Audio source localization and tracking in reverberant environments

Li, Tian

2010

Li, T. (2010). Audio source localization and tracking in reverberant environments. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/41621>

<https://doi.org/10.32657/10356/41621>

Audio Source Localization and Tracking in Reverberant Environments

Li Tian

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Master of Engineering

2010

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

Li Tian

Thesis dedication.

To those who care about me.

Acknowledgments

First of all, I would like to express my deepest gratitude to my advisor Prof. Ser Wee for his guidance and support throughout the duration of this work. Since I started this work, he has always given me motivation and encouragement to excel in research. It has been a pleasure to work under his supervision and I would like to maintain this good relationship with him.

Next, I would like to say thanks to Mr. Leng Shuang and Miss. Zhang Xin who gave me valuable help and encouragement during the working hours in the lab, as well as the joyful and non-forgetful memories in the campus life.

Thanks also go to the members of the Center for Signal Processing in the Nanyang Technological University for the useful discussions, support and fun that I have received and enjoyed while working together with them.

Last but not the least, I would like to appreciate the support and loves that originate from my beloved parents and friends. Without whom, I will not be able to come through the challenges and difficult times to finish this work.

Contents

Acknowledgments	i
List of Figures	v
List of Tables	ix
Abstract	x
1 Introduction	1
1.1 Background and Motivation	1
1.2 Major Contributions	3
1.3 Organization of the Thesis	8
2 Reverberant Room Impulse Response Simulation	10
2.1 General Concept of Image Method	11
2.2 Concept of Reverberation Time	17
2.3 Proposed Image Cube Method	18
2.4 Proposed Image Amount Threshold Estimation	20
2.5 Summary	24

3	Conventional Audio Source Localization Techniques	26
3.1	ASL Problem Scenario and System Model	27
3.2	Steered Beamformer based ASL Algorithms	30
3.3	High Resolution Spectral Estimation based ASL Algorithms . .	32
3.4	Time Difference of Arrival (TDOA) based ASL Algorithms . . .	36
3.5	Analysis of SRP-PHAT Algorithm	40
3.6	Summary	56
4	General Concept of Particle Filtering Algorithm	58
4.1	The Bayesian Filtering Problem	59
4.2	Particle Filtering based Bayesian Filtering Solution	62
4.3	The Bootstrap Particle Filter Algorithm	64
4.4	The Sequential Importance Sampling Particle Filter Algorithm .	67
4.5	The Degeneracy Problem	70
4.6	The Imperfect Particle Filtering based ASLT Implementation .	71
4.7	Summary	73
5	Proposed Hybrid Particle Filtering Algorithms	74
5.1	Existing Hybrid Particle Filtering Algorithms	75
5.2	Proposed Hybrid BS-IS Particle Filtering Algorithm	81
5.3	Proposed Statistically Weighted Particle Filtering Algorithm . .	84
5.3.1	The Quad-Stage Statistical Source Height Estimation Al- gorithm	84

<i>Contents</i>	iv
5.3.2 The Statistically Weighted Hybrid Particle Filtering Algorithm	88
5.4 Numerical Results	94
5.5 Summary	111
6 Conclusion and Future Work	114
6.1 Conclusion	114
6.2 Future Work	118
Author's Publications	124
Bibliography	125

List of Figures

2.1	Illustrations of "Direct" Image and Self-Imaged Image	13
2.2	3-D View of Simulated Images Allocation	15
2.3	2-D view of Simulated Images Allocation	15
2.4	Plot of Measured Reverberation Time with Varying Reflection Coefficients and Image Number Coefficients	22
2.5	Room Impulse Response Intensity Points Distribution for $\eta = 32$	23
2.6	Room Impulse Response Intensity Points Distribution for $\eta = 128$	24
3.1	Generic Two-Dimensional Formulation of the Audio Source Lo- calization Problem[10]	27
3.2	M-element Uniform Linear Array (ULA)	28
3.3	Comparison of Spatial Spectrum Resolution of Capon and MUSIC	37
3.4	Percentages of Anomalous TDEs by GCC versus Reverberation Time T_{60} [10]	41
3.5	Two-Dimensional Layouts of the Simulated Reverberant Room and Microphone Array Location	44
3.6	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0$	46

3.7	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.1$	47
3.8	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.2$	48
3.9	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.3$	49
3.10	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.4$	50
3.11	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.5$	51
3.12	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.6$	52
3.13	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.7$	53
3.14	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.8$	54
3.15	Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.9$	55
4.1	Symbolic Representation of One Particle Filtering Iteration [10]	64
5.1	The Logic Flow Chart of Quad-Stage Statistical Source Height Estimation Algorithm	86
5.2	An Example of the Weight Ratio (in log10 scale) and the corre- sponding Statistical Weights	92

5.3	An Illustration of the Target Speech Waveform (a), Simulated Reverberant Room (b, c), Microphone Array (o) Distributions and Acoustic Source Trajectory (arrow)	95
5.4	Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the Iterative SHE Algorithm (b), and the corresponding Binary Hypothesis Option (c) for $RT_{60} = 0.211$ seconds and SNR= 30dB	97
5.5	Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the QSS-SHE Algorithm (b), and the corresponding Statistical Weights (c) for $RT_{60} = 0.211$ seconds and SNR= 30dB	98
5.6	An Example of (x, y) Estimation Square Errors vs Source Height Estimation Square Errors for $RT_{60} = 0.565$ seconds and SNR= 30dB	101
5.7	Ratio of Frames Sharing Consistent Estimation Errors for Different Reverberation Levels	102
5.8	Acoustic Source Localization and Tracking Results Example for $RT_{60} = 0.211$ seconds and SNR= 30dB.	104
5.9	Average Tracking Root Mean Square Error vs Reverberation Time (SNR= 30dB)	107
5.10	Average Track Loss Percentage vs Reverberation Time (SNR= 30dB)	108
5.11	An Illustration of the Target Speech Waveform with Silence Gap (a), Simulated Reverberant Room (b, c), Microphone Array (o) Distributions and Acoustic Source Trajectory (arrow)	109

5.12	Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the QSS-SHE Algorithm (b), and the corresponding Statistical Weights (c) for Source Traveling with Various Height ($RT_{60} = 0.211$ seconds and SNR= 30dB)	112
5.13	Acoustic Source Localization and Tracking Results Example for $RT_{60} = 0.211$ seconds and SNR= 30dB.	113
6.1	SRP-PHAT Power Spectrum and the Corresponding Importance Region Distribution	119
6.2	An Illustration of Source Motion Validation based on Previously Estimated Source Locations	121
6.3	An Illustration of BS-PF Estimation Square Errors with Various $\bar{v}$ values	122

List of Tables

2.1	Measured Reverberation Time with Varying Wall Reflection Coefficients and Image Number Coefficients	21
-----	---	----

Abstract

Audio source localization and tracking (ASLT) in reverberant environment has been a challenging research goal for many years. The failure of conventional ASLT algorithms can be contributed to the improper modeling of reverberation as non-correlated Gaussian noise, as well as to the lack of consideration of previous estimations. Therefore, in this thesis, the Image Method based reverberant room impulse response simulation is discussed and improved to provide an accurate research basis for reverberation robustness ASLT techniques. The Particle Filtering (PF) algorithm, as an approximation to the Bayesian Filtering problem, is then studied in detail. The PF algorithm in theory provides an optimal solution to involve previous estimations to tackle the reverberation robustness problem. Two novel hybrid PF algorithms with the source height estimation based hypothesis systems are proposed in this thesis. The proposed hybrid PF algorithms outperform the existing PF algorithms in the reverberation robust ASLT implementation.

Chapter 1

Introduction

1.1 Background and Motivation

The M.Eng project aims to develop novel and efficient audio source localization and tracking (ASLT) techniques that can be applied in reverberant environment. With the usage of microphone array technology, the ASLT techniques are involved in numerous applications such as hands-free teleconferencing, automated video camera steering, smart conference and tutorial rooms, etc. Even in the audio quality sensitive fields like speech recognition and hospital surveillance, the sophisticated ASLT algorithms can be applied to help estimate the audio source location information. This provides a basis for more effective application of beamforming techniques, which in turn increases the acquisition quality of input audio signal.

One of the specific research problems associated with the project is that most conventional methods either do not assume reverberant environment or generally assume reverberation as Gaussian noise. However, in practice rever-

beration effects cannot be ignored, and the reverberations of complex speech signals are unlike background noise that can always be assumed as of white and Gaussian. Instead, the reverberation causes a target signal to arrive as multiple signals at the sensor. The multi-path propagated signals have strong correlation with the target signal and are non-Gaussian in nature. The void assumption thus greatly degrades the existing ASLT techniques' performance and restricts their application in reverberant environment. Even only moderate reverberation is already severe enough to prevent those conventional techniques from functioning effectively [30]. Therefore, proper reverberant environment modeling becomes essential in developing practically applicable reverberation robust ASLT techniques.

The other specific research issue arises from the limitation that the conventional ASLT techniques usually neglect the previous estimations made on source locations [10]. Some useful information may therefore be ignored in iteratively updating the current source location estimation as well as in the efforts of providing references in minimizing the reverberation effects. For the latter issue, the judgment of the accuracy and the reliability of the current measurement forms another possible investigation topic. Though some Bayesian filtering based ASLT techniques, for example, the Particle filtering methods [5, 6, 7, 8, 9], have been proposed during the last decade to make use of the previous estimations for enhanced reverberation robustness. The theoretically developed algorithms generally have difficulties in the practical application of fusing the previous estimations and current array measurement simultaneously, or at least adaptively, to estimate the source locations. The reason behind this common issue is that there is no feasible hypothesis system to decide whether the source location should be updated according to the previous estimation or

it should be updated according to the current array measurement. To tackle such issue, we expect to develop novel and applicable approach to judge the accuracy of the array measurement and to build a reliable hypothesis system. Based on the measurement accuracy resulted hypothesis system, the ultimate objective of this thesis is to develop novel ASLT technique that is capable of adaptively involving the previous estimations to better tackle the reverberation caused spurious effects.

1.2 Major Contributions

This thesis contains some original contributions on algorithms developed to solve the ASLT problem in reverberant environments. The main contributions can be classified into two parts - reverberant environments modeling and reverberation robust ASLT technique. Now we will present a brief summary of our works conducted on the two parts.

Reverberant Environments Modeling: Reverberant environment modeling is all about analyzing the sound behavior by simulating the room impulse response (RIP). In the history of acoustics study, there have been numerous techniques developed to analyze the sound behavior inside an enclosed room. Among those techniques, Image Method based approaches have drawn great attentions and frequent applications have been made [1, 2, 3]. For the objective of efficiently simulating the room reverberation, Image Method was adopted by Jont B. Allen and David A. Berkley [4] and generally recognized to be the most typical approach in simulating the perceptual effects of a small reverberant room. In the Image Method theory, the room model to be assumed is a rectan-

gular enclosure with a point-to-point (source-to-microphone) impulse response, or transfer function, calculated using a time domain image expansion method. For the time domain assumption, each image contributes only a pure impulse of calculated strength and delay. All together, the multi-path channel corresponding to the reflective walls could be modeled by the impulse response based on numerous images with specific strengths and delays. Therefore, the multi-path channel can be regarded as a numerous tapped finite impulse response (FIR) filter. The reverberant sample of the speech could be produced by convolving an anechoic (non-reverberant) speech sample with the calculated FIR filter.

Though the Image Method provides a rather direct and easy approach to simulate the reverberant environment impulse response. There is implementation issue due to the large quantity of images needed to simulate highly reverberant situation. Theoretically, the number of images was decided by those contributing to the impulse response. The resulted drawbacks are the increased computational complexities (e.g. the complex estimation of the distances between all the images and the microphone) and increased computational loads (e.g. the enlarged buffering memory size and the lengthened convolution time, etc). To overcome such implementation issues, we proposed an image cube method to quickly estimate the distances for images in large amount. We also proposed a reverberation time measurement based method to find the corresponding image amount threshold for different reverberation levels. A series of simulations has been conducted, the results demonstrated that the impulse responses of the additional (with respect to the threshold) images can be compensated by adding white Gaussian noise. As a result, the image amount threshold successfully reduced the computational loads. The improved implementation of the Image Method not only constitutes one of our major contributions in this thesis

but also provides a comprehensive basis to examine and testify our proposed reverberant robust ASLT techniques.

Reverberation Robust ASLT Technique: ASLT has been an active research area over the past two decades. Indeed many such techniques have been reported in the literature, including the Steered Beam Former (SBF) [11, 12, 29], the Generalized Cross Correlation (GCC) [16] based Time-Delay Estimator (TBE), the high resolution spectral estimation methods (e.g. Capon [13], MUSIC [14]), etc. One main problem faced by such techniques in practice is the reverberation effect encountered in a closed room where a target signal will arrive as multiple signals at the sensor. This makes it difficult, if not impossible, for the basic form of conventional ASLT algorithms to function effectively. One reason for the basic form of such techniques to fail is that they do not take into consideration past estimations made on the source signals and the time-varying nature of the source signals. In order to tackle this problem, the Particle Filter (PF) based tracking approaches have been proposed recently [5, 6, 7, 8, 9]. The PF approach works by applying the sequential Monte Carlo (SMC) method and the hidden Markov chain (HMC) assumption. The source location information is estimated through a collection of particles iteratively sampled from the state space (the enclosed reverberant environment in ASLT case). The PF based ASLT approach then sequentially updates the source location by taking into consideration both current array measurement (observation) and previously estimated locations. Several PF based algorithms have been proposed in the literature, and most of them are based on two general forms. One of the two PF general forms is the Bootstrap (BS) PF algorithm [9]. The BS-PF predicts the current state particles by sampling the previously estimated particles through the transition density function. The BS-PF is thus able to

minimize the reverberation effects for source tracking. However, the limitation of the BS-PF is that it does not make use of the current observations that may be useful in indicating the true source location. Another PF based tracking algorithm is the Sequential Importance Sampling (SIS) PF algorithm [7]. The SIS-PF is capable of locating the source rapidly by sampling new particles from the importance functions defined by the current observation. Yet the limitation of the SIS-PF algorithm is that it makes little use of previous source location estimations in the sampling of new particles. The importance function is very sensitive to the accuracy of the current observation and turns to be less robust against the detrimental effects of reverberation.

To overcome the limitations of the BS-PF and the SIS-PF in the ASLT application, some improved PF structures have been developed in literature, e.g. [7] and [8]. They all share the same basic idea that is to fuse, or at least partially, the BS and the SIS algorithms to combine the reverberation robustness and fast location/relocation abilities. The fusion of different PF algorithms is governed by the hypothesis system based on the accuracy judgement of current state observation. Depending on whether the current state observation originates from true source location or from spurious reverberation effects, the hypothesis system adopts corresponding option to apply suitable PF algorithm. The choices of the current state observation judgement and the hybrid hypothesis system are widely open. In this thesis, we proposed two hybrid PF structures that use the hypothesis system based on the source height estimation (SHE). One of the important assumptions is that the source height is always a constant in the entire ASLT process. The other assumption is that any inaccurate height estimation is only caused by the reverberation effects. Therefore, the variation of the current SHE from the (estimated) true source height level serves as a

direct indicator in judging the accuracy of the array's measurement. Based on the judgment and related hypothesis option, the hybrid PF algorithm is able to adaptively apply either the BS-PF or the SIS-PF to tackle the reverberation effects. The first hybrid PF structure (hybrid BS-IS PF) is governed by the binary hypothesis system based on the iterative SHE algorithm. The iterative SHE algorithm estimates the true source height by iteratively updating the mean of height estimations whose differences are within pre-defined threshold. The result of this work has been presented at the 16th *European Signal Processing Conference, EUSIPCO 2008, August. 2008*.

The second hybrid PF structure (SW-PF) that we proposed is governed by the statistical weights (SW) hypothesis system, which is based on the quad-stage statistical (QSS) based SHE (QSS-SHE) algorithm. The QSS-SHE algorithm is able to estimate and update the height information in a more responsive and reverberation-robust manner. As a result, the estimated source height information not only serves as the measure of the observation accuracy but also extends the related ASLT estimations from two dimensional (i.e. (x, y) in Cartesian plane) to three dimensional (i.e. (x, y, z) in Cartesian plane). The QSS-SHE algorithm is also able to discard any false observation even if the height of the source changes during the tracking process, making the algorithm more applicable in the practice. Another added novelty is that the hypothesis system of the hybrid PF structure has been upgraded from the binary "all-or-nothing" option system, as in our first proposed one, to a more comprehensive option system based on the SW. The SW are estimated and upgraded based on the added observation accuracy judgement decisions of the QSS-SHE algorithm. The SW based hypothesis system allows the BS-PF and the SIS-PF to be applied simultaneously and fuses the particles processed by both PF algorithms

to maintain high reverberation robustness. Moreover, as the height and the location of the source are estimated using different acoustic sensor arrays, there was argument that reverberation may have different effects on the observations of different arrays. Therefore, in our second proposed work, we also conducted a comprehensive analysis of the relationship between the observations of different acoustic sensor. The results showed that such argument is difficult, if not possible, to be valid. The result of this work has been accepted for publishing in *EURASIP Journal on Signal Processing*.

Beside the two proposed SHE based hybrid PF structures, in this thesis we also present our works on the importance region vector quantization based measurement accuracy judgement, the source motion validation based measurement accuracy judgement, and their associated hypothesis system to govern the corresponding hybrid PF structures. We also conducted a BS-PF based ASLT performance analysis with various defined steady state velocity of source state transition function. These additional works provide the basis for developing novel and effective reverberation robust ASLT techniques in the future study.

1.3 Organization of the Thesis

This thesis is divided into six chapters, and its organization is as follows.

Chapter 2 first discusses the concept of the Image Method based simulation of the reverberant room impulse response. The definition and calculation of theoretical reverberation time is then introduced. Next, we proposed the image cube method and image amount threshold estimation method to solve the computational load and computation complexity issues of implementing the

Image Method.

Chapter 3 first overviews the existing ASLT techniques and classified them into three major categories. The corresponding strength and weakness of each category has been addressed. Then we analyzed the SRP-PHAT algorithm in detail. This is done by introducing its working principle and by comparing its ASLT performance with delay and sum beamforming for various reverberation levels.

Chapter 4 focuses on the description of Particle Filtering algorithm. The Bayesian Filtering problem and its solution is first introduced. The two basic forms of Particle Filtering algorithm: the Bootstrap Particle Filtering and the Sequential Importance Sampling Particle Filtering are then discussed in details. The corresponding limitations and imperfect implementations of the two basic Particle Filtering forms are then analyzed and summarized.

Chapter 5 first introduces the concept of the existing hybrid Particle Filtering structure and the combined *likelihood function*. Then we proposed two source height estimation based hybrid Particle Filtering algorithms, including the corresponding source height estimation algorithm, the hypothesis system development and the summarized Particle Filtering algorithm. Numerical examples and results are also included to demonstrate that the proposed algorithms outperform the existing algorithms in the reverberation robust ASLT implementations.

Chapter 6 finally concludes our research findings in this thesis and presents a few topics for future works.

Chapter 2

Reverberant Room Impulse Response Simulation

To efficiently simulate the room impulse response (RIP) is not an easy task, especially if one needs to bring the reverberations of complex human speech waveform into consideration. From the literature study, we learned that the Image Method provides an easy and direct solution for computer based simulation of small room's acoustic system. Most up-to-date ASLT related researches have adopted the Image Method to examine the proposed algorithms' performance in reverberant environment. In this chapter, we first introduce the general concept of the Image Method and the concept of reverberation time. Then we move on to discuss our proposed image cube method to quickly estimate the distances for large amount of images. Last but not least, we will discuss about our methods and findings of establishing the image amount threshold for different reverberation levels. The corresponding numerical results and analysis are also included.

2.1 General Concept of Image Method

In theory, the room model assumed is a rectangular enclosure with a point-to-point (source-to-microphone) impulse response, or transfer function, calculated using a time domain image expansion method. The number of source images is decided by those contributing to the impulse response, i.e. within a radius given by the speed of sound times the reverberation time. And for the time domain assumption, each image contributes only a pure impulse of calculated strength and delay. All together, the multi-path channel corresponding to the reflective walls could be modeled by the impulse response based on numerous images with specific strengths and delays. Therefore, a reverberant sample of speech could be produced by convolving an anechoic (non-reverberant) speech sample with the calculated impulse response.

If expressed in mathematical form, the pressure wave emitted by a single frequency point source of acceleration in free space is:

$$P(\omega, \ell_s, \ell_m) = \frac{\exp[j\omega(R/c - t)]}{4\pi R} \quad (2.1)$$

where ℓ_s and ℓ_m stand for the locations of the source and the microphone, ω is the signal angular frequency, $R = \|\ell_s - \ell_m\|$ is the Euclidean distance between the source and microphone, and c is the wave propagation speed expressed in meter per second.

Assuming that the source is located inside a reverberant room and that all of the reflective surfaces are rigid (lossless), then the rigid wall boundary conditions may be satisfied by placing an image symmetrically on the far side of the wall. The resulting direct-path plus wall reflected sound waves' pressure

can be modeled as

$$P(\omega, \ell_s, \ell_m) = \left[\frac{\exp[j(\omega/c)R_+]}{4\pi R_+} + \frac{\exp[j(\omega/c)R_-]}{4\pi R_-} \right] \exp(-j\omega t) \quad (2.2)$$

Assuming that the wall is located at the vertical Cartesian surface where $x = 0$. The distance between the microphone and the source, R_- , and the distance between the microphone and the image, R_+ , can be calculated as

$$\begin{aligned} R_- &= \sqrt{(\mathbf{x}_s - \mathbf{x}_m)^2 + (\mathbf{y}_s - \mathbf{y}_m)^2 + (\mathbf{z}_s - \mathbf{z}_m)^2} \\ R_+ &= \sqrt{(\mathbf{x}_s + \mathbf{x}_m)^2 + (\mathbf{y}_s - \mathbf{y}_m)^2 + (\mathbf{z}_s - \mathbf{z}_m)^2} \end{aligned} \quad (2.3)$$

However, the sound wave emitted from a single acoustic source will always hit the walls multiple times. More echoes are thus being created and being received by the microphone. To simulate the complex reverberation case, the entire room may also be considered as being imaged symmetrically about the wall. We define the source's first image, which is being created when the source signal first hit the wall, as the "direct" image. The imaged room walls could create more self-imaged images by treating the "direct" image as the "source". Then the repeatedly reflected sound waves received by the microphone can be regarded as originally emitted from the corresponding self-imaged image. A simple illustration of the imaged room and corresponding self-imaged image is given in Figure 2.1.

If applying the Image Method to all the six surfaces inside the room, the situation would become even more complicated. There would be six imaged boundaries. Each image can get itself imaged onto all the imaged boundaries.

Figure 2.1: Illustrations of "Direct" Image and Self-Imaged Image

In such case, the number of images can go up to infinity. The resulted sound wave pressure at the microphone receiving end may be expressed as:

$$P(\omega, \ell_s, \ell_m) = \sum_{p=1}^8 \sum_{|\mathbf{r}|=-\infty}^{\infty} \frac{\exp[j(\omega/c)|\mathbf{R}_p + \mathbf{R}_r|]}{4\pi|\mathbf{R}_p + \mathbf{R}_r|} \cdot \exp(-j\omega t) \quad (2.4)$$

where $\mathbf{R}_p$ represents the eight vectors given by the eight permutations of

$$\mathbf{R}_p = (\mathbf{x}_s \pm \mathbf{x}_m, \mathbf{y}_s \pm \mathbf{y}_m, \mathbf{z}_s \pm \mathbf{z}_m) \quad (2.5)$$

$\mathbf{r}$ is the integer vector triplet (n, l, m) , denoting how many times the image has been self-imaged, the room dimensions are denoted by $\mathbf{L} \times \mathbf{W} \times \mathbf{H}$, and the self-imaged image's extended distance $\mathbf{R}_r$ is defined as,

$$\mathbf{R}_r = 2 \times (n\mathbf{L}, l\mathbf{W}, m\mathbf{H}) \quad (2.6)$$

The illustrated interpretation of (2.5) and (2.6) can be found in Figure 2.2 and Figure 2.3, where $|\mathbf{r}| \in [0, 1]$ was purposely scaled so that only twenty six images were generated. Among all of the images, the six “direct” images resulting directly from the six surfaces of the room were denoted by crossed marks, left the twenty self-imaged images indicated by circles. The original acoustic source was highlighted with a star. The bordered box in the two dimensional plane view was used to denote the room dimension boundaries respectively.

If the pressure frequency response of (2.4) is Fourier transformed, the RIP function (time domain Green's function) can obtained as,

Figure 2.2: 3-D View of Simulated Images Allocation

Figure 2.3: 2-D view of Simulated Images Allocation

$$P(\omega, \ell_s, \ell_m) = \sum_{p=1}^8 \sum_{|\mathbf{r}|=-\infty}^{\infty} \frac{\delta[t - (|\mathbf{R}_p + \mathbf{R}_r|/c)]}{4\pi|\mathbf{R}_p + \mathbf{R}_r|} \quad (2.7)$$

Furthermore, if the room surfaces are not rigid, meaning that all the waves reflected by the surfaces will be attenuated in pressure, the solution in terms of point images may no longer be exact. However, if adopt the assumption that the pressure wall reflection coefficient β is angle independent, and also that there will be no frequency variations in reflection coefficient. The modified RIP could be expressed as,

$$P(\omega, \ell_s, \ell_m) = \sum_{|\mathbf{p}|=0}^1 \sum_{|\mathbf{r}|=-\infty}^{\infty} \beta_{x1}^{|n-q|} \beta_{x2}^{|n|} \beta_{y1}^{|l-j|} \beta_{y2}^{|l|} \beta_{z1}^{|m-k|} \beta_{z2}^{|m|} \times \frac{\delta[t - (|\mathbf{R}_p + \mathbf{R}_r|/c)]}{4\pi|\mathbf{R}_p + \mathbf{R}_r|} \quad (2.8)$$

where $\mathbf{R}_p$ is now expressed in terms of the integer 3-vector $\mathbf{p} = (q, j, k)$ as,

$$\mathbf{R}_p = (\mathbf{x}_s - \mathbf{x}_m + 2q\mathbf{x}_m, \mathbf{y}_s - \mathbf{y}_m + 2j\mathbf{y}_m, \mathbf{z}_s - \mathbf{z}_m + 2k\mathbf{z}_m) \quad (2.9)$$

And the β are defined in the way that those with subscript 1 are corresponding to walls next to the coordinate origin. While those with subscript 2 are corresponding to the opposing walls. Physically the two sums of (2.8) are over a three-dimensional lattice of points. For $\mathbf{p}$ there are only eight points in the lattice, but for $\mathbf{r}$ there could be infinite. Usually in application or simulation, the number of points decided by $\mathbf{r}$ could be set according to the problem scenario and simulation requirement.

The RIP as expressed in (2.8) can be interpreted as a finite response (FIR)

filter. The number of filter taps are equivalent to the number of images, which is determined by $\mathbf{p}$ and $\mathbf{r}$ (for simulation purpose, the limits of $\mathbf{r}$ are both finite). The weight of each tap is equivalent to the corresponding microphone received sound wave pressure level, which is determined by β , $\mathbf{R_p}$ and $\mathbf{R_r}$. The delays in between the taps are proportional to the sound wave transmission time and speed, i.e. c , $\mathbf{R_p}$ and $\mathbf{R_r}$. Therefore, a reverberant sample of speech could be produced by convolving the clean speech sample with the FIR filter. This constitute the general concept of the Image Method.

2.2 Concept of Reverberation Time

The reverberant sound in an auditorium dies away with time as the sound energy is absorbed by multiple reflections with the surfaces of the room. In a more reflective room, it will take longer for the sound to die away and the room is said to be “live”. In a very absorbent room, the sound will die away quickly and the room will be described as acoustically “dead”. But the time for reverberation to completely die away will depend upon how loud the sound was to begin with, and will also depend upon the acuity of the hearing of the observer. In order to provide a reproducible parameter, a standard reverberation time has been defined as the time for the sound pressure level decays to $-60dB$ below its original level. The reverberation time hence can be denoted as RT_{60} , and can be calculated using the Sabine’s Algorithm [28] as,

$$RT_{60} = k \cdot V/A \quad (2.10)$$

where $k = (24 \cdot \ln 10)/c_{20} = 0.161(\text{metric})$, $c_{20} = 343\text{m/s}$ is the sound transmission speed at 20°C , V denotes the room volume in m^3 , $A = \alpha \cdot S = \sum_{n=1}^6 \alpha_n \cdot S_n$ is the equivalent absorption surface in m^2 , and $\alpha = 1 - \beta^2$ is the wall absorption coefficient.

The RT_{60} approximately indicates the room reverberation level given the wall reflection coefficient and the room dimensions. Moreover, the RT_{60} can be used as a benchmark for our proposed image amount threshold estimations.

2.3 Proposed Image Cube Method

The Image Method successfully adopted the image expansion approach to simulate the RIP. The Euclidean distances between the generated images and the microphone, given the coordinates of the source and room dimensions, can be estimated by $\mathbf{R}_p$ and $\mathbf{R}_r$ as expressed in (2.5) and (2.6). However, once the reverberation level goes up, there will be a large increase in the quantity of images. In such case, the distance estimation becomes a heavy duty in the simulation implementation as one needs to work out all the possible permutations of $\mathbf{R}_r$, i.e. $|\mathbf{r}| \in [-\infty, +\infty]$. Things can be even more complicated if the source is moving inside the enclosed room, as the coordinates will no longer be fixed. To overcome such implementation issues, we proposed an image cube method to quickly calculate the distances for the large quantity of images.

First we need to define a $2 \times 2 \times 2$ basic image cube consisting of the real and imaged room that the source and its seven closest images reside in, i.e. all the permutations of $(\pm \mathbf{x}_s, \pm \mathbf{y}_s, \pm \mathbf{z}_s)$. The distances between the points inside the image cube can be directly calculated using (2.5).

Then we treat the basic image cube as a point source located in the center of Figure 2.2. We can therefore have a bigger image cube consists twenty seven of basic image cubes. In this way, we can easily estimate the distances for $8 \times 27 - 1 = 215$ images by (2.6) with $|\mathbf{r}| \in [0, 1]$, i.e.

$$\mathbf{R} = \mathbf{R}_p + \mathbf{R}_{\mathbf{r}, |\mathbf{r}| \in [0, 1]} \quad (2.11)$$

Intuitively, we could treat the $3 \times 3 \times 3$ image cube as another point source and extend the image cube size to the next level consisting of $8 \times 27^2 - 1 = 5831$ images. The distances for the images can be estimated by

$$\mathbf{R} = \mathbf{R}_p + 3^{(\kappa-1)} \cdot \mathbf{R}_{\mathbf{r}, |\mathbf{r}| \in [0, 1]} \quad (2.12)$$

where κ denotes the cube extension count given the total number of images N , i.e.

$$\kappa = \log_{27} \frac{(N + 1)}{8} \quad (2.13)$$

Consequently, we could easily estimate the distances for any large quantity of images by extending the image cube dimensions repeatedly, i.e. increasing κ . We do not have to worry about the permutations of $\mathbf{R}_{\mathbf{r}}$, because $|\mathbf{r}|$ is fixed to be within $[-1, 1]$ range. In this way, we solved the computational complexity issues for implementing the Image Method simulation.

2.4 Proposed Image Amount Threshold Estimation

Once we have calculated the theoretical reverberation time RT_{60} , the total number of images required by the Image Method to simulate the corresponding RIP can be estimated as,

$$N = \eta \cdot F_s \cdot RT_{60} \quad (2.14)$$

where F_s is the sampling frequency and η is the image number coefficient that varies according to different reverberation levels.

Meanwhile, we have estimated the distances for all the images needed to simulate the room impulse response, we could easily calculate the delay and intensity level of each image relative to the microphone's location. Arrange the intensities in time sequence according to their respective delays, we have already obtained the impulse response at the microphone placed location of the room. Convolution of the impulse response with a pulse signal, we could simulate the reverberant sample of the pulse signal. Consequently, we could measure the time needed for the signal's intensity level to drop to $-60dB$ below its original level. By varying the image number coefficient at one particular reverberation level, we can have different measured reverberation time. If the measured reverberation time converges or approximates the theoretical value of RT_{60} , we conclude that we have estimated the image amount threshold for that particular reverberation level.

The designated simulation was configured as follows: the dimensions of the

simulated rectangular room are $3.83m \times 2.90m \times 2.70m$. The reflection coefficients of the room's surrounding walls, floor and ceiling are set to be uniform. We set the source as fixed, together with a single microphone, at the center of the room. The source produced waveform is just a single pulse with unity intensity at time $t = 0$. The converged (or the maximum) measured reverberation time (RT) for all the combinations of β and η have been summarized in Table 2.1 and plotted in Figure 2.4 respectively.

β	$F_s \cdot RT_{60}$	RT_{60}	RT	η
0	660	0.00825	0.00013	1
0.1	666	0.08329	0.04475	1
0.2	687	0.08589	0.05663	2
0.3	725	0.09061	0.07875	4
0.4	785	0.09816	0.10100	8
0.5	880	0.10994	0.13438	16
0.6	1031	0.12884	0.19013	64
0.7	1293	0.16168	0.25763	128
0.8	1832	0.22904	0.34663	128
0.9	3472	0.43397	0.42888	128

Table 2.1: Measured Reverberation Time with Varying Wall Reflection Coefficients and Image Number Coefficients

From the plotted results for low to moderate reverberation levels, it is obvious that the measured reverberation time can always converge. The converged reverberation time is either close to or beyond the theoretically estimated RT_{60} . The image number coefficient η , at which the measured reverberation time converges, doubles its value with every 0.1 increment of wall reflection coefficient β . However, for the high reverberation levels, i.e. $\beta > 0.7$, the measured reverberation time never converges nor exceeds the RT_{60} . Due to the computation constraint of the hardware used in the simulation, we believe that further increase η to generate more images is not a practical solution. Therefore, we

Figure 2.4: Plot of Measured Reverberation Time with Varying Reflection Coefficients and Image Number Coefficients

turned back to examine the images resulted room impulse response from another perspective. Instead of measuring the time needed for the intensity to fall below the $-60dB$ threshold, we decided to examine the entire intensity distribution along the time delay axis. The highest reverberation level in our case is obtained when we set $\beta = 0.9$. We then compared the estimated reverberation intensity levels based on $\eta = 32$ and $\eta = 128$. The results are illustrated by the log scaled plots in Figure 2.5 and Figure 2.6 respectively.

Based on the plots, we found that the intensity points corresponding to the

Figure 2.5: Room Impulse Response Intensity Points Distribution for $\eta = 32$

first $32 \times F_s \cdot RT_{60}$ images are generally distributed in a non-linear manner for both cases. While for the $\eta = 128$ plot, it displays more decayed intensity points as compared to the $\eta = 32$ plot. The extra intensity points could contribute to the increased reverberation time measurement. However, we also realized that the extra intensity points in $\eta = 128$ plot tend to decay in a linear manner. Moreover, all the extra intensity points fall below the $-30dB$ level. Therefore, we may assume that the additional images created by setting $\eta \geq 32$ will only contribute to relatively weaker intensity levels. We may therefore use white Gaussian noise to mask the multi-path effects contributed by the additional images. In this way, $\eta = 32$ readily presents the image amount threshold for efficiently simulate the highly reverberant RIP. Referring back to the converged

Figure 2.6: Room Impulse Response Intensity Points Distribution for $\eta = 128$

reverberation time measurement and corresponding image number coefficient in Table 2.1, we could draw our conclusion on our proposed image amount threshold, N_{th} .

$$N_{th} = \begin{cases} 4 \times F_s \cdot RT_{60}, & \text{if } \beta < 0.4 \\ 8 \times F_s \cdot RT_{60}, & \text{if } 0.4 \leq \beta < 0.7 \\ 32 \times F_s \cdot RT_{60}, & \text{if } 0.7 \leq \beta \end{cases} \quad (2.15)$$

2.5 Summary

In this chapter we first reviewed the general concept of Image Method for simulating reverberant room's impulse response. Based on the theory, we learned

that the multi-path effect of reverberation can be simulated by convolving the original audio signal with the images resulted tapped FIR filter. The filter tap information is determined by each individual image's corresponding delay and intensity level at the microphone.

In order to efficiently implement the Image Method based computer simulation, we proposed the image cube method and the image amount threshold estimation. The image cube method successfully reduced the computational complexity by quickly calculating the distances between large amount of images and the microphone. While the image amount threshold estimation finds the lowest number of images needed to efficiently simulate different reverberation levels. Therefore, the computational load has been greatly reduced. With such improvements, all the subsequent research studies in this thesis will adopt the Image Method based reverberant environment simulation, which provides a comprehensive basis to examine the reverberation robustness of the existing and proposed ASLT techniques.

Chapter 3

Conventional Audio Source Localization Techniques

The literature history of audio source localization (ASL) based on microphone array signal processing can be traced back since Worldwide War II. Decade by decade, numerous researchers have proposed vast amount of ASL algorithms that may be loosely classified into three categories: (i) locators based on Steered Beamformer [11, 12, 29]; (ii) locators based on High-Resolution Spectral Estimation [13, 14]; and (iii) locators based on Time Difference of Arrival (TDOA) [15, 16]. However, when the ASL is to be implemented in reverberant environment, the traditional ASL algorithms generally fail to offer reliable performance. The contributing factor of the ASL failure can be explained as that the non-correlated assumption of signal and reverberation is no longer valid. Consequently, new reverberation robust ASL algorithms were proposed, and the Steered Response Power (SRP) with Phase Transform (PHAT) algorithm (SRP-PHAT) [21] has drawn lots of attention. The PHAT weighing emphasizes

the Generalized Cross Correlation (GCC) value at the true TDOA value over the spurious peaks due to reverberation and thus effectively whitens the signal. This in turn makes the SRP-PHAT algorithm less sensitive to reverberations and increases the beamformer resolution of the true peak.

3.1 ASL Problem Scenario and System Model

Figure 3.1: Generic Two-Dimensional Formulation of the Audio Source Localization Problem[10]

The ASL problem usually assumes that a single source is present in a reverberant (and possibly noisy) environment. The source is typically defined as a person producing speech signals, or alternatively as a generic source of band-limited white noise. The reverberant environment will be assumed to be a small and regularly shaped enclosure of typical office room dimensions. A series of microphones are distributed at several locations inside the enclosure to ac-

quire (noisy and reverberant) signals in parallel. Furthermore, the microphones' positions are considered to be fixed and known exactly.

Figure 3.2: M-element Uniform Linear Array (ULA)

The ASL system model is usually considered to be based on a collection of M microphones positioned arbitrarily in a multipath environment. Assuming that only a single source is active, the discrete-time signal $s_m(\cdot)$ received at the m_{th} microphone, $m \in \{1, \dots, M\}$, can be expressed as a function of the discrete time variable $k = 1, 2, \dots$, as:

$$s_m(k) = h_{tot,m}(k) \otimes s_s(k) + v(k) \quad (3.1)$$

where $h_{tot,m}(k)$ is the complete impulse response from the source to the m th microphone, $s_s(k)$ is the source signal, $v(k)$ is additive white noise (assumed to be uncorrelated with the source signal and from microphone to microphone), and the symbol $\otimes$ denotes the convolution.

The complete impulse response $h_{tot,m}(k)$ is the result of cascading two filters: the room impulse response and the microphone channel response. The former one characterizes all the possible acoustic paths between the source and microphone locations. It is a function of both source location $\ell_s = [x_s \ y_s]^T$ and microphone location $\ell_m = [x_m \ y_m]^T$. In general, it is reasonable to assume that the room impulse response may be time-invariant for short periods when the source and microphones are spatially fixed. While the microphone channel response accounts for the electrical, mechanical and acoustical properties of the microphone system. In general, the microphone's directivity pattern makes its response a function of its orientation as well as its spatial placement relative to the noise.

As a result, the impulse response from the source to any microphone $h_{tot,m}(k)$ could be divided into direct-path and multi-path terms $h_{rev,m}(k)$,

$$s_m(k) = \frac{1}{4\pi\|\ell_s - \ell_m\|} s_s(k - \tau_m) + s_s(k) \otimes h_{rev,m}(k) + v(k) \quad (3.2)$$

where the direct-path delay from the source to the m th microphone is estimated by

$$\tau_m = \frac{\|\ell_s - \ell_m\|}{c} \quad (3.3)$$

Assuming that the signal received at microphone side is divided into frames

of L samples each, and let the vector $s_m(l)$ denotes the vector of signal received at the m th microphone for a specific time frame l , i.e.

$$s_m(l) = [s_m(lL - L + 1) \dots s_m(lL)] \quad (3.4)$$

Then, stacking the received signal vectors could form the $M \times L$ matrix,

$$Y_l = \begin{bmatrix} s_1(l) \\ \vdots \\ s_M(l) \end{bmatrix} \quad (3.5)$$

which represents all the signal received at the array during time frame l . The matrix Y_l will be referred to as the raw data. The ASL problem is hence to estimate the instantaneous location ℓ_s of the acoustic source from the raw data Y_l for $l = 1, 2, \dots$

3.2 Steered Beamformer based ASL Algorithms

The first ASL category is consisted by those approaches that are purely based upon maximizing the steered response power (SRP) of a beamformer. They are applicable for passive arrays at any situation where the location estimate is derived directly from a filtered, weighted, and summed version of the acoustic signal received by the array. For a uniform linear array shown in Figure 3.2, the frequency averaged array output power and the localization function can be expressed as

$$p(\ell) = \int \left| \sum_{m=1}^M H_m(\ell, \omega) S_m(\omega) \right|^2 d\omega \quad (3.6)$$

$$\hat{\ell}_s = \arg \max_{\ell} p(\ell) \quad (3.7)$$

where the complex valued beamformer weighting term $H_m(\ell, \omega)$ (also known as steering delay factor) at the m th microphone is a function of the steered location ℓ , the m th microphone location ℓ_m , and the signal frequency ω , i.e.

$$\begin{aligned} H_m(\ell, \omega) &= \alpha_m \exp[j\omega c^{-1}(\|\ell - \ell_m\| - d_{ref})] \\ &= \alpha_m \exp(j\omega c^{-1}|\Delta\ell \times (m-1) \times \cos(\phi)|) \end{aligned} \quad (3.8)$$

where α_m is the gain applied to the m th microphone received signal, d_{ref} is the distance from the origin of the Cartesian coordinate system to the array center, and ϕ denotes the Direction of Arrival (DOA) of the desired source.

In case that the gains at all microphones are equally set to be $1/M$, then (3.6), (3.7) and (3.8) would correspond to the simplest Delay-and-Sum beamformer (DSB). The principle states that DSB applies time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. This very basic kind of beamformer is also known as Conventional beamformer or Bartlett beamformer. As the natural extension of classical Fourier-based spectral analysis [11] to sensor array data, the DSB suffers low resolutions in accurately localize the source, especially when background noise and reverberation caused multipath interference both present up.

Therefore there exist the demands of DSB modifications to suppress the environmental influences and increase the beamformer resolution. And these modified beamformers form up the other sub-class of steered beamformer, i.e. the Filter-and-Sum beamformer (FSB), which are more sophisticated beamformers who apply filters to the array signals as well as the time alignment. It was shown in [12] that if the microphone gain is chosen according to the signal gain level of the source at the m th microphone, then the steered beamformer is said to be offering the optimal maximum likelihood solution to locate wideband signals.

In short summary, the steered beamformer based ASL algorithms provide direct means to estimate source location. They possess relatively high reverberation robustness. The steered beamformer based ASL algorithms can be applied to both single source and multiple sources cases. Moreover, these ASL algorithms are applicable to both near-field and far-field situations. However, the steered beamformers based ASL algorithms suffer from two major drawbacks. One is that they need to steer the beam power to every possible location within the considered environment. Therefore the computational load is quite high. Another drawback is that the localization resolution of the steered beamformers is relatively low as compared to the rest ASL algorithms.

3.3 High Resolution Spectral Estimation based ASL Algorithms

The ASL techniques classified under this category are those modern beamforming methods adapted from the fields of high-resolution spectral analysis. They

were based upon the spatio-spectral correlation matrix derived from the signals received at the microphones. When exact knowledge of this matrix is unknown (which is most always the case), it must be estimated from the observed data. In practical, this is usually done via ensemble averaging of the signals over an interval in which the sources and noises are assumed to be statistically stationary and their estimation parameters (locations in the ASL case) are assumed to be fixed.

The mathematical derivation could be started by defining the steering vector similar to (3.8),

$$A(\phi) = [a_1(\phi) \ a_2(\phi) \ \dots \ a_M(\phi)]^T \quad (3.9)$$

where the steering vector $a_m(\phi)$ is defined as,

$$a_m(\phi) = \exp(-j2\pi c^{-1}(m-1)\Delta\ell \cos \phi) \quad (3.10)$$

In associate of the additive noise $v(l)$, and the signal vectors $s_s(l)$, the array observed signal data could be expressed as

$$X(l) = A(\phi)s_s(l) + v(l) \quad (3.11)$$

The ensemble averaging process is then performed in means of the statistical expectation, the spatio-spectral correlation matrix R can thus be estimated as,

$$R = E[X(l)X^H(l)] = A(\phi)E[s_s(l)s_s^H(l)]A^H(\phi) + E[v(l)v^H(l)] = A(\phi)PA^H(\phi) + \sigma^2 I \quad (3.12)$$

where $P = E[s_s(l)s_s^H(l)]$ stands for the signal correlation matrix and $\sigma^2 I = E[v(l)v^H(l)]$ is the noise correlation matrix. The latter correlation structure is a reflection of the spatially white noise having a common variance σ^2 at all microphones and being uncorrelated among all array sensors.

With the estimated spatio-spectral correlation matrix, Capon [13] proposed the Minimum Variance Distortionless Response (MVDR) Filter, which tries to minimize the power contributed by noise and signals coming from all directions other than the DOA. This could be viewed as a sharp spatial bandpass filter and generalized as,

$$\min P_{Capon}(\phi) = \frac{A^H(\phi)RA(\phi)}{A^H(\phi)A(\phi)} \text{ subject to } W^H A(\phi) = 1 \quad (3.13)$$

Solving this constrained optimization equation gives:

$$W_{Capon} = \frac{R^{-1}A(\phi)}{A^H(\phi)R^{-1}A(\phi)} \text{ and } P_{Capon}(\phi) = \frac{1}{A^H(\phi)R^{-1}A(\phi)} \quad (3.14)$$

The DOA of the source signal could then be estimated at $\phi = \arg \max_{\phi} P_{Capon}(\phi)$. Capon's method demonstrated improved capability at locating the target source and was classified as high resolution spectral estimation. However, the resolution class was upgraded into super resolution level in the 1980s with the introduction of Multiple Signal Classification (MUSIC) algorithm [14]. This algorithm makes use of the Eigenvalue decomposed structure of the correlation matrix. It divides the correlation matrix into signal and noise subspaces. Hence, MUSIC is also under the class of so called Sub-Space based methods.

The derivation of MUSIC starts with performing Eigenvalue Decomposition (EVD) to the correlation matrix R , and then it stores the eigenvectors into

the unitary matrix U and forms the diagonal matrix $\Lambda = \text{diag}[\lambda_1 \dots \lambda_M]$ with all the eigenvalues sorted by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M > 0$. In such manner, any vector that is orthogonal to $A(\phi)$ will be an eigenvector of R with the eigenvalue σ^2 . For the single source case, there should be a total of $M - 1$ such independent vectors. And the only eigenvalue corresponding to the single source should be significantly larger than σ^2 . Thus the eigenvalue/eigenvector pairs can be partitioned into noise eigenvectors and signal eigenvectors, the correlation matrix could be decomposed into the form

$$R = U\Lambda U^H = U_s\Lambda_s U_s^H + U_n\Lambda_n U_n^H \quad (3.15)$$

where $\Lambda_n = \sigma^2 I$ is the diagonal matrix of the noise variance σ^2 .

Since all noise eigenvectors are orthogonal to $A(\phi)$, the columns of U_s must span the range space of $A(\phi)$ whereas those of U_n span its orthogonal complement (the null space of $A(\phi)^H$). The projection operations onto these signal and noise subspaces are defined as,

$$\Pi = U_s U_s^H = A(\phi)[A(\phi)^H A(\phi)]^{-1} A(\phi)^H \quad (3.16)$$

$$\Pi^\perp = U_n U_n^H = I - A(\phi)[A(\phi)^H A(\phi)]^{-1} A(\phi)^H \quad (3.17)$$

The orthogonality between the noise eigenvectors and the steering vector containing the DOA information leads to the followed relationship,

$$U_n^H A(\phi) = 0 \quad (3.18)$$

Therefore the spatial spectrum of MUSIC could finally be obtained as,

$$P_{MUSIC}(\phi) = \frac{A^H(\phi)A(\phi)}{A^H(\phi)U_n U_n^H A(\phi)} = \frac{A^H(\phi)A(\phi)}{A^H(\phi)\Pi^\perp A(\phi)} \quad (3.19)$$

The increased resolution capability of MUSIC as compared to Capon's in DOA estimation can be easily figured out by looking into the difference of the peaks resulting from the two spatial spectrums. Consistent simulation settings were used in obtaining Figure 3.3, the single source is positioned at 30° with respect to the uniform linear array's center.

Once the DOA estimations ϕ with respect to more than one ULA has been completed, the spectral estimation based ASL algorithms could draw lines that start from the center of the ULA and extend along ϕ direction. The potential source location could be estimated by finding the intersection points of the lines. However, in the practical implementation, it is quite problematic to average the signals sufficiently while the statistical stationary remains to be valid in short period. Again, the sensitivity to the signal and sensor modeling errors makes the high resolution approaches less robust in giving consistent ASL performance. Also, the signal correlation resulted by the reverberation could severely detriment the Eigenvalue decomposition based analysis of these high resolution based ASL algorithms.

3.4 Time Difference of Arrival (TDOA) based ASL Algorithms

Estimation techniques of the last ASL category generally adopt a two step procedure. First, they partition the microphones into pairs, and attempt to find

Figure 3.3: Comparison of Spatial Spectrum Resolution of Capon and MUSIC

the time delay estimation (TDE) for each pair of microphones. Second, the obtained TDE along with the knowledge of the separated microphone positions are then used to generate hyperbolic curves which are intersected in some optimal sense to arrive at a source location estimate. It is quite obvious that, for effective localization of acoustic source using the TDOA categorized techniques, a robust and accurate TDE is of essential importance. Therefore a variety of algorithms were proposed to compute the time delay, including the Adaptive Eigenvalue Decomposition Algorithm (AEDA) [15], the most well known Generalized Cross-Correlation (GCC) [16] and its later functionally modified variants.

For GCC, the solution of the estimation of the difference in time arrival is much related to the cross correlation of the two signals received by a pair of microphones. The principle states that to find the relative time delay of the arrived signal to the corresponding microphones is equivalent to find out the maximum of the cross correlation of the two signals. Consider the two signals received by the selected microphone pair to be

$$\begin{aligned} s_m(k) &= s_s(k) + v_m(k) \\ s_n(k) &= s_s(k - \tau) + v_n(k) \end{aligned} \tag{3.20}$$

where τ denotes the relative direct path delay. Then applying Fourier Transform to the two signals would give the Fourier pair of

$$\begin{aligned} S_m(f) &= S_s(f) + V_m(f) \\ S_n(f) &= S_s(f) \exp(-j2\pi f\tau) + V_n(f) \end{aligned} \tag{3.21}$$

Thus the cross power spectral density function can be expressed as

$$\begin{aligned}
G_{mn} &= S_m(f)S_n^*(f) \\
&= S_s(f)S_s^*(f)e^{j2\pi f\tau} + S_s(f)V_n^*(f) + S_s^*(f)e^{j2\pi f\tau}V_m(f) + V_m(f)V_n^*(f)
\end{aligned} \tag{3.22}$$

Since the noise and signals are uncorrelated to each other, thus the last three terms on the right hand side of the cross power spectral density function could be eliminated and left only

$$G_{mn} = S(f)S^*(f)e^{j2\pi f\tau} \tag{3.23}$$

Hence, by taking the Inverse Fourier Transform of G_{mn} , we could derive the Cross Correlation of the two received signal,

$$R_{mn}(k) = F^{-1}[G_{mn}(f)] = F^{-1}[S_s(f)S_s^*(f)e^{j2\pi f\tau}] = R_{SS}(k) \otimes \delta(k+\tau) = R_{SS}(k+\tau) \tag{3.24}$$

It is then quite obvious that the cross-correlation function is just the auto-correlation function of the source signal centered at delay $(-\tau)$. Ideally the cross-correlation function would have the form of a delta function, thus a sharp peak could be located at the relative time delay, i.e.

$$\tau = \arg \max_k R_{mn}(k) \tag{3.25}$$

After obtained the TDE for each microphone pair, the locus of potential source locations in a two-dimensional setting would be a hyperbola. An estimate

of the two-dimensional source location is then given as the location that best fits the potential source loci across all microphone pairs. The least square solution that finds the best fit could be found in [17][18], and the general principle is defined to minimize the sum of the differences to each intersection point of the loci with each other. In mathematical manner, once a total of $N_{int} = P(P-1)/2$ intersection points $\{\ell_{int,i}\}_{i=1}^{N_{int}}$ have been computed (on the basis of the TDEs obtained for P microphone pairs) the estimated source location would result as

$$\hat{\ell}_s = \arg \min_{\ell} \sum_{i=1}^{N_{int}} \|\ell - \ell_{int,i}\| \quad (3.26)$$

Primarily because of their relative low computational load and reasonable robust performance under amicable conditions, the bulk of passive talker localization systems in use today are based on this TDOA category. However, it is also well known that those TDOA based methods are quite sensitive to the level of reverberation in the considered environment. The multi-path effects caused by the echoes certainly violate the uncorrelated assumptions of both interfering signals and background noises. And this fact has already been clearly demonstrated in studies of [10] and [19], one of their simulation result showing the anomalous TDE is quoted here as in Figure 3.4.

3.5 Analysis of SRP-PHAT Algorithm

Before starting to describe the SRP-PHAT algorithm, it is first necessary to address the PHAT weighting function. PHAT is the short form of Phase Transform, and it was originally developed to emphasize the GCC value at the true TDOA value over the undesired local extreme, i.e. the spurious peak(s) due

Figure 3.4: Percentages of Anomalous TDEs by GCC versus Reverberation Time T_{60} [10]

to the reverberation. For realistic acoustical conditions, the PHAT weighting defined in [16] is given by

$$H_{mn}(\omega) = \frac{1}{|S_m(\omega)S_n^*(\omega)|} \quad (3.27)$$

where the cross-power density terms in the denominator are given by (3.22) with the frequency indicator changes from f to ω . With respect to GCC-based TDE, the PHAT weighting has been found to perform considerably better than its counterparts designed to be statistically optimal under specific non-

reverberant, noise conditions. The PHAT weighting whitens the microphone signals to equally emphasize all frequencies. The utility of this strategy and its extension to steered beamforming form the basis of the SRP-PHAT algorithm.

Now recall the frequency averaged power spectrum of the steered beamformer defined in (3.6), this is the simplest Delay-and-Sum beamformer (DSB), one of the direct means for locating the source. The DSB requires shorter analysis intervals and exhibits an elevated insensitivity to environmental conditions, though again, not to a degree that allows for usage under excessive multipath conditions. The filter-and-sum version of the SRP adds flexibility but the design of the filters is typically geared towards optimizing SNR in noise-only conditions and is excessively dependent on knowledge of the signal and channel content. First proposed by [20], the goal of the SRP-PHAT algorithm is to combine the advantages of the steered beamformer for source localization with the signal and condition independent robustness offered by the PHAT weighting. So the SRP of the Filter-and-Sum beamformer with the corresponding multi-channel version of the PHAT weighting is given by

$$P_{PHAT}(\ell) = \int \sum_{m=1}^M \sum_{n=1}^M \frac{S_m(\omega) S_n^*(\omega)}{|S_m(\omega) S_n^*(\omega)|} \exp(j\omega(\tau_m - \tau_n)) d\omega \quad (3.28)$$

These are the desired SRP-PHAT filters, and one can easily find that they are analogous to the individual channel filter of

$$H_n(\omega) = \frac{1}{|S_n(\omega)|} \quad (3.29)$$

Alternatively, it may be shown that (3.28) is equivalent to the sum of the GCC's of all possible N -choose-2 microphone pairings. This means that the

SRP of a 2-element array is equivalent to the GCC of those two microphones. Hence, as the number of microphones is increased, SRP naturally extends the GCC method from a pair wise to a multi-microphone technique. With the GCC of the microphone pair received signals defined in (3.24) and applying PHAT weighting, a time-domain version of SRP-PHAT functional can now be expressed as

$$p(\ell) = 2\pi \sum_{m=1}^M \sum_{n=1}^M R_{mn}(\tau_m - \tau_n) \quad (3.30)$$

This is the sum of all possible pair wise GCC permutations which are time shifted by the differences in the steering delays. Included in this summation is the sum of the M autocorrelations, which are the GCC evaluated at a lag of zero. These terms contribute only a DC offset to SRP since they are independent of the steering delays.

Moreover, it has been proved in [21] that given either method of computations, SRP-PHAT localization is performed in a manner similar to the standard SRP-based approaches. Namely, the SRP is maximized over a region of potential source locations. Relative to the search space indicative of the standard SRP approach, the SRP-PHAT functional significantly deemphasizes extraneous peaks and dramatically sharpens the resolution of the true peak. These desirable features result in a decreased sensitivity to noise and reverberations and more precise location estimates than the existing localization methods could offer. Additionally, all of them are achieved using a very short analysis interval, making the SRP-PHAT more applicable in the practical simulations.

To test the localization performance of SRP-PHAT in reverberant environment, a rectangular small room with dimensions of $4m \times 3m \times 3m$ was proposed.

And the entire (x, y) Cartesian plain was defined into grids of $1\text{cm} \times 1\text{cm}$ segmentation. The reflection coefficients of the surrounding walls, roof and ceilings were assumed to be of uniform value. The sound wave transmission speed inside the room was also assumed to be constant at 343m/s .

Figure 3.5: Two-Dimensional Layouts of the Simulated Reverberant Room and Microphone Array Location

For the two-dimensional localization purpose, two uniform linear microphone arrays each of four Omni directional microphones were used. As depicted by the small circles in Figure 3.5, the first microphone array was mounted on the 4 meters side wall at a height of 1.5m, and the inter element spacing was set to be of 0.8m. While for the other microphone array, it was mounted on the adjacent 3 meters side wall at the same height, and the inter element spacing reduced to 0.6m. In such organization, we can guarantee that both microphone arrays' center could sit at the same plain surface and at the centers of the mounted

walls.

The audio source signal used in the simulation was a pre-recorded clean male speech yielding a length of about 7.46s. The source was assumed to be fixed at the location of (1, 2) and its height was set to be the same with the microphone arrays height. In the lateral processing, the speech signal was sampled at 8000Hz and split into non-overlapped 123 frames along the source path, resulting in a frame length of 512 samples. The multi-path propagated speech received at each microphone was obtained by convolving these frames of source speech signal with the corresponding impulse response simulated by the Image Method.

In order to gain a better understanding of the ASL performances under various reverberation levels, the wall reflection coefficient β was varied from 0 to 0.9 with 0.1 increment step. This corresponds to a set of reverberation times RT_{60} ranging from 0 to 0.43s. The Delay-and-Sum beamforming (DSB) was implemented together with the SRP-PHAT under consistent settings to compare the ASL performances under reverberant influences. Their output power spectrum for increasing β values are plotted in the followed figures.

By observing the differences of power spectrums obtained by the two ASL, we can easily find that for low reverberation cases ($\beta \leq 0.4$), both SRP-PHAT and DSB can accurately locate the source position. But SRP-PHAT generally results in shaper peaks corresponding to true source location. This means that SRP-PHAT has improved estimation resolution over DSB, which would certainly match with the literature studies. With β further increasing, the localization performance of DSB first starts to degrade by showing noticeable spurious peaks in the power spectrum. While similar degradation only happens for SRP-PHAT

Figure 3.6: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0$

Figure 3.7: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.1$

Figure 3.8: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.2$

Figure 3.9: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.3$

Figure 3.10: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.4$

Figure 3.11: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.5$

Figure 3.12: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.6$

Figure 3.13: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.7$

Figure 3.14: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.8$

Figure 3.15: Output Power Spectrum of DSB (top) and SRP-PHAT (bottom) for $\beta = 0.9$

at higher reverberation levels ($\beta = 0.7$). For the even higher reverberation situations ($\beta > 0.7$), both algorithms are suffering from the multi path effect by demonstrating spurious peaks in the output power spectrum. At the most extreme reverberation case, the source location is very hard to be detected as the largest peak no longer corresponds to the true source location. However, we need to emphasize that even though the SRP-PHAT may not locate the source position very accurately for the high reverberation case in our simulation. We noticed that unlike the DSB resulted power spectrum at the same reverberation situation, where there isn't any peak indication at the source's location, the SRP-PHAT could still have some coarse peaks distributed around the source's location. This would in turn remind the system that the corresponding regions shall not be ignored in the estimation process. The potential indications could play very important roles in the implementation of the acoustic source tracking algorithms that is going to be discussed in the next chapter.

3.6 Summary

In this chapter we first reviewed the problem scenario and system models of ASL. Then we roughly went through the literature about the three categories of conventionally developed ASL algorithms. The basic principles and their respective strengths and weakness were discussed. Later we focused our vision onto the SRP-PHAT, a recently proposed Filter-and-Sum beamforming technique that successfully combined the GCC concept into the SRP methodology. Its novel hybrid structure and theoretically derived strong robustness against reverberation spurious effects drew our attention and some simulations were thus conducted to examine its performance under varying reverberation situations.

From the obtained results, we realized that SRP-PHAT do improve the estimation resolution as compared to the conventional DSB method. Also, SRP-PHAT has demonstrated stronger robustness against the multi-path effects caused by reverberation. Under same reverberant environment, SRP-PHAT could deemphasize the extraneous peaks better than the DSB. Even for the extreme reverberation settings, the SRP-PHAT could still indicate potential source location with certain depressed peaks around the true source located region. This property could be further utilized in the followed acoustic source tracking algorithm study.

Chapter 4

General Concept of Particle Filtering Algorithm

Particle Filtering, is an approximated solution to the Bayesian Filtering problem. Based on the Sequential Monte Carlo (SMC) filtering, the Particle Filter (PF) can be well applied for the non-linear and non-Gaussian cases. The Markov process assumption grants the Particle Filter (PF) the ability to estimate the hidden state variables behind a series of continuous measurement. The PF approach thus greatly matches with the acoustic source localization and tracking (ASLT) problem. As to track a moving acoustic source inside a room means that based on the microphone array obtained data, one needs to sequentially update the hidden source location information in fast and effective manner. Moreover, in the presence of room reverberation, the microphone array received signal is easily corrupted with the reverberations, which generally turns the ASLT problem into non-linear and non-Gaussian nature. The PF approach hence certainly outperforms the traditionally developed ASLT methods in better handling of

the spurious effects caused by the non-linear and non-Gaussian reverberations.

4.1 The Bayesian Filtering Problem

Let the source's instantaneous velocity along the x -axis and y -axis be $\dot{x}_k$ and $\dot{y}_k$ respectively. The source's *state variable* at time step k in the state-space can be expressed as,

$$X_k = [x_k \ y_k \ \dot{x}_k \ \dot{y}_k]^T \quad (4.1)$$

where T denotes the transpose. Please note that the state-space defined in this section focuses on a two dimensional horizontal plane to estimate the source's location (x_k, y_k) . This is to make sure that the derivations of the Bayesian filtering approaches are consistent with the earlier literature works. The height of the source, i.e. the third dimension z_k , will be estimated using different method and discussed in the corresponding section.

With equation (4.1), the objective of the Bayesian filtering based ASLT is now equivalent to estimate the source's state variable (the velocity components may not be emphasized in the current problem formulation). Let the observation variable, i.e. the array's measurement, at the k th time step be denoted as Y_k . Assuming the hidden Markov chain (HMC) model, the ASLT problem can now be formulated by the Anderson and Moore's equations [22].

$$\begin{aligned} X_k &= g(X_{k-1}, u_k) \\ Y_k &= h(X_k, v_k) \end{aligned} \quad (4.2)$$

where $g(\cdot)$ and $h(\cdot)$ are the *transition function* and the *observation function*

respectively. Both functions are not necessarily to be linear, and the noise terms u_k and v_k may not be Gaussian as well.

Let the posteriori probability density function, $p(X_k|Y_{1:k})$, be denoted as PDFp. The Bayesian filtering tackles the ASLT problem by estimating the PDFp with a recursive two stages approach, *prediction* and *update*.

Prediction: Assuming that the PDFp at time $k - 1$ has already been estimated, the prior PDF $p(X_k|Y_{1:k-1})$ of the state at time k can be predicted in the *prediction* stage via the Chapman-Kolmogorov equation [23],

$$p(X_k|Y_{1:k-1}) = \int p(X_k|X_{k-1})p(X_{k-1}|Y_{1:k-1})dX_{k-1} \quad (4.3)$$

where $p(X_k|X_{k-1})$ is the state *transition density*.

Update: The predicted prior PDF is then updated with the current available array observation Y_k via Bayes' rule to obtain the PDFp,

$$p(X_k|Y_{1:k}) \propto p(Y_k|X_k)p(X_k|Y_{1:k-1}) \quad (4.4)$$

where $p(Y_k|X_k)$ denotes the *likelihood function*.

The location of the source at time k , i.e. the *state variable* X_k , can now be estimated by computing the mean or mode of the PDFp. This constitutes the Bayesian filtering based ASLT solution, which, in principle, effectively takes into consideration both the previous source location estimations and current array observation.

For the specific case where the filtering problem is purely linear with additive Gaussian noise, the exact Bayesian solution can be computed analytically. The

resulting posterior PDF is also purely Gaussian and the well-known Kalman filter [22] can be used to propagate its mean and variance over time, conditioned on the measurements.

For the nonlinear and/or non-Gaussian cases, which often occur in practice, the Kalman filter might however fail due to poor assumptions (linearity) in the model representation. Also, the Bayesian solution consists in solving integrals for each update of the posterior density. These integrations are usually impossible to solve analytically, and there exists no closed-form solution for equations (4.3) and (4.4) in practice. Thus, practical implementations of the recursion must inevitably be approximated. Several approximation methods have been proposed as solutions of the nonlinear non-Gaussian filtering problem, including mainly the extended Kalman filter, grid-based algorithms, and Gaussian sum methods. An over view of such suboptimal solutions can be found in [23].

One approximation of special interest is Particle Filtering. The use of Sequential Monte Carlo [24] simulation offers a powerful and flexible numerical approximation of the Bayesian solution, and since the great increase in low cost computational power in the late 80's, particle filtering has been the object of a great deal of research in the literature. This principle is based on a discrete representation of the posterior PDF using a set of state samples with corresponding likelihood weights, effectively replacing integrations with discrete sums in the general Bayesian solution. Moreover, this technique is not restricted to problems involving linear functions and Gaussian noise. It also allows the representation of multi-modal and hence non-Gaussian densities. This is of particular importance in the case of ASLT where the PDFp has to reflect the multi-hypothesis character that each of the peaks or alternatively none of them in the localization

function is due to the true source, with the rest due to the reverberations.

4.2 Particle Filtering based Bayesian Filtering Solution

Particle filtering (PF) algorithm, as a sequential Monte Carlo (SMC) method, approximates the Bayesian filtering solution by representing the PDFp as a set of N samples (particles), $\alpha_k^{(n)}$, $n \in \{1, \dots, N\}$, of the *state variable* X_k with associated weights, $w_k^{(n)}$, $n \in \{1, \dots, N\}$ [23].

$$p(X_k|Y_{1:k}) \approx \sum_{n=1}^N w_k^{(n)} \delta(X_k - \alpha_k^{(n)}) \quad (4.5)$$

where $\delta(\cdot)$ is the Dirac delta function.

For the ASLT problem, because the state-space is defined as the horizontal (x, y) Cartesian plane of the room, therefore each of the particles will represent a grid point of fixed Cartesian coordinate $(x_k^{(n)}, y_k^{(n)})$. At the *prediction* stage of each iterative ASLT process, new set of particles are first sampled (predicted) to indicate a potential state-space region where the target source is located. At the followed *update* stage, based on current array's measurement (observation), each particle is associated with a weight $w_k^{(n)}$ to denote its likelihood of representing the true source location. In this way, the location of the source is estimated by computing the Mean of the approximated PDFp, i.e.

$$\begin{aligned}
\hat{X}_k &= \int X_k p(X_k | Y_{1:k}) dX_k \\
&\approx \int X_k \sum_{n=1}^N w_k^{(n)} \delta(X_k - \alpha_k^{(n)}) dX_k \\
&= \sum_{n=1}^N w_k^{(n)} \alpha_k^{(n)}
\end{aligned} \tag{4.6}$$

In a manner similar to the theoretical Bayesian solution of the filtering problem, the Particle Filtering algorithm relies on the two main steps of *prediction* and *update* to propagate the discrete representation of the posterior PDF from one time step to the next. This type of algorithm hence allows a numerical recursion over time as new observation data become available online. The symbolic representation of this principle is provided in Figure 4.1, where the particles are represented as circles, and the size of which denotes the corresponding likelihood weights. (Please note the different symbolic representation of the particles, χ_k instead of α_k .) The state space corresponds to the horizontal axis. The additional resampling step is due to the fact that in practical simulation the variance of the particle weights tends to go up as time goes by. The resulted degeneracy problem could certainly degrade the estimation accuracy. Hence the resampling step is introduced before the prediction step whenever the weight variances increased to certain threshold [25].

This constitutes the general concept of the PF algorithm and the basic application of the PF algorithm in the ASLT problem. Several PF algorithm based ASLT techniques can be found in the literature. The developed ASLT solutions all share the same iterative procedure of (4.3), (4.4) and (4.6). However, they adopt different approaches to sample new particles and to update particle

Figure 4.1: Symbolic Representation of One Particle Filtering Iteration [10]

weights.

4.3 The Bootstrap Particle Filter Algorithm

The bootstrap particle filter (BS-PF) algorithm, originally developed by Gordon et al. [9], is famous for its low demands of computational power and reduced difficulties of implementation.

In the *prediction* stage, the BS-PF predicts the current state particles based on the priori density function, i.e.

$$\alpha_k^{(n)} \propto p(\alpha_k | Y_{1:k-1}) \quad (4.7)$$

According to (4.4), it is equivalent to select a particle from the previously estimated particle set and propagate the selected particle through the transition

density function, i.e.

$$\alpha_k^{(n)} \propto p(\alpha_k | \alpha_{k-1}) p(\alpha_{k-1} | Y_{1:k-1}) \quad (4.8)$$

This forms a relatively simple and straight particle *prediction* approach. In the ASLT application, the *transition function* that governs the source state dynamics can be used to approximate the transition density function. Considering the Langevin process in [5], the source state *transition function* as defined in (4.2) can be expressed as,

$$\alpha_k = g(\alpha_{k-1}, \mathbf{u}_k) = \begin{bmatrix} 1 & 0 & aT_U & 0 \\ 0 & 1 & 0 & aT_U \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix} \cdot \alpha_{k-1} + \begin{bmatrix} bT_U & 0 \\ 0 & bT_U \\ b & 0 \\ 0 & b \end{bmatrix} \cdot \mathbf{u}_k \quad (4.9)$$

where $\mathbf{u}_k$ is the Gaussian noise variable

$$\mathbf{u}_k \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \quad (4.10)$$

in which $\mathcal{N}(\mu, \Sigma)$ represents the density of a multi-dimensional Gaussian random variable with mean vector μ and covariance matrix Σ , T_U is the time interval between the PF's two consecutive updates, and the other excitation parameters are defined as

$$a = \exp(-\beta T_U) \quad (4.11)$$

$$b = \bar{v}\sqrt{1-a^2} \quad (4.12)$$

where $\bar{v}$ is the steady-state velocity and β is the rate constant.

In the *update* stage, the non-normalized weights of the newly sampled particles will be updated according to the *likelihood function*,

$$\tilde{w}_k^{(n)} \propto p(Y_k | \alpha_k^{(n)}) \quad (4.13)$$

In the ASLT application, the choice of the *likelihood function* is widely open, and a few alternative *likelihood functions* have been applied and tested in [6]. From the results, it is suggested that the *pseudo likelihood function* based on the direct output of the source localization algorithm delivers the best performance. For example, if the steered response with phase transform (SRP-PHAT) method [21] is applied to locate the source. The *pseudo likelihood function* can be represented by the power spectrum of the SRP-PHAT:

$$p(Y_k | \alpha_k^{(n)}) = p(\hat{\theta}) = \int \sum_{l=1}^M \sum_{m=1}^M \frac{S_l(\omega) S_m^*(\omega)}{|S_l(\omega) S_m^*(\omega)|} \exp(j\omega(\tau_l - \tau_m)) d\omega \quad (4.14)$$

where $\hat{\theta}$ is the location estimation corresponding to the SRP-PHAT output, $S_l(\omega)$ and $S_m(\omega)$ are the Fourier transformed signal at the *lth* and *mth* sensor, ω is the angular frequency, the symbol $*$ denotes the complex conjugate, and τ_l and τ_m are the respective direct path delays.

The non-normalized weights are then normalized so that the summation of all the weights equals to unity,

$$w_k^{(n)} = \tilde{w}_k^{(n)} / \sum_{n=1}^N \tilde{w}_k^{(n)} \quad (4.15)$$

The source location at current state can be estimated by computing the Mean of the updated particles according to (4.6).

4.4 The Sequential Importance Sampling Particle Filter Algorithm

Another basic form of the PF algorithm is based on the general concept of Sequential Importance Sampling (SIS) [24], the resulted PF algorithm is named as SIS-PF.

In the *prediction* stage, the SIS-PF samples new particles according to a so-called *importance function*, $q(\alpha_k|Y_{1:k})$:

$$\alpha_k^{(n)} \propto q(\alpha_k|Y_{1:k}) \quad (4.16)$$

The *importance function* can be interpreted as a conditional PDF. Therefore, it can be updated iteratively according to the Bayes' rule,

$$q(\alpha_k|Y_{1:k}) = q(\alpha_{k-1}|Y_{1:k-1})q(\alpha_k|\alpha_{k-1}, Y_k) \quad (4.17)$$

The latter expression in (4.17) suggests that the particles should be pre-

dicted based on both the previously estimated particles and the current state observation. The involvement of Y_k constitutes the most significant improvement of the SIS-PF over the BS-PF. Theoretically, because of the indication of the current available observation, the *importance function* predicted particles are more efficient in representing the true source state.

However, this idealized *prediction* scheme has no general solution in the real time implementation. Alternatively, the application of the *importance function* in the ASLT problem can be interpreted as: rather than a fine-scale and accurate representation of the particle sampling areas, the *importance function* is typically meant to give a coarse indication of where the particles should be sampled in the state-space [7]. The beam pattern of the steered beamforming (SBF) is hence suggested to be a good reference of the *importance function*. As a result, the SRP-PHAT method's output in (4.14) may again be used to approximate the *importance function*, i.e.

$$q(\alpha_k|Y_{1:k}) \approx p(\alpha_k|Y_k) \propto p(Y_k|\alpha_k^{(n)}) = p(\hat{\theta}) \quad (4.18)$$

from which it is quite clear that the previous estimated particles α_{k-1} have been excluded from the approximated *importance function*.

While in the *update* stage, according to [7], [23] and [26], for the approximation of the PDFp in (4.5) to be valid, the computation of the weights must be re-formulated to:

$$\begin{aligned}
\tilde{w}_k^{(n)} &\propto \frac{p(\alpha_k^{(n)}|Y_{1:k})}{q(\alpha_k^{(n)}|Y_{1:k})} \\
&\propto p(Y_k|\alpha_k^{(n)}) \cdot \frac{p(\alpha_k^{(n)}|Y_{1:k-1})}{q(\alpha_k^{(n)}|Y_{1:k})}
\end{aligned} \tag{4.19}$$

The so-called importance weights are hence defined as the product of the *likelihood function* and a correction term that compensates for a potentially uneven distribution of the particles sampled from the *importance function*. The prior PDF in the correction term can be approximated according to (4.3) and (4.8), i.e.

$$p(\alpha_k^{(n)}|Y_{1:k-1}) \approx \sum_{j=1}^N w_{k-1}^{(j)} p(\alpha_k^{(n)}|\alpha_{k-1}^{(j)}) \tag{4.20}$$

where the transition density function can be implemented by the Langevin process defined in (4.9). But some of the *importance function* sampled particles may fall within the null state-space that is not governed by the transition density function. In this case, the corresponding particles will have zero weights. To avoid such problem, the transition density function needs to be updated to

$$p(\alpha_k|\alpha_{k-1}) = (1 - \psi) \cdot g(\alpha_{k-1}, \mathbf{u}_k) + \psi \cdot \mathcal{U}(\alpha_k) \tag{4.21}$$

where $\mathcal{U}(\cdot)$ denotes the uniform distribution over the considered state-space, and the background probability ψ is defined as a small constant to account for the fact that some of the *importance function* predicted particles do not coincide with the particles that have been propagated through the Langevin process.

Finally, the importance weights are normalized according to (4.15), and the

source location is computed as the Mean of the weighted particles using (4.6).

4.5 The Degeneracy Problem

The *degeneracy* problem can be encountered by most of the basic forms of the PF algorithm. It is observed that, after several PF iterations, the variance of the particles' weights tends to increase significantly. This will lead to the fact that in the end all but one particle will have negligible weights. As a result, the particles are no longer effective at representing the desired PDFp. A suitable measure of *degeneracy* problem is to estimate the effective particle size, as suggested in [23].

$$N_{eff} = 1 / \sum_{n=1}^N (w_k^{(n)})^2 \quad (4.22)$$

If N_{eff} is below certain threshold value, then the particles are considered as ineffective. To tackle such problem, an additional *resampling* stage can be applied. The *resampling* stage samples N new particles from the existing particle set $\{\alpha_k^{(n)}\}_{n=1}^N$ according to the corresponding weights $\{w_k^{(n)}\}_{n=1}^N$. The weights of these “resampled” particles are reset to a uniform value of $1/N$. This is equivalent to break the heavily weighted particles into several uniformly weighted particles and discard any particles with negligible weights. (The actual implementation of the *resampling* algorithm can be performed by using the cumulative weighting scheme that has been discussed in [23].) In this way, the particles are re-distributed in a more effective manner to approximate the PDFp.

4.6 The Imperfect Particle Filtering based ASLT Implementation

In theory, the PF algorithm provides a superior approximation to the Bayesian filtering solution. However, in the practical implementation, the PF algorithm suffers from a few drawbacks that have been addressed in the literature. The major drawback of the basic forms of PF algorithm is the sub-optimal nature of the *importance function* that has been adopted in the corresponding particle *prediction* process.

When first proposed in [27], the *optimal importance function* is designed to minimize the weights variance that were contributed to the *degeneracy* problem and is defined as

$$q_{opt}(\alpha_k|Y_{1:k}) = p(\alpha_k|\alpha_{k-1}, Y_k) \quad (4.23)$$

It is also considered as “optimal” for the Bayesian filtering solution because that it takes into consideration of both previous estimations and current observation. However, the *optimal importance function* suffers from two major limitations. It requires the ability to sample from $p(\alpha_k|\alpha_{k-1}, Y_k)$ and to evaluate $p(Y_k|\alpha_{k-1}) = \int p(Y_k|\alpha_k)p(\alpha_k|\alpha_{k-1})d\alpha_k$. These are generally not possible or have no analytic form in the practical case [9]. Due to such issues, the BS-PF and SIS-PF algorithms have to implement the *optimal importance function* from two different sub-optimal approaches.

Referring back to (4.7) and (4.8), particles are predicted according to the prior density of the *state variable* in the *prediction* stage of the BS-PF. While the

source state transition density function played a crucial role in determining the prior density. If treat the BS-PF algorithm as another “transformed” SIS-PF approach, the sub-optimal *importance function* adopted by the BS-PF algorithm can be expressed as,

$$q_{BS}(\alpha_k|Y_{1:k}) = p(\alpha_k|Y_{1:k-1}) \propto p(\alpha_k|\alpha_{k-1}) \quad (4.24)$$

The BS-PF hence omits any useful information provided by the current observation Y_k . Due to such limitation, in the ASLT implementation, the BS-PF algorithm cannot efficiently spread the particles into state-space regions with potentially high PDFp. Despite the fact that the BS-PF algorithm demonstrates relatively high reverberation robustness, it is generally slow in locating the source or weak in relocating the source after some silence gap.

While for the SIS-PF based ASLT implementation, the *importance function* is directly related to the current observation. According to (4.18), the sub-optimal *importance function* used by the SIS-PF can be interpreted as

$$q_{SIS}(\alpha_k|Y_{1:k}) = p(\alpha_k|Y_k) \propto p(Y_k|\alpha_k) \quad (4.25)$$

which clearly indicates that only current observation is involved in the *prediction* of new particles. In the ASLT implementation, the accuracy of the observation can be easily affected by the reverberation effects, e.g. spurious peaks in the output of the SRP-PHAT. As a result, the “observation-only” *importance function* is no longer a valid indication of the state-space regions with potentially high PDFp. The original idea of efficiently distribute particles into potential state-space is seriously challenged. This contributes to the reduced reverbera-

tion robustness of the SIS-PF based ASLT implementation.

4.7 Summary

In this chapter, we studied the Particle Filtering approach for solving the ASLT problem in reverberant environment. The problem scenario of Bayesian Filtering and its proposed solutions were first briefly discussed. Followed is the in detail description of the Particle Filtering concepts and the two famous PF basic forms, the BS-PF and the SIS-PF. Though in theory, the PF algorithm provides a superior approximation to the Bayesian filtering solution. However, in the practical implementation, the two PF basic forms suffer a few drawbacks. The drawbacks can be contributed to the sub-optimal nature of the adopted *importance function*. The BS-PF omits the current observation in prediction of new particles. On the other hand, the SIS-PF only predict new particles according to the current observation. Therefore, the BS-PF based ASLT implementation is slow in locating the target source location. While the SIS-PF based ASLT suffers from reduced reverberation robustness. As a result, improved PF structures need to be developed in order to overcome the limitations of the existing PF basic forms, and if possible, to combine their respective advantages in the ASLT based implementations.

Chapter 5

Proposed Hybrid Particle Filtering Algorithms

Though the *optimal importance function* cannot be implemented in the practical ASLT applications, the drawbacks of the rather sub-optimal BS-PF and SIS-PF implementations could still be improved. The hybrid PF, which adaptively applies the BS-PF or the SIS-PF, has been generally considered as a straight and workable approach. Several hybrid PF based ASLT implementations can be found in the literature, see [7] and [8]. (Please take note that [8] did not explicitly propose a hybrid PF algorithm. However, it does discussed a very useful combined *likelihood function* based on the fusion of two different measurements. The combined *likelihood function* plays a significant role in the development of the hybrid PF algorithm based ASLT implementations. Therefore, we include [8] into the discussion of the hybrid PF algorithm based ASLT implementations.)

5.1 Existing Hybrid Particle Filtering Algorithms

The key idea of the hybrid PF algorithm is to choose the proper particle sampling method based on the judgement of whether the current observation, i.e the peaks of the localization algorithms' output, is originated from the true source location or from the spurious reverberation effects. Intuitively, the *likelihood function* (especially the *pseudo likelihood function*), which measures the probability of the peaks correspond to potential source location, draws the attention. An excellent and comprehensive investigation on the *likelihood function* has been carried out by E.A. Lehmann and A.M. Johansson, and been summarized in [8]. The very basic and generic approach is to count the number of peaks, N_p , in the observation, i.e.

$$Y_k = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{N_p}\} \quad (5.1)$$

where $\hat{\theta}_v$ is the location estimation corresponding to the local peak.

Without any prior information, there are a total of $N_p + 1$ hypotheses to be considered (assuming single active source case). Each of the peaks could be potentially indicating the true source location, the hypothesis of representing the true location is denoted by $\mathcal{H}_t, t \in \{1, \dots, N_p\}$. Alternatively, there could be one hypothesis, $\mathcal{H}_0$, that none of the peaks are corresponding to the true source location. The *likelihood function* is therefore expressed as,

$$p(Y_k|X_k) = \sum_{t=0}^{N_p} p(\mathcal{H}_t|X_k) \cdot p(Y_k|X_k, \mathcal{H}_t) \quad (5.2)$$

where $p(\mathcal{H}_t|X_k)$ denotes the prior probability of the hypothesis. And it can be generally assumed to be equal and independent, i.e.

$$p(\mathcal{H}_t|X_k) = \frac{1 - p(\mathcal{H}_0|X_k)}{N_p} \quad (5.3)$$

By further assuming statistical independence between the peaks in the location estimation, the conditional term on the right-hand side of (5.2) can be expressed as,

$$p(Y_k|X_k, \mathcal{H}_t) = \prod_{v=1}^{N_p} p(\hat{\theta}_v|X_k, \mathcal{H}_t), t \in \{1, \dots, N_p\} \quad (5.4)$$

The *likelihood function* based on (5.2) to (5.4) provides a comprehensive basis for the development of hybrid or improved PF algorithm. However, due to its rather high computational complexity, the actual implementation of the PF algorithms adopts relatively simple but reasonable alternative hypothesis to represent the accuracy of the observation.

In [7], the developed hybrid PF algorithm, SBF-IS, applies the SIS-PF or the BS-PF sampling method based on three option hypothesis system: reinitialization, importance sampling, and bootstrap. The three hypotheses are closely related to the accuracy of the current observation. If the current observation is accurate in indicating potential source location region, the SIS-PF based particle sampling method should be adopted. It effectively accelerates the source location speed (importance sampling option) or relocates the source after some silent period (reinitialization option). Otherwise, if the current observation is inaccurate due to the reverberation effects, the particles cannot be sampled according to the indication of the observation. Instead, the particles should be propagated through the *transition density function* of the BS-PF to ensure the reverberation robustness (bootstrap option). The corresponding prior probabil-

ities of the three hypothesis are denoted as: P_R , P_S and P_B , and are defined to be

$$P_R + P_S + P_B = 1 \quad (5.5)$$

To ensure a true representation of the observation accuracy, P_R and P_S are updated at each PF iteration according to the number of peaks, N_p , i.e.

$$P_R = \frac{\bar{P}_R}{N_p} \text{ and } P_S = \frac{\bar{P}_S}{N_p} \quad (5.6)$$

where $\bar{P}_R$ and $\bar{P}_S$ are predefined constant values for each hypothesis.

The hypothesis system of the SBF-IS based ASLT implementation makes direct use of the observation information, particularly the number of the local peaks. This approach is a straight and efficient measure of observation accuracy, again, assuming there is only one active source. However, the limitation of this method is that it still requires a computation across the entire state-space region to detect all the peaks. The detection of the peak quantities involves a series of local maxima estimations, which certainly increase the computational load. There is also no standard method to predefine $\bar{P}_R$ and $\bar{P}_S$, which are used to update the corresponding prior probabilities. As a result, there are no consistent prior probability models that can be optimized for any arbitrary acoustic conditions. The actual ASLT performance may vary due to improperly defined prior probabilities.

To tackle the computational load problem, a hypothesis system based on the *pseudo likelihood function*, $p(\hat{\theta})$, was developed in [8]. Because the *pseudo likelihood function* implicitly redefines the observation Y_k as the localization al-

gorithm output, which can be used as a measure of the likelihood. The observation Y_k is no longer treated as a set of peaks in (5.1). Therefore, the detection of local maxima is waived to save the computational load. Furthermore, the likelihood of none peaks correspond to the true source location can be interpreted as a random distribution uniformly distributed across the state-space. The *likelihood function* of (5.2) can now be re-expressed as,

$$p(Y_k|X_k) = p(\mathcal{H}_0|X_k) \cdot \mathcal{U}_D + \gamma(1 - p(\mathcal{H}_0|X_k)) \cdot (p(\hat{\theta}))^r \quad (5.7)$$

where $\mathcal{U}_D(\cdot)$ denotes uniform PDF over the entire state-space $D = \{(x, y) | x \in [x_{min}, x_{max}], y \in [y_{min}, y_{max}]\}$, γ is a normalization constant that ensures the *pseudo likelihood function* is suitable to be used as a density function, and r is used to help shape the localization algorithm output for the source tracking purpose. The detailed derivation of γ and r can be found in [8].

It is very obvious that there are only two hypotheses according to the *likelihood function* of (5.4). The first hypothesis is that the current observation is not accurate, the probability of estimating the true source location is equally distributed over the entire considered domain. The other hypothesis suggests that the current observation is accurate for the indication of potential source location region. The probability of estimating the true source location according to such indication is proportional to the prior probability of the hypothesis and to the strength level of the localization algorithm output, i.e. the *pseudo likelihood function*.

Based on such hypotheses, the next naturally raised question is the judgement of the accuracy of the current observation. Though the number of peaks in the observation can be used as a direct indication of the accuracy, the cor-

responding computational load is still a challenging issue. There are other less computationally expensive approaches to detect inaccurate measurements, one of which is the fusion of two different measurements. The signal to noise ratio (SNR) based voice activity detection (VAD) has been suggested by [8] to help detect the inaccurate measurements in the output of the localization algorithm. The scheme works on the basis of the expected noise power spectral density, which is estimated during nonspeech periods. The estimated noise level is then used during periods with speech activity to estimate the SNR from the observed signal. The hypothesis, which can be denoted as $\mathcal{H}_1$, is that the speaker is active when the SNR is sufficiently higher than a threshold. Alternatively, the speaker is inactive when the SNR is below the threshold, such hypothesis can be denoted as $\mathcal{H}_0$. The two-option hypotheses system of the VAD can be linked to the two-option hypotheses system of the *pseudo likelihood function* (5.4) in the followed manner,

$$p(\mathcal{H}_0|X_k) = p(\mathcal{H}_0|\varphi_k) = 1 - p(\mathcal{H}_1|\varphi_k), \quad p(\mathcal{H}_0|\varphi_k) \propto f(\varphi_k) \quad (5.8)$$

where φ_k is the instantaneous SNR estimation at time k , and $f(\varphi_k) \in [0, 1]$ is statistical decision derived from the state of VAD.

The fusion of the two different measurements hence defined a combined *likelihood function*, i.e. by substituting (5.5) back into (5.4). This representation implicitly suggests that if the current VAD indicates that the source is active, the corresponding observation is most likely to be originated from the true source location. Otherwise, the observation could be quite problematic in indicating the source location if the source is considered as inactive by the VAD. The resulted PF algorithm is named as PF-VAD, which is mainly based on the BS-

PF approach, i.e. it predicts the particles according to the transition density function. However, based on the VAD indication, the PF-VAD takes special caution to update the particles' weights when the source is inactive. Thus, the PF-VAD based ASLT implementation is able to reduce the probability of being misled by the spurious reverberation effects. The limitation of the PF-VAD algorithm is that it requires the prior information of the background noise, which is generally not available for the ASLT applications. Therefore, the algorithm relies on its own VAD output to determine when to start estimate the noise level. Any false indication of source activity status may cause detrimental effects to the followed SNR estimation. Another limitation is that the background noise level must be assumed to be constant during the entire tracking process. If the noise changes rapidly, the VAD is unable to give reliable indication of the observation accuracy.

Despite the corresponding limitations, the SBF-IS and the PF-VAD algorithms have obtained promising results in their respective ASLT applications. The PF-VAD outperforms the SBF-IS with the less computational expensive hypothesis system. However, the SBF-IS algorithm demonstrates a more comprehensive hybrid PF structure that allows fast source localization and robust tracking performance against reverberation. The hybrid PF structure and the combined *likelihood function* using fusion of different measurement serve as a basis for improved PF performance in the ASLT implementation.

5.2 Proposed Hybrid BS-IS Particle Filtering Algorithm

The first proposed hybrid PF algorithm adopts the iterative source height estimation (SHE) based hybrid hypothesis, which can be fused with the localization algorithm output to measure the accuracy of the observation. The SHE based scheme works on the assumption that the height of the source can be considered as a constant within a relatively short time period. Ideally, the array's measurement of the SHE should be a horizontal line within the given time period. Any reverberation caused spurious measurements all tend to differ significantly from the true height level. Therefore, the inaccurate measurement can be easily detected. Furthermore, it is generally true that the reverberation effects have same impact on all the three dimensional estimations of the source location. As a result, if the SHE indicates an inaccurate measurement, the corresponding (x, y) estimations at the localization algorithm output is most probably to be inaccurate as well.

Assuming that any spurious SHE is only caused by the reverberation, the spurious SHE should be distributed randomly within $[0, H]$, where H is the height of the room. The Mean of accurate SHE, $\bar{h}$, can be first estimated as,

$$\bar{h} = (h_k + h_{k-1})/2 \quad (5.9)$$

where h_k and h_{k-1} , $k \in [1, \dots, K]$, are the first two consecutive SHE, whose difference is within a predefined error threshold, δ_h , i.e. $|h_k - h_{k-1}| \leq \delta_h$. In the followed iterations, any SHE that differs significantly from $\bar{h}$, i.e. $|h_k - \bar{h}| > \delta_h$, will be regarded as inaccurate. The corresponding hypothesis is again denoted

as $\mathcal{H}_0$. On the other hand, any SHE satisfies $|h_k - \bar{h}| \leq \delta_h$ will be regarded as accurate. Such hypothesis is denoted as $\mathcal{H}_1$. Meanwhile, $\bar{h}$ is also being updated as the moving average of all the accurate SHE up to the current iteration. The SHE based hypothesis is finally defined similar to (5.8),

$$p(\mathcal{H}_0|X_k) = p(\mathcal{H}_0|h_k) = 1 - p(\mathcal{H}_1|h_k) \quad (5.10)$$

$$\text{where } \begin{cases} p(\mathcal{H}_0|h_k) = 0, & \text{if } |h_k - h_{k+1}| \leq \delta_h \\ p(\mathcal{H}_0|h_k) = 1, & \text{if } |h_k - \bar{h}| > \delta_h \end{cases}$$

Substituting (5.10) back into (5.7) and considering the sub-optimal *importance function* defined in (4.24) and (4.25), intuitively, the SHE based hybrid *importance function* can be derived as,

$$q_{hd}(\alpha_k|Y_{1:k}) = p(\alpha_k|\alpha_{k-1}) \cdot p(\mathcal{H}_0|h_k) + p(\alpha_k|Y_k) \cdot (1 - p(\mathcal{H}_0|h_k)) \quad (5.11)$$

Such hybrid structure allows the adaptive application of either the BS-PF or the SIS-PF according to the SHE based observation accuracy hypothesis. This approach, named as hybrid BS-IS PF is summarized in algorithm 1.

INITIALIZATION:

Assuming that at time $k - 1$, the particle set $\{\alpha_{k-1}^{(n)}\}_{n=1}^N$ and the weights $\{\omega_{k-1}^{(n)}\}_{n=1}^N$ have already been obtained. The iterative particle *prediction* and weight *update* process at time k should be carried in the followed procedure.

ITERATION:

1. **Resampling:** if N_{eff} is less than threshold, resample N new particles with uniform weights of $1/N$ from the existing particle set $\{\alpha_{k-1}^{(n)}\}_{n=1}^N$ according to the associated weights $\{w_{k-1}^{(n)}\}_{n=1}^N$.
2. **Source Height Estimation:** estimate h_k and update $\bar{h}$ using the iterative SHE algorithm as in (5.9).
3. **Hypothesis Option Update:** based on the iterative SHE algorithm estimated accuracy judgement, update the corresponding hypothesis option $p(\mathcal{H}_0|X_k)$ using (5.10).
4. **Particle Prediction:** sample α_k according to BS-PF: $\alpha_k \propto p(\alpha_k|\alpha_{k-1})$, if $p(\mathcal{H}_0|h_k) = 1$; sample α_k according to SIS-PF: $\alpha_k \propto p(\alpha_k|Y_k)$, if $p(\mathcal{H}_0|h_k) = 0$.
5. **Weight Update:** update the new sampled particles' weights according to the designated weight update function of either BS-PF ($p(\mathcal{H}_0|h_k) = 1$) using (4.13) or SIS-PF ($p(\mathcal{H}_0|h_k) = 0$) using (4.19). The weights of all the particles will also be normalized (4.15), so that the summation equals to unity.

ESTIMATION:

The (x, y) location of the source at time k can be estimated with the newly obtained particle set $\{\alpha_k^{(n)}\}_{n=1}^N$ and the corresponding weights:

$$\hat{X}_k = \sum_{n=1}^N w_k^{(n)} \alpha_k^{(n)}.$$

Algorithm 1: Hybrid BS-IS Particle Filtering Algorithm

The hybrid BS-IS PF successively combines the advantages of the SBF-IS and the PF-VAD algorithm. The hybrid BS-IS PF requires less computational

complexity, as compared to the SBF-IS algorithm, in the detection of inaccurate observation. The hybrid BS-IS PF algorithm also does not require any prior information about the acoustic environment. Furthermore, because the same localization algorithm, e.g. SRP-PHAT, which has been used to estimate source location, can be used to estimate the height information in a parallel way. The SRP-PHAT based SHE only needs two additional acoustic sensor arrays to be located along the ceiling and one side wall of the room. So that the SRP-PHAT algorithm can be directly applied to estimate the height h in the (x, z) Cartesian plane. (Please note that, in this report it is still preferred to use h instead of z to denote the height of the source.) The SHE based hypothesis judgement is hence relatively easier and more straight forward as compared to the VAD based hypothesis judgement, which still requires additional computation of signal and noise powers. This constitutes another significant improvement when compared to the PF-VAD algorithm.

5.3 Proposed Statistically Weighted Particle Filtering Algorithm

5.3.1 The Quad-Stage Statistical Source Height Estimation Algorithm

The SHE algorithm that has been proposed in section 5.2 relies on two important assumptions. The first one is that any two consecutive inaccurate SHE must differ from each other significantly, i.e. $\Delta h \gg \delta_h$. The other assumption is that the height of the source is a constant during the entire ASLT process.

Though the first assumption is valid for most, if not all, of the situations, it may still be challenged if δ_h is improperly defined. In such case, the initial estimation of $\bar{h}$ in (5.9) is very problematic and may lead to complete failure of followed estimation accuracy judgement. The second assumption can be, however, easily challenged if the trajectory process is long or complicated enough for the source to change its height, e.g. the source may need to pass by some obstacles or needs to sit down. (However, the source also cannot continuously change its height in a way both significantly, e.g. > 10% in amplitude, and abruptly, e.g. within 0.2 second.) In order to tackle such challenges, additional judgement hypotheses are needed to evaluate the estimation accuracy in a more comprehensive manner. The quad-stage statistical(QSS) SHE algorithm (QSS-SHE) is therefore proposed here. The logic flow chart of the QSS algorithm for a single SHE iteration is illustrated in Figure 5.1. The principle of the QSS-SHE algorithm has been summarized in Algorithm 2.

Figure 5.1: The Logic Flow Chart of Quad-Stage Statistical Source Height Estimation Algorithm

-
1. The initial SHE ($h_k, k = 1$) is always considered as inaccurate, because there is no reference estimation to be compared with.
 2. For any two adjacent height estimations whose differences are within δ_h , the corresponding iteration index and estimation value will be stored (denoted as h_j) as reference for further verification.

$$h_j = h_k, \text{ if } |h_k - h_{k-1}| \leq \delta_h \quad (5.12)$$

3. Before the mean of accurate SHE, $\bar{h}$, can be initialized, the SHE accuracy hypothesis can be defined as:

$$\begin{aligned} \text{Inaccurate, } \mathcal{H}_0 : & \begin{cases} h_k, & \text{if } k = 1 \\ |h_k - h_{k-1}| > \delta_h, & \text{if } h_j \in \emptyset \\ |h_k - h_{k-1}| > \delta_h \ \& \ |h_k - h_j| > \delta_h \end{cases} \\ \text{Semi-inaccurate, } \mathcal{H}_1 : & \begin{cases} |h_k - h_{k-1}| \leq \delta_h, & \text{if } h_j \in \emptyset \\ |h_k - h_{k-1}| \leq \delta_h \ \& \ |h_k - h_j| > \delta_h \end{cases} \end{aligned}$$

$$\text{Semi-accurate, } \mathcal{H}_2 : |h_k - h_{k-1}| \leq \delta_h \ \& \ |h_k - h_j| \leq \delta_h$$

where $\emptyset$ denotes the empty set.

4. The mean of accurate SHE, $\bar{h}$, is only initialized after the first four consecutive* estimations (CE), whose differences are all within the defined threshold δ_h , have been obtained.

$$\bar{h} = \frac{h_k + h_{(k-1)*} + h_{(k-2)*} + h_{(k-3)*}}{4} \quad (5.13)$$

5. The initialized $\bar{h}$ serves as the reference for the followed SHE accuracy judgement, and it keeps being updated as the mean of the latest four (and only four) consecutive* accurate SHE. The accurate hypothesis is defined as:

$$\text{Accurate, } \mathcal{H}_3 : |h_k - \bar{h}| \leq \delta_h$$

6. If the height of the source changes, the update of $\bar{h}$ can be considered as a “re-initialization” with four new CE, whose differences are within the threshold δ_h .

Algorithm 2: Quad-Stage Statistical Source Height Estimation Algorithm

(*Note that the term “consecutive” and the subscript k^* used here are rather coarsely defined. As there can be more than one inaccurate SHE measurement within any two “consecutive” accurate SHE measurements.)

5.3.2 The Statistically Weighted Hybrid Particle Filtering Algorithm

The objective of the hybrid PF algorithm is to be both fast in locating the source in the initial stage and robust against the reverberation effects in the followed tracking. The algorithm hence needs to sample the particles from

potential state-space regions effectively and carefully. This requires a comprehensive evaluation of whether the microphone array's measurement, i.e. the observation, is originated from the true source location.

The binary hypothesis system (5.10) and the associated hybrid *importance function* (5.11) suggest that either applying the BS-PF to sample particles according to the source state transition function when current observation is inaccurate or applying the SIS-PF to sample particles from potential state-space regions when current observation is accurate. Due to the “all or nothing” restriction of the binary hypothesis system, only one PF algorithm can be applied at each iteration. Useful source location information may be neglected by the BS-PF algorithm when a partially corrupted observation has been judged as inaccurate. Alternatively, it is also possible that observation with spurious source location information may be judged as accurate. In such case, the probability of being misled by such erroneous information can be amplified if only the SIS-PF is applied. The harsh decision may therefore lead to degraded ASLT performance. To tackle such problem, the hybrid PF structure needs to be more cautious in applying the corresponding BS-PF/SIS-PF algorithm. It requires a more comprehensive likelihood measure of the observation accuracy. Referring to the hypothesis system of the QSS-SHE algorithm, $\mathcal{H}_0$ certainly corresponds to the case that current observation has lowest likelihood of containing true source location information. While $\mathcal{H}_3$ corresponds to the case that current observation is most likely to be accurate. The hypothesis system hence statistically weights the likelihood in ascending order. As a result, the statistical weight (SW) of each hypothesis can be naturally developed as

$$p(SW_k|\mathcal{H}_t) \propto \sum_{v=0}^t p(\mathcal{H}_v|h_k) \quad (5.14)$$

where $p(\mathcal{H}_v|h_k)$ is the prior probabilities of each hypothesis defined by the QSS-SHE. Ideally, the summation of all the prior probabilities should equal to one. However, an additional probability, p_{er} , is needed to count for the fact that observation with spurious location information may be erroneously judged as accurate. i.e.

$$\sum_{v=0}^3 p(\mathcal{H}_v|h_k) + p_{er} = 1 \quad (5.15)$$

The SW, $p(SW_k|\mathcal{H}_t) \in [0, 1)$, hence effectively estimates the confidence level of sampling particles from the current observation indicated region. The values of SW have been optimized on the basis of numerous MATLAB simulations as $p(SW_k|\mathcal{H}_0) = 0$, $p(SW_k|\mathcal{H}_1) = 0.4$, $p(SW_k|\mathcal{H}_2) = 0.6$, and $p(SW_k|\mathcal{H}_3) = 0.8$. The SW based hybrid *importance function* can be defined as,

$$q_{SW}(\alpha_k|Y_{1:k}) = p(\alpha_k|\alpha_{k-1}) \cdot (1 - p(SW_k|\mathcal{H}_t)) + p(\alpha_k|Y_k) \cdot p(SW_k|\mathcal{H}_t) \quad (5.16)$$

which can be interpreted as: if there are a total of N particles and defining $SW_k = p(SW_k|\mathcal{H}_t) \times 100$, the hybrid PF structure in (5.16) should sample $SW_k\%$ of N particles using the SIS-PF algorithm. While the rest $(100 - SW_k)\%$ of N particles should be sampled using the BS-PF algorithm. In this way, the SW based hybrid PF structure allows both the BS-PF and the SIS-PF to be applied simultaneously. More importantly, the percentage of particles sampled using the corresponding algorithm is adaptively updated according to the confidence level

of the accuracy judgement. When confidence level is low, which corresponds to a lower SW value, a large percentage of the particles will be sampled using the BS-PF algorithm to emphasize the reverberation robustness. When confidence level raises, the SW value also grows, a large portion of the particles will be sampled using the SIS-PF algorithm to effectively explore the potential state-space regions, which in turn boosts the localization speed. Nonetheless, there will always be a small percentage of particles, due to p_{er} , to be sampled using the BS-PF to maintain the reverberation robustness.

However, the contributions of the BS-PF sampled particles can be quite problematic if the weights of the SIS-PF sampled particles are significantly larger than the weights of the BS-PF sampled particles. Defining the weights ratio (WR) as,

$$WR = \frac{\bar{w}_{SIS}}{\bar{w}_{BS}} \quad (5.17)$$

where $\bar{w}_{SIS}$ and $\bar{w}_{BS}$ is the mean weight of the particles sampled using the SIS-PF and the BS-PF respectively. An example of the WR and the corresponding SW has been given in Figure 5.2. (Please take note that, for the frames with zero valued SW, all of the particles are sampled using the BS-PF algorithm. Therefore, the WR at such frames are all zero. Moreover, the WR that have values larger than 5 have been rounded to 5 for the better illustration purpose.)

Several frames with $WR \gg 1$ can be observed and interpreted by two different scenarios. The first scenario is that at the initial stage, the values of SW are generally low, most of the particles are randomly distributed across the entire state-space by the BS-PF. The random distribution has limited overlapping area with the observation indicated regions. As a result, the corresponding

Figure 5.2: An Example of the Weight Ratio (in log10 scale) and the corresponding Statistical Weights

weights of the BS-PF sampled particles tend to differ significantly from those of the SIS-PF sampled particles. In such case, the over valued WR may help locate the source by emphasizing the contributions of the SIS-PF sampled particles. While in the second scenario, significantly large WR can be observed when the value of SW has been relatively stable at a higher level. This usually corresponds to the fact that the observation with spurious location information has been erroneously judged as accurate. The BS-PF sampled particles again fall within the regions with underestimated weights. The BS-PF sampled particles hence have very limited contribution in maintaining the reverberation

robustness. Moreover, the particles have almost nil chance to survive after the resampling process. To tackle the drawbacks caused by the overvalued WR in this scenario, the weights of the BS-PF sampled particles need to be compensated by multiplying with WR. In this way, the BS-PF sampled particles can contribute to the reverberation robustness and stand equal chance to survive in the resampling as compared to the SIS-PF sampled particles.

Finally, the resultant SW based hybrid PF algorithm (SW-PF) is summarized in Algorithm 3.

INITIALIZATION:

Assuming that at time $k - 1$, the particle set $\{\alpha_{k-1}^{(n)}\}_{n=1}^N$ and the weights $\{\omega_{k-1}^{(n)}\}_{n=1}^N$ have already been obtained. The iterative particle *prediction* and weight *update* process at time k should be carried in the followed procedure.

ITERATION:

1. **Resampling:** if N_{eff} is less than threshold, resample N new particles with uniform weights of $1/N$ from the existing particle set $\{\alpha_{k-1}^{(n)}\}_{n=1}^N$ according to the associated weights $\{w_{k-1}^{(n)}\}_{n=1}^N$.
2. **Source Height Estimation:** estimate h_k and the corresponding observation accuracy hypothesis using the QSS-SHE algorithm.
3. **Statistical Weights Update:** based on the QSS-SHE algorithm estimated accuracy hypothesis, update $p(SW_k|\mathcal{H}_t)$ using (5.14).
4. **Particle Prediction:** sample $(100 - SW_k)\%$ of N particles using the BS-PF algorithm, i.e. $\{\alpha_{BS,k}^{(n)}\}_{n=1}^{N \cdot (100 - SW_k)\%} \propto p(\alpha_k|\alpha_{k-1})$; sample another

$SW_k\%$ of N particles using the SIS-PF algorithm, i.e.

$$\{\alpha_{SIS,k}^{(n)}\}_{n=1}^{N \cdot SW_k\%} \propto p(\alpha_k | Y_k).$$

5. **Weight Update:** update the weights $\{w_{BS,k}^{(n)}\}_{n=1}^{N \cdot (100-SW_k)\%}$ and $\{w_{SIS,k}^{(n)}\}_{n=1}^{N \cdot SW_k\%}$ of the two new particle sets using the respective update algorithms in (4.13) and (4.19).
6. **Weight Compensation:** compute the weights ratio (WR) of the two particle sets. If $WR > 1$ and $p(SW_k | \mathcal{H}_t) \simeq 0.8$, the weights of the particles selected from $\{\alpha_{BS,k}^{(n)}\}_{n=1}^{N \cdot (100-SW_k)\%}$ will be compensated by multiplying with WR. The weights of all the particles will also be normalized so that the summation equals to unity.

ESTIMATION:

The (x, y) location of the source at time k can be estimated with the newly obtained particle set $\{\alpha_k^{(n)}\}_{n=1}^N$ and the corresponding weights:

$$\hat{X}_k = \sum_{n=1}^N w_k^{(n)} \alpha_k^{(n)}.$$

Algorithm 3: Statistically Weighted Hybrid Particle Filtering Algorithm

5.4 Numerical Results

The enclosed reverberant environment in this simulation is defined as a typical rectangular room with dimensions of $6m \times 4m \times 3m$. Both the (x, y) and (x, z) Cartesian planes of the room have been divided into grid segmentations of $0.01m \times 0.01m$. The reverberations inside the room are simulated by the Image Method for a range of reverberation times $RT_{60} \in [0, 0.565]$ seconds.

Figure 5.3: An Illustration of the Target Speech Waveform (a), Simulated Reverberant Room (b, c), Microphone Array (o) Distributions and Acoustic Source Trajectory (arrow)

White Gaussian noises are also added at each microphone to approximate an average input SNR of 30dB.

There are a total of $M = 16$ microphones, grouped into 4 linear microphone arrays. The distribution of the microphone arrays is illustrated in Figure 5.3. The two microphone arrays mounted in the (x, y) Cartesian plane have a constant height of $1.5m$ and serve to estimate the (x, y) Cartesian coordinates of the acoustic source. The rest two microphone arrays are allocated along the vertical central axis of a side wall and the horizontal central axis of the ceiling respectively. The latter two arrays are used to estimate the height of the source. The distance between two adjacent microphones in the same microphone array is $1.2m$ (along x -axis), $0.8m$ (along y -axis), or $0.6m$ (along z -axis).

The acoustic source used in the simulation is a male speaker with a fixed height of $1.75m$, and the source is assumed to move straightly from $(1, 1)$ to $(5, 3)$ in (x, y) plane with constant speed. The acoustic signal used here is a pre-recorded clean male speech sample with duration of about 3.2 seconds. The signal's transmission speed is set to be $343m/sec$. The microphone array received signals are sampled at a frequency of $16kHz$ and decomposed into non-overlapping frames of $L = 512$ samples each.

For the source height estimation, the initial* value of $\bar{h}$ is set to be $-0.1m$, and the error threshold δ_h is set to be $0.03m$. (* Please note that this is not equivalent to the initialization of $\bar{h}$. The negative initial value used here is to avoid any false accuracy judgement before $\bar{h}$ has been initialized.)

Figure 5.4 and Figure 5.5 show an example of the SHE results obtained from both the iterative SHE algorithm and the QSS-SHE algorithm, as well as the corresponding binary hypothesis options and statistical weights. Based

Figure 5.4: Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the Iterative SHE Algorithm (b), and the corresponding Binary Hypothesis Option (c) for $RT_{60} = 0.211$ seconds and $SNR = 30dB$

Figure 5.5: Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the QSS-SHE Algorithm (b), and the corresponding Statistical Weights (c) for $RT_{60} = 0.211$ seconds and $SNR = 30dB$

on the plots of Figure 5.4(a) and Figure 5.5(a), it clearly illustrates that the reverberation resulted spurious estimations tend to differ significantly from the true height level. The random distribution of the spurious height estimations also justifies the assumption that any consecutive false estimations tend to differ from each other significantly. While from the plots of Figure 5.4(b) and Figure 5.5(b), it can be easily observed that $\bar{h}$ successfully converges to the true height level after a few accurate SHE measurements have been observed in Figure 5.4(a) and Figure 5.5(a). This means that based on the correct estimation accuracy judgement, the initialization of $\bar{h}$ can be completed in a fast and promising manner. The iterative SHE algorithm is even faster in estimating $\bar{h}$ due to the loosely defined estimation criteria. For both SHE algorithms, the followed updates of $\bar{h}$ only tend to fluctuate along the true height level in a negligible manner. As a result, the robustness of using $\bar{h}$ to track the source in the third dimension has also been justified. In plot of Figure 5.4(c), the binary hypothesis options effectively respond to the variation of SRP-PHAT estimations. Whenever there is erroneous estimation detected, the binary hypothesis option switches to zero to indicate spurious observation. While in plot of Figure 5.5(c), the distribution of the SHE based statistical weights shows a close correspondence to the distribution of the SHE in Figure 5.5(a). For example, whenever the SHE in Figure 5.5(a) differs from the true height level significantly, the corresponding statistical weight is either 0 or 0.4. Hence, the statistical weights truly represent the confidence level of whether the current observation is accurate.

However, because the SHE and the (x, y) estimation is carried out with different microphone arrays. The assumption that the reverberation effects have same impacts on the different array's measurements needs to be justified as well.

For the assumption to be valid, ideally the square errors of both estimations need to be consistent in scale. For example, at time k , if the square error of the SHE is larger than a defined error threshold, say $0.0025m^2$, the square error of the corresponding (x, y) estimation should also be larger than $0.0025m^2$. Otherwise, an accurate (x, y) estimation may be indicated as spurious by the corresponding SHE, which is spurious indeed. Same logic applies that a spurious (x, y) may be indicated as accurate because of the corresponding SHE is accurate indeed. As a result, the square errors of both the SHE and (x, y) estimations should be measured and compared under different reverberation cases.

The plot of Figure 5.6 (a) shows an example of the actual measured square errors of the (x, y) estimation and SHE. Because of the large scale of the square errors, it is hard to observe the consistency of the two estimations. The square errors hence need to be rounded according to the defined error threshold, i.e. any square error that is larger than $0.0025m^2$ will be rounded to be $0.0025m^2$. The rounded errors are plotted in Figure 5.6 (b). From this plot, it can be observed that the two estimations share consistent square errors in a large proportion, though not all, of the entire estimation process. This means that the assumption that the reverberation effects have same impacts on the different array's measurements can be treated as valid for most of the estimation iterations.

A more in-depth analysis of the consistency is conducted by calculating the averaged ratio of frames that the two estimations share consistent estimation errors. Based on 50 simulation runs for each defined reverberation time RT_{60} , the mean, maximum and minimum ratio value of frames sharing consistent estimation errors are summarized in Figure 5.7. From the plot, the mean value of the frame ratio indicates that the two estimations can share consistent estimation

Figure 5.6: An Example of (x, y) Estimation Square Errors vs Source Height Estimation Square Errors for $RT_{60} = 0.565$ seconds and $SNR = 30dB$

Figure 5.7: Ratio of Frames Sharing Consistent Estimation Errors for Different Reverberation Levels

errors for about 80% to 85% of the entire estimation process. It is also observed that the ratio of frames sharing consistent estimation error does not vary much, within 75% minimum to 91% maximum, according to different reverberation levels. From the practical implementation perspective, this result, though not 100%, provides a very comprehensive justification of the assumption. As a result, the proposal of using SHE based accuracy hypothesis to indicate whether current observation is originated from true source is justified. Furthermore, the averaged 80% to 85% consistency ratio also justifies that $SW = 0.8$ is a truthful representative to indicate accurate observation.

Figure 5.8 shows an example of the ASLT results obtained by different PF algorithms, including the two basic forms (BS-PF and SIS-PF) of the PF algorithm, the Hybrid BS-IS PF algorithm, and the SW-PF algorithm. For the applied PF algorithms, the number of particles is $N = 50$, the effective particle size threshold is set to be $N_{th} = 38$. SRP-PHAT is applied to generate the arrays' measurement (*observation function*) Y_k , and the output of SRP-PHAT is used as the *pseudo likelihood*, $p(Y_k|\alpha_k^{(n)})$. For the BS-PF algorithm, the Langevin process in (4.9) is used to define the source state *transition function*. For the SIS-PF algorithm, the threshold of *importance function* is set to be 70% of largest measured SRP-PHAT output. For both the Hybrid BS-IS PF algorithm and the SW-PF algorithm, the SHE error threshold δ_h is set to be $0.03m$.

From the corresponding plots in Figure 5.8, it is clear that the SIS-PF is able to locate the source very fast, but there are quite a lot noticeable errors in the followed tracking process. However, the BS-PF demonstrates much fewer errors in the tracking process but suffers from a slow localization speed in the initial

Figure 5.8: Acoustic Source Localization and Tracking Results Example for $RT_{60} = 0.211$ seconds and $SNR = 30$ dB.

stage. The contradictory performance of these two basic forms of PF algorithm can be explained by their sub-optimal *importance function* discussed in section 4.6. By implementing the hybrid PF structure to best approximate the *optimal importance function*, the hybrid BS-IS PF algorithm and the SW-PF algorithm demonstrate quite promising ASLT performance from the plots. It is obvious that these two hybrid PF algorithms successfully combined the advantage of both the SIS-PF algorithm and the BS-PF algorithm. They are fast in locating the source in the initial stage, and they can firmly track the source with few occasional tracking errors.

It is also clear that the proposed SW-PF algorithm outperforms the proposed hybrid BS-IS PF algorithm in both the initial localization speed and the followed tracking reverberation robustness. The advanced performance can be contributed to the more comprehensively defined hypothesis system of the SW-PF algorithm, as compared to the “all or nothing” binary hypothesis system of the hybrid BS-IS PF algorithm. For the hybrid BS-IS PF algorithm, it needs to establish $\bar{h}$ in the initial stage to make accuracy judgement of the followed SHE. Only after $\bar{h}$ has been established, the hybrid BS-IS PF algorithm will be able to adaptively apply the SIS-PF algorithm to allocate particles into potential state-space regions to boost the localization speed. Otherwise, the hybrid BS-IS PF algorithm will sample all the particles using the BS-PF algorithm, which is slow in the localization speed. For the SW-PF algorithm, some of the SW can be assigned non-zero values even before the initialization of $\bar{h}$. The SW-PF hence allows certain amount of the particles to be sampled using the SIS-PF algorithm, which accelerates the localization speed. During the tracking process, the reverberation caused spurious observations, i.e. inaccurate (x, y) measurement, tend to confuse the PF algorithm. This is particularly detrimen-

tal when the reverberation effects on the SHE are not consistent with those on the array's (x, y) measurement. For example, an inaccurate (x, y) measurement may be indicated as accurate because of the reverberation effects on the corresponding SHE is not serious. In such case, the hybrid BS-IS PF algorithm will still sample all the particles using the SIS-PF algorithm. As a result, particles are allocated into wrong state-space regions and severe tracking errors occur. Instead, the SW-PF algorithm always involves some BS-PF sampled particles at all PF iterations. With proper weight compensation, those BS-PF sampled particles effectively suppresses the misleading effects caused by the reverberation. The reverberation robustness of the SW-PF algorithm is hence guaranteed to certain extent.

To get a more comprehensive comparison of the four PF algorithms, the tracking root mean square errors (RMSE) and the track loss percentage (TLP)

$$TLP = \frac{\text{number of frames with estimation error} \geq 0.1m}{\text{total number of frames}} \times 100\% \quad (5.18)$$

are calculated and averaged over 50 simulation runs each for a range of reverberation levels, $RT_{60} \in [0, 0.565]$ seconds. The results are summarized in Figure 5.9 and Figure 5.10 respectively. From the plots, it is clear that the SW-PF algorithm demonstrates the best overall performance with the lowest RMSE and TLP for almost all the reverberation levels. The hybrid BS-IS PF algorithm shares almost same TLP with the SW-PF algorithm, and demonstrates even lower RMSE for certain relatively low reverberation levels. While for the two basic forms of PF algorithm, the SIS-PF algorithm is better than the BS-PF algorithm in both RMSE and TLP results. The worst performance of the BS-

Figure 5.9: Average Tracking Root Mean Square Error vs Reverberation Time (SNR= 30dB)

Figure 5.10: Average Track Loss Percentage vs Reverberation Time (SNR=30dB)

PF algorithm can be contributed to the fact that it has difficulties in locating the source due to its sub-optimal *importance function*. The more iterations the BS-PF algorithm takes to locate the source, the higher the RMSE and TLP will be resulted. This makes the performance of the BS-PF algorithm very ad-hoc for each single simulation run. As a result, the overall performance of the BS-PF algorithm may degrades to a large extent.

Figure 5.11: An Illustration of the Target Speech Waveform with Silence Gap (a), Simulated Reverberant Room (b, c), Microphone Array (o) Distributions and Acoustic Source Trajectory (arrow)

Last but not least, the simulation setup was modified and presented in Figure 5.11. The target speech was looped twice and padded with some zeros

in-between to simulate a noticeable silence gap. Before the silence gap, the source is defined to move from $(1, 1)$ to $(3, 2)$ with a constant height of 1.75m . After the silence gap, the source becomes active again and restarts from $(3.5, 3)$ to $(5.5, 1.5)$. However, during this process, the height of the source changes from 1.75m to 1.2m . The dimension and the location of the microphone arrays are kept unchanged. The Image Method was applied to simulate a moderate reverberation level, $RT_{60} = 0.211\text{seconds}$. The array's input SNR was maintained to be 30dB .

Based on such simulation settings, the QSS-SHE algorithm was applied to locate and track the source in the third dimension and the SW-PF was applied to track the source in the (x, y) Cartesian plane. The results have been plotted in Figure 5.12 and Figure 5.13 respectively. From Figure 5.12(b), it is easy to observe that shortly after the source became active again from the temporary silence gap, the QSS-SHE algorithm successfully relocated the source (i.e. the re-initialization of $\bar{h}$). The QSS-SHE algorithm then accurately followed the source trajectory in the third dimension. The SHE based SW also showed close correspondence with the SRP-PHAT estimations in Figure 5.12(a). The variation of SW between the 120th and 140th frame gives a clear indication of how the SW based hypothesis system judges the SHE during the silence gap: height estimations were selectively discarded (those due to the silence gap and reverberation) or gradually emphasized (those contributing to the re-initialization of $\bar{h}$). As a result, the SW-PF algorithm demonstrates promising re-localization capability and robust tracking performance in Figure 5.13. With the SW based hypothesis system, the SW-PF algorithm is able to decide when to apply the SIS-PF algorithm to boost the localization speed in the beginning of the entire process. Once the source became silent, the SW-PF algorithm sticks

to the BS-PF algorithm to maintain track of possible source trajectory. After the QSS-SHE algorithm has re-initialized $\bar{h}$, the SW based hypothesis system will guide the SW-PF algorithm to allocate particles into potential state-space regions to relocate the source. For both the tracking process before and after the silence gap, the SW-PF algorithm demonstrates high reverberation robustness by maintaining the occasional tracking errors in manageable scales. Again, this can be contributed to the more comprehensively defined hybrid PF structure that involves some BS-PF sampled particles at all PF iterations. The BS-PF sampled particles greatly suppressed the misleading effects of the reverberation.

5.5 Summary

In this chapter, we studied the hybrid PF algorithm for better approximation to the “*optimal importance function*”. The existing hybrid PF algorithms and the associated combined *likelihood function* have been first reviewed. Then we proposed our own hybrid hypothesis option systems based on the accuracy detection of the source height estimation. We proposed two algorithms, namely the iterative SHE algorithm and the QSS-SHE algorithm to estimate and update the mean of accurate SHE. Based on the SHE results, we proposed the hybrid BS-IS PF algorithm that adopts a binary hypothesis option structure and the SW-PF algorithms that adopts the more comprehensive statistical weights system. Both hybrid PF algorithm based ASLT implementations are able to adaptively apply either BS-PF algorithm or SIS-PF algorithm whenever necessary. The numerical results have demonstrated that the proposed hybrid PF algorithms have successfully combined the advantages of fast location speed and reverberation robustness.

Figure 5.12: Source Height Estimation Results Obtained from the SRP-PHAT Algorithm (a) and the QSS-SHE Algorithm (b), and the corresponding Statistical Weights (c) for Source Traveling with Various Height ($RT_{60} = 0.211$ seconds and $SNR = 30dB$)

Figure 5.13: Acoustic Source Localization and Tracking Results Example for $RT_{60} = 0.211$ seconds and SNR= 30dB.

Chapter 6

Conclusion and Future Work

This chapter summarizes the results of this thesis and describes open topics for future research.

6.1 Conclusion

This thesis has addressed issues regarding reverberant environments modeling and reverberation robust ASLT techniques. Various problems are discussed and new algorithms are proposed to overcome the existing problems. Simulation examples and comparison with existing algorithms are given to illustrate the effectiveness of the proposed algorithms. The following is a summary of the contributions.

Reverberant Environments Modeling: In this thesis, the Image Method for simulating room impulse responses has been studied. From the literature study, we realized that though the Image Method provides a rather direct and easy approach to simulate the reverberant environment impulse response. There

is implementation issue due to the large quantity of images needed to simulate highly reverberant situation. To overcome such issue, we proposed the image cube method to fast expand the image amount into large scales and to quickly calculate the distances between all the images and the microphone. The image cube method hence solved the computational complexity issue. Additionally, we proposed the image amount threshold estimation method to find the most efficient image amount needed to simulate the reverberation effects for different reverberation levels. The proposed method examines the distribution of the decayed sound wave intensity levels. From the numerical results, it has demonstrated that white Gaussian noise can be used to mask the reverberation effects when the image amount is beyond certain threshold for various reverberation levels. Therefore, the computational load has been greatly reduced. With such improvements, all the subsequent research studies in this thesis will effectively adopt the Image Method based reverberant environment simulation, which provides a comprehensive basis to examine the reverberation robustness of the existing and proposed ASLT techniques.

Reverberation Robust ASLT Techniques: In chapter 3, we reviewed the existing audio source localization (ASL) techniques, which can be classified into three categories: the steered beamformer based, the high resolution spectral estimation based and the time difference of arrival based. The most representative algorithms under each category has been discussed. The corresponding strength and weakness of each category are also summarized accordingly. Additionally, we analyzed in detail the working principles and simulation performances of SRP-PHAT algorithm, which has been well recognized as a promising reverberation robust ASL technique. Based on our analysis, we realized that even for the extreme reverberation settings, the SRP-PHAT could still indicate

potential source location with certain depressed peaks around the true source located region. This property could be further utilized in the followed audio source localization and tracking (ASLT) algorithm study.

In chapter 4, we studied the Particle Filtering (PF) approach for solving the ASLT problem in reverberant environment. The problem scenario of Bayesian Filtering and its proposed solutions were first briefly discussed. Followed is the in detail description of the Particle Filtering concepts and the two famous PF basic forms, the BS-PF and the SIS-PF. Though in theory, the PF algorithm provides a superior approximation to the Bayesian filtering solution. However, in the practical implementation, the two PF basic forms suffer a few drawbacks. The drawbacks can be contributed to the sub-optimal nature of the adopted *importance function*. The BS-PF omits the current observation in prediction of new particles. On the other hand, the SIS-PF only predict new particles according to the current observation. Therefore, the BS-PF based ASLT implementation is slow in locating the target source location. While the SIS-PF based ASLT suffers from reduced reverberation robustness. As a result, improved PF structures need to be developed in order to overcome the limitations of the existing PF basic forms, and if possible, to combine their respective advantages in the ASLT based implementations.

Finally, in chapter 5, we discussed the hybrid PF algorithm based ASLT solutions. The main idea is to fuse, or at least partially, the BS and the SIS algorithms to combine the reverberation robustness and fast location/relocation abilities. The fusion of different PF algorithms is governed by the hypothesis system based on the accuracy judgement of current state observation. Depending on whether the current state observation originates from true source location or

from spurious reverberation effects, the hypothesis system adopts corresponding option to apply suitable PF algorithm. The choices of the current state observation judgement and the hybrid hypothesis system are widely open. In this thesis, we proposed two hybrid PF structures that use the hypothesis system based on the source height estimation (SHE). One of the important assumptions is that the source height is always a constant in the entire ASLT process. The other assumption is that any inaccurate height estimation is only caused by the reverberation effects. Therefore, the variation of the current SHE from the (estimated) true source height level serves as a direct indicator in judging the accuracy of the array's measurement. Based on the judgment and related hypothesis option, the hybrid PF algorithm is able to adaptively apply either the BS-PF or the SIS-PF to tackle the reverberation effects. The first hybrid PF structure (hybrid BS-IS PF) is governed by the binary hypothesis system based on the iterative SHE algorithm. The iterative SHE algorithm estimates the true source height by iteratively updating the mean of height estimations whose differences are within pre-defined threshold. The second hybrid PF structure (SW-PF) that we proposed is governed by the statistical weights (SW) hypothesis system, which is based on the quad-stage statistical (QSS) based SHE (QSS-SHE) algorithm. The QSS-SHE algorithm is able to estimate and update the height information in a more responsive and reverberation-robust manner. As a result, the estimated source height information not only serves as the measure of the observation accuracy but also extends the related ASLT estimations from two dimensional (i.e. (x, y) in Cartesian plane) to three dimensional (i.e. (x, y, z) in Cartesian plane). The QSS-SHE algorithm is also able to discard any false observation even if the height of the source changes during the tracking process, making the algorithm more applicable in the practice. Another added novelty

is that the hypothesis system of the hybrid PF structure has been upgraded from the binary “all-or-nothing” option system, as in our first proposed one, to a more comprehensive option system based on the SW. The SW are estimated and upgraded based on the added observation accuracy judgement decisions of the QSS-SHE algorithm. The SW based hypothesis system allows the BS-PF and the SIS-PF to be applied simultaneously and fuses the particles processed by both PF algorithms to maintain high reverberation robustness. Various numerical examples have been conducted and the results all demonstrated that the two proposed hybrid PF algorithms successfully combined the advantages of the BS-PF (reverberation robustness) and the SIS-PF (fast localization speed). We have also showed that the SW-PF successfully relocates the source location in three dimensions after significant silence gap.

6.2 Future Work

Besides the major contributions that have been summarized above, we also present a few research topics for the future work.

Importance Region Vector Quantization based Hybrid Particle Filtering: To determine whether the current microphone array’s measurement is accurate, one direct approach is to look at the localization algorithm estimated power spectrum. Based on our previously estimated SRP-PHAT simulation results, we find that if there is only one sharp peak or several peaks but located in relative squeezed manner, then most likely that at this time step the source is located at the peak(s) indicated region, performing SIS-PF is not a problem in this case. Otherwise, if there are several peaks spread around the entire state

space, then most probably the estimated observation information at this step is not so reliable to be used to indicate the source location. The BS-PF is more suitable to be applied at this situation due to the fact that SIS-PF has relative high possibility to sample particles from incorrect importance regions.

Figure 6.1: SRP-PHAT Power Spectrum and the Corresponding Importance Region Distribution

However, the computation of the number of peaks and their respective locations can be quite problematic and result high computational load. Hence, we need to find some alternative solution to release the computation complexity. Recall that the state space is just the two-dimensional steering surface defined by the microphone arrays. The steering surface can be divided into fine grid points and each grid point has a unique one-to-one correspondence to the array estimated SRP-PHAT power spectrum. This relationship is depicted in Figure 6.1, for this case we no longer need to know the exact information of the peak quantities and distribution information. Instead we could just set certain threshold according to the peak value of the overall power spectrum, and gather the location information of all the grid points that have corresponding SRP-PHAT power levels beyond the threshold. These grid points formed up the importance region from which the SIS-PF could sample the particles. And

later by performing the vector quantization of all the gathered grid points, we are able to estimate the distribution of the importance region. If there is only one converged region, then applying SIS-PF to sample particles from this region is safe. If the quantized regions turn out to be diverged, then the sampling of particles should better be performed in the way defined by the source state transition function, i.e. the BS-PF should be applied.

Source Motion Validation based Hybrid Particle Filtering: Since that the source that we are going to track can only move inside the relative small enclosure, and that the source's movement is generally "mild" (i.e. the source is very unlikely to accelerate, decelerate or change moving direction in dramatic manner) most of the time, so with such assumption we can make judgment of the observation reliability based on the estimated source motion status.

Because in audio source tracking, the source motion status is unknown to us, so we need to learn from the previously estimated source locations to help decide whether the current state source location validates with the source motion or not. One simple but effective approach to help determine the source motion validation is based on the famous Newton's Laws of Motion, and an illustration of the source motion validation that is based on previously estimated source locations is shown in Figure 6.2.

Upon estimated a series of source locations, the source's past motion status including the velocity and acceleration can be determined by the Newton's Laws of Motion. As a result, we can then define a circular source motion validation zone, by again applying the Newton's Laws of Motion, to predict the current step source potential location region. The source motion validation zone should

Figure 6.2: An Illustration of Source Motion Validation based on Previously Estimated Source Locations

have its origin located at the previous one step estimated source location. The radius R of the circular zone measures the possible distance between the current state source location and the previous one step estimation. If the current step observation is accurate, the corresponding source location estimation should fall inside the circular zone, and the SIS-PF can be applied to sample particles according to the observation. However, if the current observation indicates that the source is located outside the circular zone, then most probably the current observation is unreliable, the BS-PF should be applied to sample particles through past estimations and the state transition density instead.

Velocity Factors in Improving BS-PF Performance: We may also investigate the performance of BS-PF with varied excitation parameters of the source state transition function. Because BS-PF samples new particles according to the Langevin process defined in (4.9), the randomness of the particles being sampled are largely determined by the steady-state velocity $\bar{v}$ (4.12) based random process (i.e. the noise). Moreover, the steady-state velocity is a predefined parameter that may or may not correctly represent the actual source motion status. Therefore, it is possible for us to estimate the most suitable

steady-state velocity value for certain source motion status. The direct approach is to examine the BS-PF estimation square errors with various $\bar{v}$ values. One example of the examination is plotted in Figure 6.3.

Figure 6.3: An Illustration of BS-PF Estimation Square Errors with Various $\bar{v}$ values

From the plot, it is obvious that the initial BS-PF estimation errors are quite large due to the fact that the actual source velocity is faster than $1m/sec$. Gradually, with $\bar{v}$ converging to the actual source velocity, the estimation errors of BS-PF keep reducing. However, once $\bar{v}$ goes beyond the actual source velocity, the errors of the BS-PF estimation starts to increase again.

Our proposed interpretation on such observations is that if $\bar{v}$ is slower than the source actual velocity, the BS-PF sampled particles do not have enough

excitation power to follow the source trajectory. On the other hand, if $\bar{v}$ is set to be too large, the particles will be given excessive excitation power so that they will be re-allocated into random spaces within the room. Again, the over excitement will result in larger estimation errors. Therefore, we need some careful investigation works to set the most suitable $\bar{v}$ value for improved BS-PF performance.

Author's Publications

Journal Paper

1. T. Li and W. Ser, "Three dimensional acoustic source localization and tracking using statistically weighted hybrid particle filtering algorithm", *EURASIP Journal on Signal Processing*, 90: pp.1700-1719, 2010.

Conference Paper

1. T. Li and W. Ser, "Hybrid BS-IS particle filter based acoustic source tracking algorithm", 16th *European Signal Processing Conference, EUSIPCO 2008*, Lausanne, Switzerland, August, 2008.

Bibliography

- [1] D. Mintzer, "Transient Sounds in Rooms," *Journal of the Acoustical Society of America*, 22: pp. 341-352, 1950.
- [2] R.H. Bolt, P. EDoak, and P.J. Westervelt, "Pulse Statistics Analysis of Room Acoustics," *Journal of the Acoustical Society of America*, 22: pp. 328-339, 1950.
- [3] J.M. Berman, "Behavior of Sound in a Bounded Space," *Journal of the Acoustical Society of America*, 57: pp. 1275-1291, 1975.
- [4] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 107, pp. 384-391, 1979.
- [5] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. ICASSP 2001*, vol.5, pp. 3021-3024.
- [6] D.B. Ward, E.A. Lehmann and R.C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 826-836, 2003.

-
- [7] E.A. Lehmann and R.C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP Journal on Applied Signal Processing, special issue on Advances in Multi-Microphone Speech Processing*, vol. 2006, pp. 168-176, 2006.
 - [8] E.A. Lehmann and A.M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 28-38, 2007.
 - [9] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *Proc. Radar and Signal Processing*, vol. 140, pp.107-13...3, 1993
 - [10] E.A. Lehmann, "Particle Filtering Methods for Acoustic Source Localisation and Tracking," *PhD Thesis, Australian National University*, Australia, July 2004
 - [11] M.S. Bartlett, "Something Periodograms from Time Series with Continuous Spectra," *Nature*, vol. 161, pp. 686-687, 1948.
 - [12] J.C. Chen, R.E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Transactions on Signal Processing*, vol. 50(8). pp.1843-1854, August 2002.
 - [13] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *IEEE Proceedings*, vol. 85, pp. 1031 C 1060, Jan. 1969.

-
- [14] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, pp. 276-280, Mar. 1986
 - [15] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107(1), pp. 384-391, January 2000.
 - [16] G.C. Carter, A.H. Nuttall and P.G. Cable, "The smoothed coherence transform (SCOT)," *Naval Underwater Systems Center, New London Lab.*, New London, CT, Tech, Memo TC-159-72, August, 1972.
 - [17] E.E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, pp. 1321-1324, USA, 1996.
 - [18] J.O. Smith and J.S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 1661-1669, December 1987.
 - [19] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol.4, pp. 148-152, March 1996.
 - [20] J. Dibiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments," *PhD Thesis, Brown University, USA*, May 2000

- [21] J.H. DiBiase, H.F. Silverman and M.S. Brandstein, "Robust Localization in Reverberant Rooms," *Microphone Arrays Signal Processing Techniques and Applications*, pp.157-180, January 2001.
- [22] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, N.J, 1979.
- [23] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174-188, 2002.
- [24] A. Doucet, N. Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*, Springer-Verlag, New York, 2000.
- [25] G. Kitagawa, "Monte Carlo filtering and smoother for non-Gaussian non linear state space models," *Journal of Computational and Graphical Statistics*, vol. 5(4), pp. 1-25, December 1996.
- [26] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197-208, 2000.
- [27] V.S. Zaritskii, V.B. Svetnik, and L.I. Shimelevich, "Monte Carlo technique in problems of optimal data processing," *Automation and Remote Control*, vol. 12, pp. 95-103, 1975.
- [28] W.C. Sabine, *Collected Papers on Acoustics*, Cambridge: Harvard University Press, 1922.
- [29] M. Brandstein and D. Ward, editors. *Microphone arrays: signal processing techniques and applications*. Springer-Verlag, Berlin, 2001.

-
- [30] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation." *IEEE Transactions on Speech and Audio Processing*, vol. 4, no.2, pp.148-152, March 1996.