming that the noise signals and the source signal are uncorrelated, substituting Eq. 5.10, 5.11 into Eq. 5.9 yields:

\[
\mu(W(\omega)) = \left| W^H(\omega)d(\omega) - 1 \right|^2 \Phi_S(\omega) + W^H(\omega)\Phi_N(\omega)W(\omega) 
\]  

(5.12)

The minimum-variance distortionless response (MVDR) beamformer looks to find an optimal weight \( \hat{W}_{(MVDR)} \) under the constraint that the signal in the look direc-
tion is distortionless, i.e., being passed through with unity gain:

$$W^H(\omega)\tilde{d}(\omega) = 1 \quad (5.13)$$

Under such constraint, the minimization of the mean square error $\mu(W(\omega))$ in Eq. 5.12 is equivalent to minimizing the noise power $W^H(\omega)\Phi_N(\omega)W(\omega)$. Hence, the optimal MVDR weight vector is,

$$\tilde{W}(\text{MVDR})(\omega) = \arg\min_W W^H(\omega)\Phi_N(\omega)W(\omega) \quad (5.14)$$

The MVDR, in this perspective, can also be seen as the beamformer that maximizes the signal-to-noise ratio (SNR) as it tries to minimize the noise power while maintaining the possible maximum gain for the desired signal. This also resembles the property of the generalized sidelobe canceller (GSC).

Define the CPSD matrix of the microphone signals, $\Phi_Z$, as,

$$\Phi_Z(\omega) \equiv E\{Z(\omega)Z^H(\omega)\} = \begin{bmatrix}
Z_0(\omega)Z_0^H(\omega) & \ldots & Z_0(\omega)Z_{M-1}^H(\omega) \\
\vdots & \ddots & \vdots \\
Z_{M-1}(\omega)Z_0^H(\omega) & \ldots & Z_{M-1}(\omega)Z_{M-1}^H(\omega)
\end{bmatrix}, \quad (5.15)$$

From Eq. 5.2 and the assumption that the source signal and noise signals are uncorrelated, $\Phi_Z$ can also be obtained as follows:

$$\Phi_Z(\omega) = \tilde{d}(\omega)\Phi_S(\omega)d^H(\omega) + \Phi_N(\omega) \quad (5.16)$$

Multiplying both the left and right hand sides of Eq. 5.16 with vector $W^H(\omega)$ and
then \( \mathbf{W}(\omega) \) and using the unity gain condition in Eq. 5.13 yield:

\[
\mathbf{W}^H(\omega) \Phi_\mathbf{Z}(\omega) \mathbf{W}(\omega) = \left( \mathbf{W}^H(\omega) \bar{d}(\omega) \right) \Phi_S(\omega) \left( \bar{d}^H(\omega) \mathbf{W}(\omega) \right) + \mathbf{W}^H(\omega) \Phi_N(\omega) \mathbf{W}(\omega)
\]

\[
= \Phi_S(\omega) + \mathbf{W}^H(\omega) \Phi_N(\omega) \mathbf{W}(\omega)
\]

(5.17)

Since \( \Phi_S(\omega) \) is a constant, from Eq. 5.14 and 5.17, I have:

\[
\hat{\mathbf{W}}_{(\text{MVDR})}(\omega) = \arg\min_{\mathbf{W}} \mathbf{W}^H(\omega) \Phi_\mathbf{Z}(\omega) \mathbf{W}(\omega)
\]

(5.18)

The optimal weight vector in Eq. 5.18 under the constraint in Eq. 5.13 can be solved by the method of Lagrange multipliers [15, 90] to yield:

\[
\hat{\mathbf{W}}_{(\text{MVDR})}(\omega) = \frac{\Phi_\mathbf{Z}^{-1}(\omega) \bar{d}(\omega)}{\bar{d}^H(\omega) \Phi_\mathbf{Z}^{-1}(\omega) \bar{d}(\omega)}
\]

(5.19)

The corresponding MVDR power at the steering location \( \bar{x}_h \) is,

\[
P_{(\text{MVDR})}(\bar{x}_h) = \frac{1}{\left| \bar{d}^H(\omega) \Phi_\mathbf{Z}^{-1}(\omega) \bar{d}(\omega) \right|}
\]

(5.20)

5.2 The steered enhanced MVDR-PHAT response power

The phase transform (PHAT)[53] is a robust weighting function for cross-correlation-based functionals in reverberant environments, as it has been studied earlier in Chapter 2. From Eq. 5.15, it can be seen that the CPSD matrix, \( \Phi_{\mathbf{Z}} \), is a constituent of the MVDR beamformer, hence, the PHAT idea can be applied to construct a phase
cross-power spectral density matrix,

\[
\Phi_Z(\omega) \equiv \begin{bmatrix}
    \frac{Z_0(\omega)Z_0^*(\omega)}{|Z_0(\omega)Z_0^*(\omega)|} & \cdots & \frac{Z_0(\omega)Z_{M-1}^*(\omega)}{|Z_0(\omega)Z_0^*(\omega)|} \\
    \vdots & \ddots & \vdots \\
    \frac{Z_{M-1}(\omega)Z_0^*(\omega)}{|Z_{M-1}(\omega)Z_0^*(\omega)|} & \cdots & \frac{Z_{M-1}(\omega)Z_{M-1}^*(\omega)}{|Z_{M-1}(\omega)Z_0^*(\omega)|}
\end{bmatrix}
\] (5.21)

When the prealignment enhancement (see chapter 4.2) is applied, the steering vector \(\vec{d}(\omega)\) becomes a constant \((M \times 1)\) vector of propagation attenuation coefficients:

\[
a = [a_0, a_1, \ldots, a_{M-1}]^T,
\] (5.22)

Eq. 5.21 and 5.22 allow us to constitute the steered enhanced MVDR-PHAT response power in the frequency-domain for a hypothesized location \(\vec{x}_h\):

\[
P^{(\omega)}_{(MVDR-PHAT)}(\vec{x}_h) = \frac{1}{a^T \Phi_Z^{-1} a}
\] (5.23)

If the phase cross-power spectrum matrix in Eq. 5.21 is replaced by the phase cross-correlation matrix, i.e., GCC-PHAT matrix, \(\bar{R}_{zz}\), the real-valued steered enhanced MVDR-PHAT response power can be computed as:

\[
P^{(t)}_{(MVDR-PHAT)}(\vec{x}_h) = \frac{1}{a^T \bar{R}_{zz}^{-1} a},
\] (5.24)

where:

\[
\bar{R}_{zz} = \begin{bmatrix}
    \bar{R}_{00}[0] & \cdots & \bar{R}_{0(M-1)}[0] \\
    \vdots & \ddots & \vdots \\
    \bar{R}_{(M-1)0}[0] & \cdots & \bar{R}_{(M-1)(M-1)}[0]
\end{bmatrix},
\] (5.25)

and each element \(\bar{R}_{uv}[0]\) is the GCC-PHAT value of microphone pair \(\{u, v\}\) yielded from a prealignment enhancement as defined in Eq. 4.7 in chapter 4.2.2.2.
In practice, to avoid the singularity problem of the matrix inversion in Eq. 5.23 & 5.24, to alleviate the problem of inadequate number samples to estimate such matrices, and to partially account for the steering mismatch, a small real-valued diagonal loading term $[68, 47, 16, 58]$, $\delta$, is added to the phase cross-power spectrum matrix, $\Phi_Z$, in the frequency-domain case and to the GCC-PHAT matrix, $\bar{R}_{zz}$, in the time-domain case as follows,

$$
\Phi_Z = \Phi_Z + \delta I
$$

$$
\bar{R}_{zz} = \bar{R}_{zz} + \delta I,
$$

(5.26)

where $I$ is the identity matrix. The steered response power defined in Eq. 5.24 is calculated for each hypothesized location $\hat{x}_h$ in the same manner as the SRP-PHAT functional. In the next chapter, two methods for detecting and locating multiple sound sources using just a single frame of data will be presented. Both proposed methods belong to the steered beamforming-based approach, thus, any of the proposed functionals (nominal SRP-PHAT, enhanced SRP-PHAT, or enhanced MVDR-PHAT) can be used as the functional.
Chapter 6

Methods for multiple-source localization using the proposed functionals

Locating multiple talkers using microphone arrays has many applications, such as: teleconferencing, speech enhancement and data acquisition. However, this problem still remains a challenging task especially under high noise and reverberation, and when only a small segment of data is available. I restrict the problem to only a single segment to see if it is feasible to achieve high-quality results over independent frames. However, the proposed output can be used to drive any source tracker. As it has been discussed in Chapter 1.2, there are two main approaches to solve this problem. The multiple-stage approach [60, 70] first estimates the TDOA’s corresponding to multiple sources from the information given by the unmixing matrix of a blind-source separation (BSS) procedure. The final stage is to locate multiple sources from these TDOA estimates. While this approach has been shown to be promising, it requires accumulation of results over a few seconds of speech as in the case of [70], and a significant increase of computational complexity in solving the permutation problem.
of the BSS procedure as the number of sources increases. For the problem of locating simultaneous, multiple sources using only a local segment of data without averaging/tracking, a single-stage, steered beamforming-based approach is a better option. It avoids all the burden of going through a BSS pre-processing and error-prone TDOA estimation in a highly-reverberant environment. In the steered beamforming-based approach, any of the PHAT-based functionals (nominal SRP-PHAT, enhanced SRP-PHAT or enhanced MVDR-PHAT) can be used. In the rest of this chapter, the chosen functional will be called the “PHAT functional”, and the functional value at any hypothesized location will be called the “PHAT value”.

For a single frame of data $f$, assume $K$ point sources at $(x, y, z)$ locations $\vec{s}^{(f)}(k)$, $k \in [1, K]$, are active. The real-valued PHAT functional for the 3-D spatial vector $\vec{x}$, denoted as $P^{(f)}(\vec{x})$ has been fully described in [97, 30]. A typical slice for $P^{(f)}(\vec{x})$ at a fixed height $y = constant$ is shown in Fig. 6.1 for 5 talkers. It is desirable to isolate and estimate exactly $K$ spatially separated peaks of $P^{(f)}(\vec{x})$ at locations $\vec{\lambda}^{(f)}(k)$ such that the set $\vec{\lambda}^{(f)}$ is the same as the true source location set $\vec{s}^{(f)}$. The basic components of both new algorithms for determining $\vec{\lambda}^{(f)}(k)$ are:

1. Evaluate $P^{(f)}(\vec{x})$ on a large set of $R$ randomly selected points, keeping the highest $N$ of them.

2. Cluster these $N$ points:

   A. **GMM method**: Use the expectation maximization (EM) algorithm to train a GMM and fit the GMM on the $N$ points. Obtain an estimate of $K$ and the $K$ cluster volumes.

   B. **Region Zeroing (RZ) method**: Find all the points $\vec{d}$ with the highest PHAT values among $N$ points that are $l$-meters apart from each other, where $l$ is a minimal distance between two talkers in real life. Zero the rest of the points. Cluster these high points $\vec{d}$ to obtain $K$ and $K$ cluster volumes.
3. Apply stochastic region contraction (SRC)\cite{30} on each volume to find $\lambda^\cup(k)$, $1 \leq k \leq K$.

![SRP-PHAT 3D illustration for 5 talkers](image)

**Figure 6.1**: SRP-PHAT 3D illustration for 5 talkers

### 6.1 Gaussian Mixture Model (GMM) method

If the number of talkers, $K$, is unknown, I can fit a GMM of $q$ components on $N$ data points for $q = 1, ..., Q$ where $Q > K$. The optimal value of $q$ is chosen such that,

$$\hat{q} = \arg\min_q (\text{IC}) \quad (6.1)$$

where IC can be either the Bayes Information Criterion (BIC)\cite{88} or the Akaike Information Criterion (AIC)\cite{2}. However, this often results in an overfitting, i.e., $\hat{q} \geq K$, because of the fact that there are spurious peaks due to noise. This can be further corrected by a post-processing stage.

Let $V_0$ be the boundary vector of the rectangular focal volume with volume $V_{\text{room}}$ containing the sources. SRC parameters depend considerably on the environment’s conditions, such as the room dimensions, and are defined in\cite{28}. The algorithm is:

**A. Initialization**

1. **Evaluate**: $R$ random points in $V_0$.

2. **Select**: The best $N << R$ points.
B. GMM method

1. Given: \( Q > K \), for \( q = 1 \rightarrow Q \):
   - **Initialize GMM:** Apply k-means clustering to obtain \( q \) initial clusters and \( q \) initial sets of GMM coefficients \((\mu_0(k), \Sigma_0(k), \Pi_0(k))\), \( k = 1, \ldots, q \).
   - **Train GMM:** Use EM algorithm \( \rightarrow \) mixture coefficients: \((\mu_q(k), \Sigma_q(k), \Pi_q(k))\), \( k = 1, \ldots, q \).
   - **Compute:** The BIC (or AIC), denoted as \( IC(q) \).

2. **Select:** the optimal number of GMM components,

   \[ \hat{q} = \arg\min_q (IC) \]

3. **Define:** \( \hat{q} \) clusters (GMM hills) with their corresponding coefficients \((\mu_q(i), \Sigma_q(i), \Pi_q(i))\), \( i = 1, \ldots, \hat{q} \).

C. Post processing

1. **Apply:** SRC in each cluster \( i \) in \( \hat{q} \) to get \( \hat{q} \) location estimates. Boundary vectors for SRC in cluster \( i \) are: \( \mathbf{B}_{\text{lower}}(i) = [\mu_q(i) - \delta_q(i)] \), \( \mathbf{B}_{\text{upper}}(i) = [\mu_q(i) + \delta_q(i)] \), where \( \delta_q(i) \) are the standard deviations derived from \( \Sigma_q(i) \).

2. **Cluster:** \( \hat{q} \) location estimates using Euclidean distance until they are at least \( l \)-meters apart \( \rightarrow C^{GMM} \) clusters.

3. **Select:** the highest point in each cluster \( j \) of \( C^{GMM} \) to get \( C^{GMM} \) location estimates.

4. **Keep:** only \( S^{GMM} \) location estimates, \( s^{GMM} \), where \( S^{GMM} \subseteq C^{GMM} \), such that their PHAT values are above the noise level: \( P(s^{GMM}) > P_{\text{noise}} \).

5. **Output:** \( S^{GMM} \) location estimates, \( s^{GMM} \), as the final source locations and their PHAT values.
Note: It has been shown in [28] that $R = 15000$ and $N = 500$ are reasonable values for the focal area $V_{room}=4m \times 1m \times 6m$.

Fig. 6.2 shows (a) an example 2-D energy map given by 1-cm resolution grid-evaluation of the SRP-PHAT values and (b) the corresponding GMM peaks given by the GMM method. Clearly, the GMM peaks not only correspond correctly to the clusters of high SRP-PHAT valued points of the grid-search surface, but also indicates well the heights of the SRP-PHAT peaks in the grid-search surface.

6.2 Region Zeroing (RZ) method

The region zeroing (RZ) method shares the same Initialization stage (A) as in the GMM method. The RZ algorithm is:
I. RZ method

1. Initialize: \( k = 1 \). Select \( \mathbf{d}^{RZ}(1) \) as the location of the highest peak. For each subsequent \( k \):
   - Zero: region in \( l \)-meters radius of \( \mathbf{d}^{RZ}(k) \).
   - Find: the remaining highest peak \( \rightarrow k = k + 1, \mathbf{d}^{RZ}(k) \).
   - Repeat: until all \( N \) points in the space accounted for. Let \( \hat{k} = k \).

2. Select: \( \hat{M} \) points, \( \mathbf{d}^{RZ}(\hat{m}), \hat{m} = 1, ..., \hat{M} \), which have the PHAT values above the median PHAT value of all \( \mathbf{d}^{RZ}(k) \) points.

II. Post processing

1. Apply: SRC in each volume \( \hat{m} \) (of total \( \hat{M} \) volumes) defined as: \( V_{\hat{m}} \equiv [\mathbf{d}^{RZ}(\hat{m}) \pm d_{tol}] \).

2. Cluster: \( \hat{m} \) location estimates using Euclidean distance until they are at least \( l \)-meters apart \( \rightarrow \mathcal{C}^{RZ} \) clusters.

3. Select: the highest point in each cluster \( j \) of \( \mathcal{C}^{RZ} \) to get \( \mathcal{C}^{RZ} \) location estimates.

4. Keep: only \( S^{RZ} \) location estimates, \( \mathbf{s}^{RZ} \), where \( S^{RZ} \subseteq \mathcal{C}^{RZ} \), such that their PHAT values are above the noise level: \( P(\mathbf{s}^{RZ}) > P_{\text{noise}} \).

5. Output: \( S^{RZ} \) location estimates, \( \mathbf{s}^{RZ} \), as the final source locations and their PHAT values.

Note: A typical “minimal” comfortable distance between two talkers in a room is about \( l = 0.5 \) m. In step II.1, I choose \( d_{tol} = 0.2 \) m.

In the next chapter, I will compare the performance of the nominal SRP-PHAT, enhanced SRP-PHAT, and enhanced MVDR-PHAT when being used as the functionals in the representative multiple-source localization method, region-zeroing (RZ), using real-data from a 10-talker recording experiment.
Chapter 7

Evaluations of multiple-source detection and localization

7.1 The Huge Microphone Array (HMA) and the room

The array system can support up to 512 microphones. The hardware, called the Huge Microphone Array (HMA), has been fully described in [92, 93] and the software and some algorithms in [94, 95, 96]. The front-end hardware has 32 modules, each of which does analog-to-digital conversion (16-bit) for 16 microphones and multiplexes that data onto a fiber optic link to a central processing console. The console has 96 floating point DSP microprocessors that perform a sustained total computation rate of approximately 4 GFlops. All processors have access to all the microphone data and can exchange data with any other processor. Data transfer is directed by a PC workstation and the system output is available as either PC files or as outputs from 12 DACs. The room has dimensions $8m \times 8m \times 3m$ and a reverberation time $T_{60}$ of 450ms. Virtually all the surfaces are hard and the HMA console produces a great deal of fan noise and has substantial low-frequency components. The array itself
has 14 1.34m × 0.67m panels of 3cm thick foam in aluminum frames. Thirty-two omnidirectional electret elements are placed on each panel randomly on the vertices of a 3cm × 3cm grid. The panels are hung high on walls or from the ceiling and surround a portion of the room of about 5m × 6m and there are from 8 to 44 discrete reflections above -20dB in each channel. In this thesis, due to the computational and memory requirements of recording longer speech segments, a full-aperture subset of 182 microphones are used to provide longer recordings in all experiments. Current capacities are that HMA can record only for 4–5 seconds for all 448 microphones, but can record up to about 8 minutes for 182 microphones. The 182 microphones consist of 13 microphones from each of the 14 panels so the microphones are still surrounding the same focal area. In all experiments, frames of 102.4 ms, advancing 25.6 ms with a sampling frequency $F_s = 20$ KHz were used.

### 7.2 The grand ten-talker experiment

This is the first time ever that a grand recording of 10 human talkers was made (with 10 respective close-talking channels) using the HMA. This experiment is an example of a cocktail-party problem, where the “noise” from interferers, reverberant speech from different talkers, and background noise was extremely large. In this experiment, 10 talkers broke up into 3 groups of conversations. Group 1 consisted of 5 talkers standing and carrying out a conversation (except talker 10 who was reciting a poem by himself). Group 2 consisted of 3 talkers sitting around a table and having a conversation. Group 3 had two talkers standing and having a conversation. All talkers spoke in a normal manner and loudness. Fig. 7.1 shows the top view of the room, 10 talkers divided into 3 groups with arrows indicating orientations, and the 182 microphones surrounding the focal volume. The microphone-array data, microphone locations and talker locations are available for download online at http:
7.3 Ground-truth establishment

The task is to detect and locate active talkers in each frame. In order to accurately evaluate the performance of the multiple-source localization algorithms, a ground truth of how many talkers and which talker(s) are active in each frame is needed. This ground truth can be established by hand-labeling where each person’s speech starts and stops in each of the close-talking channels. Utilizing Ken Silverman’s program\(^1\), HMAView (a latest version available for download at http://www.lems.brown.edu/array/download.html#Other_Stuff), one can easily hand label such information on the spectrograms of the close-talking channels. Fig. 7.2 shows a snapshot of a typical hand-labeled spectrogram of the close-talking channel for talker 1.

\(^{1}\)Personal conversation with Ken Silverman
Once the close-talking channels are hand labeled, the next step is to determine which talker(s) are active at the far-field microphone channels. From the measured (known) locations of the sound sources, one can calculate the travel time in samples, $d_u(s_i)$, from a source $s_i$, $i = 1, \ldots, 10$, to each microphone channel $u$, $u = 1, \ldots, 182$, as follows:

$$d_u(s_i) \equiv \frac{||\vec{x}_{s_i} - \vec{x}_u||}{F_s}.$$  \hspace{1cm} (7.1)

A “speech event”, i.e., a segment of the signal where $i^{th}$ talker is active (between the ON flag and the closest OFF flag to the right) in the close-talking channel $i$, can be projected into the far-field microphone channel $u^{th}$ by appropriately shifting by $d_u(s_i)$ samples. While $i^{th}$ talker will be labeled as “active” (i.e., having probability $p_{\text{active}} = 1$) in most of the frames within this “speech event” segment, which do not contain either ON or OFF flags, only the transitional frames (beginning frames, $f_b$, etc.)
which contain only the ON flags, and end frames, \( f_e \), which contain only the OFF flags) in microphone channel \( u \) will need extra investigation. Different microphones correspond to different travel times \( d_u(s_i) \) from a particular source \( i \). Hence, the beginning sample, \( n_b(u, i) \) (the location of the projected ON flag), and end sample, \( n_e(u, i) \) (the location of the projected OFF flag), of the “speech event” for source \( i \) will vary at microphone channels \( u \). A probability measure of a talker’s activity can be computed for the transitional frames and dual-transitional frames (containing both ON and OFF flags) as follows,

\[
p_{\text{active}}(f_b, i, u) = \frac{|L_{N-1}(f_b, u) - n_b(u, i)|}{N} \tag{7.2}
\]

\[
p_{\text{active}}(f_e, i, u) = \frac{|n_e(u, i) - L_0(f_e, u)|}{N}, \tag{7.3}
\]

\[
p_{\text{active}}(f_b, f_e, i, u) = \frac{|n_e(u, i) - n_b(u, i)|}{N}, \tag{7.4}
\]

where \( L_0(f, u), L_{N-1}(f, u) \) are the beginning and end samples of frame \( f \) for microphone \( u \), and \( N \) is the frame length. Basically, the probability in Eq. 7.2 and 7.3 indicates how much of a transitional frame the speech activity takes place. The speech activity probability corresponding to talker \( i \) for a particular frame \( f \) can be computed as follows,

\[
p_{\text{active}}(f, i) = \frac{1}{M} \sum_{u=1}^{M} p_{\text{active}}(f, i, u) \tag{7.5}
\]

Fig. 7.3 shows an example of labeling speech activity for transitional frames. The top of the figure shows the signal at the close-talking channel for talker \( i \), and the bottom shows the signal received at a far-field microphone \( u \) after some propagation delay. The three (orange) overlapped lines represent three consecutive frames containing the observed speech event. The black segment of each line represents the segment of the frame where speech activity is present, and its length is equal to \( N \times p_{\text{active}} \). The first
two frames show the case of computing Eq. 7.3, and the third frame shows the case of computing Eq. 7.2. Note that due to the time delay from source $i$ to microphone $u$, while the 2nd frame has $p_{active} = 1$ in the close-talking channel, it has $p_{active} < 1$ at the far-field microphone $u$.

Figure 7.3: Example of labeling speech activity for transitional frames

Clearly, for talker $i$, all speech frames, i.e., frames that do not contain either ON flags or OFF flags and are within the two flags, will always have probability value 1. Only the transitional frames will have positive, less than 1 probabilities. A sensible threshold, $\Theta_{SAD}$, can be chosen to select transitional frames that have sufficient speech content to be labeled as speech frames. The labeling rule for talker $i$ is:

$$f \equiv \begin{cases} 
\text{active} & p_{active}(f, i) \geq \Theta_{SAD} \\
\text{not active} & p_{active}(f, i) < \Theta_{SAD}
\end{cases}$$  \quad (7.6)
7.4 Performance of the proposed functionals

7.4.1 Improved performance in detecting and enhancing multiple sources of the enhanced SRP-PHAT and MVDR-PHAT functionals over the nominal SRP-PHAT

In this section, the improved performance of the new functional, enhanced MVDR-PHAT, and the enhanced SRP-PHAT functional relative to the nominal SRP-PHAT functional in the tasks of detecting and enhancing the functional peaks corresponding to the true talkers when a large number of talkers are active in a frame will be shown. Using the hand-labeled ground truth, a frame where the most talkers (6 talkers: T1, T2, T4, T6, T9, and T10) are simultaneously active is identified. For presentation purposes, using the microphone-array data of this frame, a slice of the grid search (1-cm resolution) through the average height of the talkers is plotted using the three functionals (nominal SRP-PHAT, enhanced SRP-PHAT, and enhanced MVDR-PHAT). Note that because we have the smallest aperture in the height direction, our slice will have the worst error in 3-D position for the sitting talkers. Fig. 7.4 shows the top views of (a) 6 active talker locations and 182 microphones, and (b) the nominal SRP-PHAT surface, (c) the enhanced SRP-PHAT surface, and (d) the enhanced MVDR-PHAT surface. It is expected that the sitting one(s) would have lower PHAT values, and indeed T9 does. In Fig. 7.4(b), the nominal SRP-PHAT functional can only detect two talkers T6 and T4 clearly, while two talkers T10 and T9 are just somewhat higher than the background level. In Fig. 7.4(c), the enhanced SRP-PHAT functional detects T4, a merged peak of T1 and T2, and T6. T10 is a bit obscure but still distinguishable from the background level. In Fig. 7.4(d), the enhanced MVDR-PHAT clearly detects six separable peaks corresponding to six active talkers. T1 and T2 are actually separated, not a merged peak as in the enhanced SRP-PHAT surface, as one will see in the 3-D views of the three functional surfaces.
in Fig. 8.5.

Figure 7.4: Top views of (a) the ground-truth talkers and 182 microphones and the surfaces of (b) the nominal SRP-PHAT functional, (c) the enhanced SRP-PHAT functional, and (d) the enhanced MVDR-PHAT functional

Fig. 7.5(a) shows that the nominal SRP-PHAT functional surface is not smooth, and T1 and T2 are not detected. Fig. 7.5(b) shows the enhanced SRP-PHAT functional has a smoother surface and enhances the talker peaks. Also, it detects T1 and T2, although these two talkers merge into a large peak. The enhanced MVDR-PHAT functional in Fig. 7.5(c) shows six clear, enhanced peaks with a smooth background. This indicates the improved performance of the enhanced MVDR-PHAT functional
in detecting and enhancing multiple talkers compared to that of the classic nominal
SRP-PHAT and the enhanced SRP-PHAT. In the next section, the performance of
the three functionals in locating multiple sources using the ten-talker recording will
be investigated.
Figure 7.5: 3-D views of the surfaces of (a) the nominal SRP-PHAT functional, (b) the enhanced SRP-PHAT functional, and (c) the enhanced MVDR-PHAT functional.
7.4.2 Improved performance in locating multiple sources of the enhanced SRP-PHAT and MVDR-PHAT over the nominal SRP-PHAT

In this section, the performance of the three functionals in locating multiple sources over 300 frames (7.75 seconds of the ten-talker recording) is evaluated. The multiple-source locationing algorithm employing the three functionals is the region zeroing (RZ) algorithm, presented in Chapter 6.2. Note that the number of active talkers in a frame was being unknown. A 3-D location estimate was considered “correct” if it was within 20-cm of the measured 3-D location of the true source. The size of this allowance is primarily due to the smaller aperture in the height dimension, $Y$. Also, the talkers participated in the experiment were not completely stationary throughout the experiment, thus, this allowance somewhat compensates for their movements. A frame was labeled as “correct” if it completely detected all active talkers labeled in the ground truth, and if all the location estimates matched the hand-measured locations of the ground-truth talkers, i.e., the ground-truth locations. Note that, in all frames that are evaluated, the largest number of active talkers labeled by the ground truth in a single frame was 6 talkers, and the fewest was 2 talkers. There were 10, 92, 94, 64, and 40 frames having 2, 3, 4, 5, and 6 active talkers, respectively. Errors resulted when an algorithm-derived location did not match the ground-truth (“extra”) or when the ground-truth locations were missed (“missed”). The following percentage measures were computed:

- Percent correct frames ($\% \text{corr}_{fr}$): How many “correct” frames over all frames
- Percent correct estimates ($\% \text{corr}_{est}$): How many “correct” location estimates over all estimates and all talkers
- Percent missed estimates ($\% \text{missed}_{est}$): How many “missed” location estimates
over all estimates and all talkers

- Percent extra estimates ($\% \text{ extra}_{est}$): How many “extra” location estimates over all estimates and all talkers

- Percent correct estimates of $n$-talker frames ($\% \text{ corr}_{n-tk}$): How many “correct” estimates over all estimates of frames having $n$ active talker(s), $n = 2, \ldots, 6$

- Percent missed estimates of $n$-talker frames ($\% \text{ missed}_{n-tk}$): How many “missed” estimates over all estimates of frames having $n$ active talker(s), $n = 2, \ldots, 6$

- Percent extra estimates of $n$-talker frames ($\% \text{ extra}_{n-tk}$): How many “extra” estimates over all estimates of frames having $n$ active talker(s), $n = 2, \ldots, 6$

Table 7.1 summarizes the performances of the three functionals over all frames and all estimates. It can be seen that in general, the enhanced MVDR-PHAT is better than the enhanced SRP-PHAT and the nominal SRP-PHAT in 3 out of 4 performance factors (except the $\% \text{ extra}_{est}$, in which the nominal SRP-PHAT is slightly better), although not by much. This raises the question, is it worth going through extra work in computing the enhanced MVDR-PHAT and enhanced SRP-PHAT to obtain just a marginal gain relative to the nominal SRP-PHAT? This question will be answered by examining the performance relative to the number of active talkers in a frame as shown in Fig. 7.6.

<table>
<thead>
<tr>
<th>Percent</th>
<th>Nominal SRP-PHAT</th>
<th>Enhanced SRP-PHAT</th>
<th>Enhanced MVDR-PHAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$% \text{ corr}_{fr}$</td>
<td>36</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>$% \text{ corr}_{est}$</td>
<td>73.41</td>
<td>73.78</td>
<td>75.79</td>
</tr>
<tr>
<td>$% \text{ missed}_{est}$</td>
<td>21.40</td>
<td>19.62</td>
<td>18.40</td>
</tr>
<tr>
<td>$% \text{ extra}_{est}$</td>
<td>5.43</td>
<td>6.60</td>
<td>5.83</td>
</tr>
</tbody>
</table>

Table 7.1: Multiple source localization performance of the three functionals
In Fig. 7.6, when the number of active talkers is small, i.e., fewer or equal to three, the performance of the nominal SRP-PHAT is actually slightly better than the enhanced SRP-PHAT and MVDR-PHAT. However, when the number of talkers increases (from 4 to 6), the performance of the nominal SRP-PHAT drops while more correct estimates were obtained by using the enhanced SRP-PHAT and the enhanced MVDR-PHAT. This can be explained by the fact that as more talkers are active simultaneously, there is likely to be more weak talkers. These weak talkers were better enhanced by the prealignment of the enhanced SRP-PHAT, and especially by the null-steering beamformer MVDR-PHAT, which nulls out the dominant talkers while aiming at weak ones. Hence, the enhanced SRP-PHAT and MVDR-PHAT do improve the detection and enhancement of the weak talkers compared to the nominal SRP-PHAT. Therefore, these two functionals are preferable when a large number of active talkers is present, whereas a nominal SRP-PHAT is preferred when only one or two talkers are active. However, this observation is made by using only a limited number of frames and from a single experiment. A general confirmation would require much more data and many experiments with more talkers in a variety of room conditions, which will be reserved for future work.
Figure 7.6: Percent correct estimates of the three functionals for $n$-talker frames, $n = 2, \ldots, 6$

In Fig. 7.7, the percent missed estimates of the three functionals for $n$-talker frames ($n = 2, \ldots, 6$) is shown. Again, the enhanced SRP-PHAT and MVDR-PHAT exhibit fewer missed estimates than the nominal SRP-PHAT when the number of active talkers is high (4 to 6).

Figure 7.7: Percent missed estimates of the three functionals for $n$-talker frames, $n = 2, \ldots, 6$
Fig. 7.8 presents the percent extra estimates of the three functionals for $n$-talker frames ($n = 2, \ldots, 6$). Surprisingly, the nominal SRP-PHAT, overall, makes fewer extra estimates than the other two functionals. This shows that the nominal SRP-PHAT is more “conservative” than the other two functionals, and nominally finds fewer talkers with the “optimal” threshold developed for the algorithm of Sec. 6.2.

Figure 7.8: Percent extra estimates of the three functionals for $n$-talker frames, $n = 2, \ldots, 6$

### 7.4.3 Conclusion

In this Chapter, the performance of the three functionals: nominal SRP-PHAT, enhanced SRP-PHAT, and enhanced MVDR-PHAT, in the task of detecting, enhancing and locating multiple sound sources is studied. The following conclusions can be made:

- The enhanced MVDR-PHAT detects and locates ground-truth talkers better than both the enhanced SRP-PHAT and the nominal SRP-PHAT when more than three talkers are simultaneously active in a frame.

- The nominal SRP-PHAT shows good performance when the number of active
talkers is small (fewer than three)

- The nominal SRP-PHAT is more “conservative” than the enhanced SRP-PHAT and MVDR-PHAT. It misses more ground-truth talkers but also rejects more extra estimates than the other functionals

- The performance at this level of the three functionals is not perfect, but is useful as an input to a multiple-source tracking algorithm, such as those described in [104, 112]
Chapter 8

Sound-source separation – An application of the enhanced MVDR-PHAT functional

8.1 Overview of source-separation methods

Sound-source separation aims to extract the speech signal of the desired source from a mixture of signals of multiple sound-sources. As I discussed in chapter 1.2, traditional approaches include:

- Blind source separation (BSS)
- Beamforming (BF)
- Computational auditory scene analysis (CASA)

Perhaps among the three major approaches, blind source separation (BSS) is the most popular since it requires very little information (hence comes the term “blind”) about the mixture of signals, the sources, the sensors, and the environment. Therefore, the applicability of this approach is broad and versatile. In return, since BSS
uses very few resources, it does not take advantage of all available discriminative features into the separation process. In essence, classical BSS uses the statistical properties, i.e., maximizing the “non-Gaussianity” or some measure of statistical independence of the mixture of signals, to separate individual signals. This principle works well for instantaneous mixtures [74, 73], however, for convolutive mixtures as in the case of reverberant speech, there always exists some limitation. Numerous BSS solutions have been proposed for the convolutive case [98, 3, 66, 13]. Most of these solutions tackle the convolutive obstacle by doing the separation in narrow frequency bins. That is, the classical instantaneous BSS solution, i.e., independent component analysis (ICA) is performed separately in each frequency bin. However, the unmixed signals in different frequency bins are not aligned. Hence, this raises the problem of permutation ambiguity, which could cause the recovered signal in the time-domain to contain frequency components from a different source. As the number of sources and sensor channels increase, the order of permutation increases significantly. Hence, the computational complexity is quite large.

Another popular approach is beamforming (BF), which essentially is the process of applying some kind of filter in the time-domain or frequency-domain or in both, i.e., temporal-frequency domain, to the mixture of signals in order to extract individual signals. The most straightforward implementation would be the time-domain delay-and-sum beamforming aimed at the desired source location. More sophisticated beamformers, such as the null-steering beamformers, can be used to gain more spatial separation. Most often, these null-steering beamformers are carried out in narrow frequency-bands [15, 71, 50, 90]. Since beamforming utilizes some spatial information about the sources, it is sometimes combined with the BSS to yield better BSS solutions [82, 107].

Computational auditory scene analysis (CASA) originates from the idea of mimicking the mechanism of the human auditory system in perceiving sounds. It is often
considered in a two-sensor framework so as to resemble human ears. How does a human perceive a mixture of sounds? A mixture of sounds is often broken up into small “elements” (time-frequency points in machine world), and elements are grouped into “sources” based upon some perceptual cues. The cues could be common onset/offset/modulation, harmonicity, spatial location, pitch frequency, etc [36, 108]. In its essence, CASA is somewhat similar to BSS, where small elements of different sources are separated at each time-frequency bin. When spatial information obtained from a beamforming procedure is used as the cue to group the elements, the result is a hybrid between CASA and BF, as reported in [34, 56].

In short, the three approaches, while having unique properties, still exhibit some mutual relations, and can be combined in a supportive manner to boost the separation in a more robust way.

8.2 Proposed source-separation algorithm using the enhanced MVDR-PHAT functional

As one has seen the improved performance of the enhanced MVDR-PHAT functional in detecting and locating multiple, simultaneous sound-sources in chapter 7, a conjecture can be made that the enhanced MVDR-PHAT functional could be a robust, discriminative feature to separate sound-sources. While the SRP-PHAT functional, which is essentially a delay-and-sum beamformer weighted by the phase-transform, has been shown to be quite robust in selecting time-frequency bins of the desired source [56] from the composite time-frequency bins of the delay-and-sum beamformer, it mainly contributes to the separation in the temporal-frequency domain. If the enhanced MVDR-PHAT functional, which is a null-steering beamformer weighted by the phase-transform, is used, an extra dimension of separation in the spatial domain is expected. Hence, I propose a source-separation algorithm using the same CASA
framework of the one proposed in [56] but employing the enhanced MVDR-PHAT functional. Denote $x_s$ as the location of the desired source, $x_{ref}$ as some reference (non-source) location. $\alpha$, $\beta$ are two threshold values, and $G$ is an attenuation factor. The proposed algorithm is,

1: Delay-sum beamforming (DSBF) to $x_s$ and $x_{ref}$, obtain two signals $y^{(x_s)}(n)$ and $y^{(x_{ref})}(n)$, where $n$ is the time-sample index
2: Take the short-time Fourier-transform (STFT) of the two DSBF signals, using 51.2-ms frames, 5-ms advance, and Hamming window, obtain $Y^{(x_s)}(t, f)$ and $Y^{(x_{ref})}(t, f)$, where $t$ is the frame index and $f$ is the frequency index
3: Compute the magnitude, $M^{(x_s)}(t, f)$, and the phase, $\theta^{(x_s)}(t, f)$, for each $(t, f)$ of $Y^{(x_s)}(t, f)$
4: Compute the average magnitude of the silence frames (in the beginning of $Y^{(x_s)}(t, f)$) for each frequency bin $f$, $M^{(x_s)}_{\text{silence}}(f)$
5: Compute the enhanced MVDR-PHAT functional for each time-frequency bin, $(t, f)$, of $Y^{(x_s)}(t, f)$ and $Y^{(x_{ref})}(t, f)$ using Eq. 5.23, obtain $P^{(x_s)}_{\text{MVDR-PHAT}}(t, f)$ and $P^{(x_{ref})}_{\text{MVDR-PHAT}}(t, f)$
6: for each $(t, f)$ do
7: if $\left(\frac{P^{(x_s)}_{\text{MVDR-PHAT}}(t, f)}{P^{(x_{ref})}_{\text{MVDR-PHAT}}(t, f)} > \alpha\right)$ AND $\left(\frac{M^{(x_s)}(t, f)}{M^{(x_s)}_{\text{silence}}(f)} > \beta\right)$ then
8: Keep the $(t, f)$: $\hat{Y}^{(x_s)}(t, f) = Y^{(x_s)}(t, f)$
9: else
10: $\hat{Y}^{(x_s)}(t, f) = GM^{(x_s)}_{\text{silence}}(f)\theta^{(x_s)}(t, f)$
11: end if
12: end for
13: Reconstruct the extracted desired signal in the time-domain: $\hat{y}^{(x_s)}(n) = \text{ISTFT}(\hat{Y}^{(x_s)}(t, f))$

Fig. 8.1 shows the basic idea of the proposed source-separation algorithm using the enhanced MVDR-PHAT functional.
Figure 8.1: Flow-chart of the proposed source-separation algorithm

8.3 Experimental evaluations

8.3.1 Experimental conditions

The HMA and the room having $T_{60} = 450$ ms have been described in chapter 7. In the source-separation experiment, I used two Advent AV009 speakers, each playing out a unique 10-second recording of a talker from TIMIT database [37]. Both speakers faced panels D and E of the HMA, see Fig. 8.2. I used 16 microphone channels from the two panels to evaluate the proposed algorithm, see Fig. 8.2. The recordings, microphones, and the sources are the same as the ones used in [56]. The threshold
values used were $\alpha = 1.8$, $\beta = 2.2$, and $G = \frac{1}{10}$.

![Figure 8.2: The two sources (red dots) and 16 microphones (blue dots): (a) 3-D view and (b) top view](image)

**8.3.2 Quality evaluations**

The proposed algorithm is compared against three baseline algorithms: (1) delay-sum beamformer (DSBF), (2) minimum-variance distortionless response (MVDR) beamformer, and (3) CASA using SRP-PHAT recently proposed in [56].

**8.3.2.1 Listening evaluation**

The first, and probably most informative, evaluation method is direct listening test evaluation. The audio outputs of the mixed signal in a single microphone channel, the three baseline outputs, and the proposed algorithm’s output are available for listening at [http://www.lems.brown.edu/~hdo/AudioSamples.htm](http://www.lems.brown.edu/~hdo/AudioSamples.htm). The performance of the two baselines, DSBF and MVDR, are similar. This has been confirmed in [5] for a reverberant environment. Baseline 3 (CASA + SRP-PHAT) exhibits a better separation than (1) (DSBF) and (2) (MVDR). However, the presence of the interferer
still can be heard in a few places, especially when the desired talker is silent. The proposed algorithm effectively suppresses the interferer, and thus, separates better the desired talker’s speech than CASA + SRP-PHAT.

8.3.2.2 Visual inspection

In this evaluation, the spectrograms of the outputs of the three baselines and the proposed algorithm will be compared against the spectrograms of the clean speech (close-talking) signals of the desired talkers. Fig. 8.3 shows the spectrograms of the mixed signal, clean-speech signal, baseline 1 (DSBF) and 2 (MVDR) for source 1. Fig. 8.4 shows the spectrograms of the mixed signal, clean-speech signal, baseline 3 (CASA+SRP-PHAT) and the proposed algorithm for source 1. While the presence of the interferer (source 2) is clearly observed in baseline 1 and 2 as being seen in Fig. 8.3, baseline 3 significantly eliminates the interferer. However, some artifacts generated by the interferer can still be seen in baseline 3, as pointed out in Fig. 8.4. Also, the harmonic structure of the spectrogram of baseline 3 is not well-defined as the one of the proposed algorithm.
Figure 8.3: Spectrograms of (a) mixed signal, (b) clean-speech signal of source 1, (c) baseline 1 (DSBF), and (d) baseline 2 (MVDR) for source 1
Figure 8.4: Spectrograms of (a) mixed signal, (b) clean-speech signal of source 1, (c) baseline 3 (CASA using SRP-PHAT), and (d) the proposed algorithm for source 1.

Similarly, Fig. 8.5 shows the spectrograms of the mixed signal, clean-speech signal, baseline 1 and 2 for source 2. Fig. 8.6 shows the spectrograms of the mixed signal, clean-speech signal, baseline 3 and the proposed algorithm for source 2.
Figure 8.5: Spectrograms of (a) mixed signal, (b) clean-speech signal of source 2, (c) baseline 1 (DSBF), and (d) baseline 2 (MVDR) for source 2
Figure 8.6: Spectrograms of (a) mixed signal, (b) clean-speech signal of source 2, (c) baseline 3 (CASA using SRP-PHAT), and (d) the proposed algorithm for source 2

8.3.2.3 Quantitative measure

It is often desired to have an objective, quantitative measure of the performance of speech enhancement/noise reduction systems. Numerous objective measures have been proposed to evaluate speech quality: segmental SNR (segSNR), perceptual eval-
uation of speech quality (PESQ) [72, 79], log-likelihood ratio (LLR) and Itakura-Saito (IS) distance measure [75], etc. Recently, Ma and Loizou [62] proposed a new, more robust objective measure (SNRloss) for predicting intelligibility of processed speech data in different SNR conditions and different types of noise. This measure uses a critical-band spectral representation of the clean and separated signals and is based on the measurement of the SNR loss incurred in each critical band after the mixed signal goes through a speech separation algorithm. The proposed measure is flexible in that they can provide different weights to the two types of spectral distortions introduced by enhancement algorithms, namely spectral attenuation and spectral amplification distortions. In this work, I utilize the three measures studied in [46] (segSNR, modified PESQ, LLR) and the newly proposed SNRloss in [62] to evaluate the performance of the source separation algorithms. In all cases, the closer resemblance the separated speech is to the clean speech of the desired talker, the higher the measure value is. Table 8.1 shows the four objective measures for the three baseline and the proposed algorithms when the desired source is source 1.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>SNRloss</th>
<th>segSNR</th>
<th>PESQ</th>
<th>LLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1 (DSBF)</td>
<td>0.9374</td>
<td>-6.0403</td>
<td>2.012</td>
<td>2.2273</td>
</tr>
<tr>
<td>Baseline 2 (MVDR)</td>
<td>0.9388</td>
<td>-6.1517</td>
<td>1.0796</td>
<td>2.4304</td>
</tr>
<tr>
<td>Baseline 3 (CASA+SRP-PHAT)</td>
<td>0.9506</td>
<td>-4.7636</td>
<td>1.9052</td>
<td>2.7935</td>
</tr>
<tr>
<td>Proposed Alg. (CASA + MVDR-PHAT)</td>
<td>0.9550</td>
<td>-4.3379</td>
<td>1.9759</td>
<td>2.9211</td>
</tr>
</tbody>
</table>

Table 8.1: Objective measures of the performance of source separation algorithms for source 1. Highest value in each column in bold type

Similarly, table 8.2 shows the objective measures for source 2.
<table>
<thead>
<tr>
<th>Algorithms</th>
<th>SNRloss</th>
<th>segSNR</th>
<th>PESQ</th>
<th>LLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1 (DSBF)</td>
<td>0.9514</td>
<td>-6.1267</td>
<td>1.9678</td>
<td>2.9730</td>
</tr>
<tr>
<td>Baseline 2 (MVDR)</td>
<td>0.9518</td>
<td>-6.2086</td>
<td>1.9490</td>
<td>2.9984</td>
</tr>
<tr>
<td>Baseline 3 (CASA+SRP-PHAT)</td>
<td>0.9592</td>
<td>-4.9128</td>
<td>1.9943</td>
<td>3.7636</td>
</tr>
<tr>
<td>Proposed Alg. (CASA + MVDR-PHAT)</td>
<td><strong>0.9628</strong></td>
<td><strong>-4.4318</strong></td>
<td><strong>2.0871</strong></td>
<td><strong>3.5029</strong></td>
</tr>
</tbody>
</table>

Table 8.2: Objective measures of the performance of source separation algorithms for source 2. Highest value in each column in bold type

In all cases, the proposed algorithm has higher SNRloss and segSNR measures than all the baselines. While the PESQ (in source 1 case) and the LLR (source 2) measures of the proposed algorithm are not the highest value, they are marginally smaller than the highest values. This again verifies the better separation quality of the proposed algorithm compared to the three baselines. However, this better separation performance comes at a high price of an expensive computational cost \(O(M^3N\log_2(N)T)\), as opposed to \(O(MN\log_2(N)T)\) of the CASA + SRP-PHAT baseline, where \(M\) is the number of microphones, \(N\) is the frame length, and \(T\) is the number of frames), as the proposed algorithm has to do a complex matrix inversion for each time-frequency bin.

### 8.3.2.4 Performance of sound-source separation algorithms versus the number of microphones

In this section, the performance of the three baselines and proposed algorithm with respect to the number of microphones is studied. The SNRloss is chosen as the measure for evaluation. Fig. 8.7 shows the SNRloss scores of the three baseline (DSBF, MVDR, CASA+SRP-PHAT) algorithms and the proposed algorithm as the number of microphones varies when the desired source is source 1, and Fig. 8.8 shows the performance for source 2. Both figures show the proposed algorithm gives better...
SNRloss scores than the three baselines in all cases.

Figure 8.7: SNRloss scores of the three baselines and proposed algorithm as a function of the number of microphone used for source 1. Note: Higher score means better intelligibility/separation

8.4 Conclusion

In this chapter, a new source-separation algorithm has been proposed using the new, robust functional MVDR-PHAT introduced in chapter 5. The proposed algorithm exploits the robust separation in the spatial domain of the null-steering beamformer MVDR-PHAT to further improve the separation in the time-frequency domain given by the CASA procedure. The algorithm was compared against three baselines (delay-sum beamformer (DSBF), minimum-variance distortionless response (MVDR) beamformer, and the recently proposed algorithm, CASA+SRP-PHAT) using a recording of two sources and 16 microphones in a real, adverse room. Experimental results show that:

- DSBF and MVDR exhibit similar performance in most cases, which agrees with previous work in [6]
<table>
<thead>
<tr>
<th>Number of microphones</th>
<th>SNRloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.945</td>
</tr>
<tr>
<td>4</td>
<td>0.95</td>
</tr>
<tr>
<td>6</td>
<td>0.955</td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
</tr>
<tr>
<td>10</td>
<td>0.965</td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
</tr>
<tr>
<td>14</td>
<td>0.975</td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.8: SNRloss scores of the three baselines and proposed algorithm as a function of the number of microphone used for source 2

- The proposed algorithm shows superior performance to all three baselines when using 16 microphones. It effectively eliminates the interferer while the presence of the interferer can still be heard/observed in the three baselines, especially when the desired source is silent.

- The performance of the proposed algorithm saturates to that of the third baseline (CASA+SRP-PHAT) when using very few microphones, for example, two microphones. This can be explained that when only two microphones are available, both of these two algorithms employ the GCC-PHAT value of a only single microphone pair, thus, not much improvement can be achieved (no effective nulls could be formed) by using the proposed algorithm from this limited amount of data.
Chapter 9

Conclusions and Future work

9.1 Conclusions

In this thesis, the following work has been presented:

- An analytic solution of SRP-PHAT that is useful in explaining the robustness of PHAT against reverberation (chapter 2)

- Three computationally efficient SRP-PHAT methods to locate a single sound source (chapter 3)

- A new enhancement (prealignment) to augment SRP-PHAT functional for the task of locating multiple sound sources (chapter 4)

- A new, robust functional, enhanced MVDR-PHAT (MVDR-PHAT with prealignment and using the maximum GCC-PHAT value in a close neighbor of the 0th-lag to account for miscalculation of the travel times and singularity problem of matrix inversion required to compute the steered MVDR response power), for multiple-source localization (chapter 5)

- Two methods of locating multiple-sources using only a single frame of a large-aperture microphone-array data (chapter 6)
• An application of the newly proposed enhanced MVDR-PHAT functional in separating sound sources in an adverse environment (chapter 8)

All the proposed ideas, except the analytic SRP-PHAT solution (because a controlled environment to test different levels of reverberation noise without effects of background noise was needed), were tested using real data recorded using the HMA system and in a real room with hard surfaces having $T_{60} = 450\text{ms}$. In the room, there was also a lot of background noise generated from the ventilation, computer fans, and the HMA console itself.

Experimental results indicated the following:

• The analytic SRP-PHAT solution made under some convenient set of assumptions in a reverberant environment is essentially a sum of many cosines. When steering to the true point-source location, the cosine values corresponding to the true location are all positive, while the remaining cosines add up incoherently and converge to a near-zero value. When steering to a location other than the true point-source one, all cosines add up incoherently to converge to a near-zero value. Thus, the analytic SRP-PHAT solution reasonably predicts enhancement at the true source and attenuation elsewhere

• The three single-source localization SRP-PHAT methods reduce the cost of the brute-force grid search by three orders of magnitude, while maintaining accuracy relative to that of the grid search

• The newly proposed, enhanced MVDR-PHAT functional shows improved performance compared to the enhanced SRP-PHAT functional and the nominal SRP-PHAT functional in detecting, enhancing and locating multiple talkers (more than three talkers) from a preliminary experiment using a 10-talker recording and 182 HMA microphones in a real, adverse room

109
The proposed source separation algorithm, CASA+ MVDR-PHAT, which brings separation in all three dimensions: spatial, temporal and frequency domains, is robust and in a first test, was “better” in several measures than previously introduced method, CASA+SRP-PHAT, and beamformers like MVDR and delay-sum beamformer (DSBF) in separating two sources recorded using 16 microphones in our adverse environment.

While the presented work in this thesis was done for a large-aperture microphone array, it is still applicable to small-aperture arrays, with the exception of the microphone pair selection enhancement.

9.2 Future work

Observing the improved performance of the enhanced MVDR-PHAT, which is essentially a null-steering beamformer, to the tasks of locating and separating multiple sound sources, a question one might ask is how are the performances of other null-steering beamformers in such tasks? It would be interesting to see if the directions (locations) of the interferers are known and if nulls are specifically put in those directions (locations), how the steered beamforming-based functional would behave. While MVDR puts nulls generally in all directions except the target one, other null-steering beamformers, such as the linearly constrained minimum variance (LCMV) [71] and its variations would allow nulls to be put in specific directions (locations). This might enhance the performance of the multiple-source locator and separator. Hence, a further study on such beamformers for the tasks of multiple source localization and separation deems necessary.

Second, the steered response power of the enhanced MVDR-PHAT involves a matrix inversion, see Eq. 5.23, 5.24. This calculation, when using 182 microphones as in the case of Eq. 5.24, requires an inversion of a $181 \times 181$-matrix for $Q$ evaluated.
points, $\hat{z}_b$, in the focal volume, which is very expensive computationally. In the case of Eq. 5.23 used in the proposed source-separation algorithm, the inversion is done for a complex matrix of size $16 \times 16$, where the number of complex matrices per frame is the number of frequency bands used, in our case, it is 512 bands. As the number of frames $f$ increases, the computational complexity also increases. Hence, a “hard” matrix inversion should be avoided if possible, especially if a real-time implementation is the goal. An adaptive estimation of the steered MVDR-PHAT response power is desirable in such case. Many adaptive methods [5, 67, 43] that utilize some form of the matrix inversion lemma [103] have been proposed. These methods can be applied for computing the steered MVDR-PHAT response power in an adaptive manner to speed up the computation.

Third, the multiple-source locationing algorithms presented in this work only use a single frame of data. If data from consecutive frames are available, a more robust and accurate multiple-source locator can be designed to exploit the statistical and/or signal properties of the sources from one frame to another. Statistical approaches like Kalman filtering [69, 40], particle filtering [106, 104] have been used extensively and have shown promising results. These methods, if appropriately incorporated with the localization algorithms and functionals proposed in this thesis, have the potential for improving the performance of the multiple-source locator. Speech-signal properties, such as: fundamental and formant frequencies, rapidity, breathing patterns of individual talkers, so far have not been exploited effectively in the task of tracking multiple sound sources. These properties could be utilized by a speaker-characterization procedure [77] to contribute to the effort of tracking multiple talkers.

Fourth, the proposed sound-source separation algorithm essentially creates a time-frequency masking, using the steered MVDR-PHAT response power to make a “hard” decision on the time-frequency bins (either keep it, i.e., decision = 1, or replace it by an appropriately scaled silence spectrum, i.e., decision = 0). While this scheme has been
shown to work effectively in our experiments, it might not be that effective when more
than two sound sources need to be separated. In such cases, a hard decision would
usually result in more distorted speech. A “softer” decision, e.g., a probabilistic one,
might be a better option in that case. This can be done by computing the probability
of the presence of the desired talker in a time-frequency bin, using the probabilities
in neighboring time-frequency bins. Maybe a Markov-chain with the states being the
time-frequency points, and the state’s probability is the probability that the desired
talker is active in that point is a potential direction. Also, other perceptual cues, such
as pitch frequencies, can be combined with the MVDR-PHAT functional to group the
time-frequency bins in a more effective way. A hybrid algorithm of the traditional
blind-source separation (BSS) and the MVDR-PHAT is also worth investigating, as it
brings statistical and signal processing properties of the speech signals together into
a single source separator.

Fifth, the thesis presented two major applications of a microphone array: multiple-
source localization and separation. These applications are very essential to a video-
conferencing system, where the locations of the participants, as well as their (high-
quality, unmixed) speech signals are often highly desired. A speaker identification
(speaker ID) (or being able to “label” a particular talker) application would be a
useful functionality to include into and complete the system. My ambitious goal is
to build a video-conferencing system with the source locator, separator introduced in
this work and a source identifier, such as the one presented in [31], if needed.

Finally, all the work presented here was in the context of using a large-aperture
microphone array. However, the applicability and scalability to a smaller array is
feasible. It would be interesting to see the performance of the proposed work when
using a smaller microphone array.
Bibliography


118


posium and USNC/URSI National Radio Science Meeting (AP-S/URSI 2005),
July 2005.


[86] R. Schmidt. A new approach to geometry of range difference location. IEEE

[87] B. Scholling, M. Heckmann, F. Joublin, and C. Goerick. Structuring time do-
main blind source separation algorithms for casa integration. In Proc. ISCA Tu-
torial and Research Workshop on Statistical and Perceptual Audition (SAPA),


[89] A. Sehr and W. Kellermann. Speech and Audio Processing in Adverse Environ-
ments, chapter Towards Robust Distant-Talking Automatic Speech Recognition


2002.


[112] S. Xu, M. Bugallo, and P. Djuric. Maneuvering target tracking with simplified

[113] Z. Yermeche, N. Grbic, and I. Claesson. Speech enhancement of multiple mov-


localization and beamforming for directional microphone arrays in distributed

[116] C. Zhang, D. Florencio, and Z. Zhang. Why does phat work well in low noise,