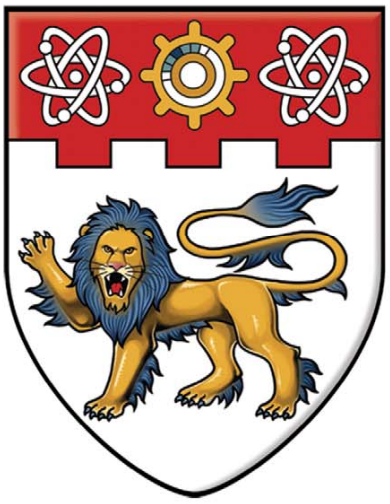
Localization and Tracking of Acoustic Sources in Room Environment



Wu Kai

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University in fulfilment of the requirement for the degree of Doctor of Philosophy

2017

To my parents Wu Lianhui, Ling Xiulan.

To my supervisor Prof. Andy W. H. Khong

Acknowledgment

First and foremost, I would like to sincerely express my gratitude to my advisor Prof. Andy W. H. Khong for his continuous support during my PhD study. His guidance and encouragement, together with his willingness to unremittingly improve the research paper quality, have tremendously benefited myself. His enthusiasm and mentorship have made my postgraduate education a very rewarding experience.

I would like to thank Dr. Vaninirappuputhenpurayil Gopalan Reju and Dr. Shu Ting Goh for many productive discussions and collaborations. Their valuable suggestions and comments on parts of this work has been deeply appreciated. Many thanks are also addressed to my lab-mates for technical discussions in meetings and casual talks during lunches and dinners. Thanks to my roommates who have been together and made the school life a little more colorful.

Specially, I am grateful to Ms. Chiyun, for being with me and inspiring me to go through the last miles of this journey.

Finally, I would like to express my gratitude to my parents, who encouraged and supported me to go through all the frustrations I encountered. I am grateful for their full understanding during my PhD study.

Wu Kai

Abstract

This thesis addresses two areas of the acoustic source localization and tracking (ASLT) problem in a room environment, namely, tracking of acoustic sources using multiple omni-directional microphone arrays and DOA estimation of the acoustic sources using a single acoustic vector sensor (AVS). The challenges for the ASLT problem, in both aforementioned applications may include room reverberation, background environmental noise, sound interference, as well as the presence of multiple speakers.

For multiple omni-directional arrays, the thesis focuses on the source tracking problem where the source position is estimated sequentially across several time frames using the particle filter (PF) framework. To achieve single-source tracking in a reverberant and noisy environment, an algorithm which utilizes the well-known sequential importance resampling PF (SIRPF) framework is proposed. As will be shown in this thesis, this proposed algorithm derives the measurement likelihood which is robust to reverberation and noise. For single-source tracking in the presence of sound interference, another SIRPF-based algorithm is proposed. This algorithm exploits the harmonicity feature of a speech signal for deriving the measurement likelihood. Due to the use of distinctive speech feature, speech-sensitive tracking can be achieved in the presence of sound interference. The performance of these two algorithms have been verified through simulation.

The problem of tracking of alternating speakers will then be discussed in which the speech sources are active in turns. For solving this problem, a novel swar-

m intelligence based PF (SWIPF) which jointly exploits the advantages of PF and particle swarm intelligence is proposed. The PF framework is used as sequential state estimation framework which suits for the tracking problem. The limitation of PF, which lies in the particle sampling, is addressed by incorporating the particle swarm intelligence. By using the swarm intelligence, particles are associated with interaction and memory mechanisms. When alternation occurs, particles can be directed toward the true source location by interacting and sharing the fitness information among themselves. In addition, the memory mechanism allows particles to retain their previous best-fit positions when signals are corrupted by noise and reverberation. The proposed SWIPF is verified using both simulations and real experiments.

The thesis finally considers the multi-source DOA estimation problem using an AVS. Unlike the conventional microphone arrays which requires inter-spacing between microphones, the co-location of sensor elements in an AVS can be exploited to achieve robust DOA estimation in a reverberant environment. As opposed to the existing multi-source DOA estimation algorithms using AVS, the proposed algorithm is developed from a reverberant received signal model. By exploiting the co-location of the sensor elements in an AVS, the low-reverberant-single-source (LRSS) zones of the received signals, where only one source is dominant with a high signal-to-reverberation ratio, can be identified. By using only these identified LRSS zones followed by a clustering step, multi-source DOA estimation in reverberant environment can therefore be achieved. Simulation is conducted to verify the performance of the proposed algorithm.

Contents

Abbreviations

Part I

ASLT:	Acoustic source localization and tracking
AST:	Acoustic source tracking
DOA:	Direction-of-arrival
DTFT:	Discrete Fourier transform
EKPF:	Extended Kalman filter
EKPF-TDOA:	TDOA-based EKPF
GCC:	Generalized cross-correlation
HSRP:	Harmonicity based steered response power
IS:	Importance sampling
MBE:	Multi-band excitation
pdf:	Probability density function
PF:	Particle filter
PHAT:	Phase transform
PSO:	Particle swarm optimization
RMSE:	Roor-mean-square error
RSRP:	Regional steered response power
SBF:	Steered beamformer
SIR:	Signal-to-interference ratio
SIRPF:	Sequential importance resampling particle filter

SIRPF-SRP:	SRP-based SIRPF
SIRPF-TDOA:	TDOA-based SIRPF
SIRPF-RSRP:	RSRP-based SIRPF
SIRPF-HSRP:	HSRP-based SIRPF
SNR:	Signal-to-noise ratio
SRP:	Steered response power
SRR:	Signal-to-reverberation ratio
SWIPF:	Swarm intelligence based particle filter
SWIPF-TDOA:	TDOA based SWIPF
TDOA:	Time-difference-of-arrival
TF:	Time-frequency

Part II

AVS:	Acoustic vector sensor
DOA:	Direction-of-arrival
FCM:	Fuzzy c-means
LRSS:	Low-reverberant-single-source
MUSIC:	multiple signal classification
RMSAE:	Root-mean-square angular error
SSP:	Single-source point
STFT:	Short-time Fourier transform
TF:	Time-frequency

List of Notations

Part I

$i,\ j = 1 \dots N$	Microphone index
N	Number of microphones
Υ	Microphone pair collection, $\Upsilon = \{(1,2),\ \dots,\ (N-1,N)\}$
c	Speed of sound
t	Sample index
k	Frame index
ω	Angular frequency bin index
Ω	Frequency range of interest
$s(t)$	Source signal
$\underline{s}(\omega,k)$	Source signal in the time-frequency (TF) domain
$y(t)$	Microphone received signal
$\underline{y}(\omega,k)$	Microphone received signal in the TF domain
$n(t)$	Noise signal
$\underline{n}(\omega,k)$	Noise signal in the TF domain
$\mathbf{r}^{\text{src}}$	Source position
$\widehat{\mathbf{r}}^{\text{src}}$	Estimated source position

$\mathbf{r}^{\text{mic}}$	Microphone position
$\mathbf{r}$	Steered (look) position
$h(\mathbf{r}^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$	Channel impulse response
Δt_i	Propagation time-delay for the ith microphone
$\tau_k^{(i,j)}$	TDOA for microphone pair (i, j) at frame k
$\widehat{\tau}_k^{(i,j)}$	Estimated TDOA for microphone pair (i, j) at frame k
$\mathcal{T}(\mathbf{r})$	TDOA function defined as $\mathcal{T}(\mathbf{r}) = (\|\|\mathbf{r}_k - \mathbf{r}_j^{\text{mic}}\|\| - \|\|\mathbf{r}_k - \mathbf{r}_i^{\text{mic}}\|\|)/c$
$\Psi_k^{(i,j)}(\tau)$	GCC function for microphone pair (i, j) at frame k
$\mathcal{P}_k(\mathbf{r})$	Steered response power for the steered position $\mathbf{r}$ at frame k
$w(\omega, k)$	Weight function for GCC/SRP function
$\mathbf{x}_k$	State variable at frame k
$\widehat{\mathbf{x}}_k$	State estimate at frame k
$\mathbf{z}_k$	Measurement variable at frame k
$\mathcal{G}(\cdot)$	State transition process
$\mathcal{H}(\cdot)$	Measurement function
$\mathbf{u}_k$	Process noise
$\mathbf{w}_k$	Measurement noise
$p = 1 \dots N_p$	Particle index
N_p	Number of Particles
$\mathbf{x}_k^{(p)}$	The pth particle at frame k
$w_k^{(p)}$	The pth particle weight at frame k
$p(\mathbf{x}_k \mid \mathbf{z}_k)$	Posterior probability density function
$p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$	Prior propagation density function
$p(\mathbf{z}_k \mid \mathbf{x}_k)$	Measurement likelihood function
$p^{(\text{IS})}(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{z}_k)$	Importance sampling density function

$\mathcal{N}(\mathbf{x};\mu,\sigma^2)$	Gaussian distribution function with mean μ and variance σ^2
$\mathcal{U}_{\mathcal{D}}(\mathbf{x})$	Uniform distribution function
$\mathcal{D}$	Enclosure domain of interest

Part II

$l = \dots L$	source index
k	Frame index
ω	Angular frequency
$s(t)$	Source signal
$\underline{s}(\omega,k)$	Source signal in the TF domain
$y_{\mathrm{p}}(t)$	Monopole sensor-element received signal
$\underline{y}_{\mathrm{p}}(\omega,k)$	Monopole sensor-element received signal in the TF domain
$\mathbf{y}_{\mathrm{v}}(t)$	3×1 vector consists of dipole sensor-element received signals
$\underline{\mathbf{y}}_{\mathrm{v}}(\omega,k)$	3×1 vector consists of dipole sensor-element received signals in the TF domain
$n_{\mathrm{p}}(t)$	Noise signal at monopole sensor-element
$\underline{n}_{\mathrm{p}}(\omega,k)$	Noise signal at monopole sensor-element in the TF domain
$\mathbf{n}_{\mathrm{v}}(t)$	3×1 vector consists of noise signals at dipole sensor-element
$\underline{\mathbf{n}}_{\mathrm{v}}(\omega,k)$	3×1 vector consists of noise signals at dipole sensor-element in the TF domain
Δt	prorogation time from the source to the AVS
$h_{\mathrm{p}}(t)$	impulse response from the source to the monopole sensor-element
$\mathbf{h}_{\mathrm{v}}(t)$	impulse responses from the source to the dipole sensor-elements
$[\phi^{\mathrm{src}},\ \psi^{\mathrm{src}}]^{\mathsf{T}}$	The azimuth and the elevation angles of the source
$\mathbf{u}^{\mathrm{src}}$	unit vector pointing towards the source
$\mathbf{u}$	unit steering vector

$\mathbf{q}^{\text{src}}$	$\mathbf{q}^{\text{src}} = [1,\ \mathbf{u}^{\text{src}\mathsf{T}}]^{\mathsf{T}}$
$\mathbf{q}$	$\mathbf{q}^{\text{src}} = [1,\ \mathbf{u}^{\mathsf{T}}]^{\mathsf{T}}$
$\mathbf{R}$	Covariance matrix computed in the time domain
$\underline{\mathbf{R}}$	Covariance matrix computed in the TF domain
k'	TF-zone time index
ω'	TF-zone frequency index
$\mathcal{Z}(\omega', k')$	TF zone for index (ω', k')
$\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$	Covariance matrix computed for the TF zone $\mathcal{Z}(\omega', k')$
$\mathbf{\Theta}_{\mathcal{Z}(\omega',k')}$	Hermitian angle features for the TF zone $\mathcal{Z}(\omega', k')$

Chapter 1

Introduction

Acoustic source localization and tracking (ASLT) refers to the problem of estimating the position or direction from which a source signal originates with respect to the microphone array geometry. It is a fundamental problem in many speech and audio applications. In recent decades, research has focused on localization and tracking of acoustic sources in an adverse room environment which may contain considerable amount of room reverberation, competing sound interference signal and background noise. In this thesis, a number of proposed algorithms for ASLT will be investigated. This introductory chapter first presents the motivation of this thesis in Section 1.1. In Section 1.2, the organization of this thesis is discussed. Finally, the author's contributions are summarized in Section 1.3.

1.1 Motivation

Recently, ASLT has received increased attention due to its widespread applications in hands-free speech communication and teleconferencing systems [1]. As depicted in Fig. 1.1, unlike a conventional speech recording system which requires the speaker to hold a microphone, the hands-free sound recording system allows the speaker to

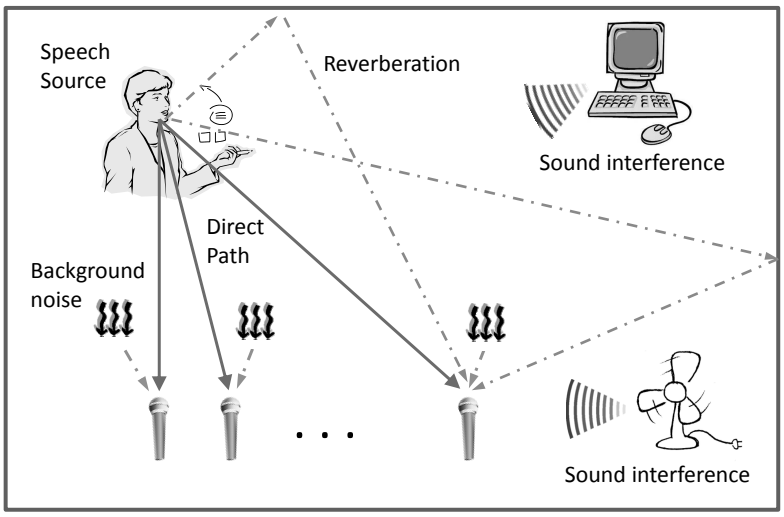


Figure 1.1: A hands-free speech communication environment.

move freely without any physical constraint. By deploying a microphone array which consists of multiple microphones, the problem of ASLT is defined as the estimation of either positions or directions from which the speech signals originate. In general, estimation of the source position/direction is often required before enhancing the quality of a speech signal emanating from a distant location via the use of beamforming algorithms [2], or to improve the speech recognition performance given the enhanced signal. In a teleconference scenario, one may also be interested in steering a camera automatically to the person who is speaking [3, 4]. The estimated locations could be useful for diarization applications. In the domain of robotics, ASLT is used to imitate the human auditorial system and steers the robot's eyes to the person during a robot-human conversation [5]. In addition, the ASLT system can be applied for a surveillance purpose. The microphones can be deployed in a public place to detect and locate, for example, a gunshot or a human scream [6].

This thesis considers the problem of ASLT in a room environment, in the presence of room reverberation, background noise and sound interference. Although many applications have been derived, localization and tracking of acoustic sources in a room environment is still an open problem. As shown in Fig. 1.1, the challenges

of ASLT arise given the fact that the microphones capture not only the direct-path component of the source signal but also the reverberant components reflected at room boundaries. In addition, the background noise and competing sound interference may also severely degrade the received signals and hence the performance of localization and tracking. Due to the non-stationarity of the speech signal, ASLT algorithms may suffer from performance degradation during frames with low signal-to-noise ratios (SNRs) and/or signal-to-reverberation ratios (SRRs). In multiple-source scenarios, localization and tracking may be more challenging due to the fact that the source spectra can be overlapped with each other, and the number of sources at each time instance may be unknown.

In order to localize and track an acoustic source, a microphone array that consists of multiple omni-directional microphones is often used. Different microphone array configurations have been employed in recent literature. For instance, the linear array and circular array have been widely used in time-difference-of-arrival (TDOA) [7] and beamforming-based approaches [8, 9]. To achieve better estimate of the two-dimensional source position $[x, y]^{\mathsf{T}}$ in a tracking scenario, the distributed microphone array configuration is commonly used where several microphone arrays are placed along the room perimeters [10, 11]. In addition to the omni-directional microphones, an acoustic vector sensor (AVS), which consists of one monopole pressure sensor element collocated with three orthogonally oriented dipole elements, has also drawn much interest. Due to the directivity property of the differential microphones, a single AVS can achieve direction-of-arrival (DOA) estimation. Due to its small size, the AVS can be advantageous in surveillance applications.

In this thesis, the localization and tracking problems are distinguished by specifically referring localization as estimating the position/direction using the signal independently at each time frame, while tracking refers to the estimation of the position/direction sequentially over several consecutive time frames. Therefore,

source tracking can be viewed as a fusion of localization measurements from independent frames by considering the consistency of these measurements. Although both terms can be used for stationary or moving source, the difference essentially lies on whether the consistency of source motion is considered.

1.2 Organization of the thesis

The scope of this thesis is divided into two parts: the ASLT problem using the conventional distributed microphone arrays is discussed in Part I; in Part II, the focus is shifted to the DOA estimation of acoustic sources using the emerging acoustic vector sensor technology. The details of each chapter is summarized as follows:

Part I

Chapter 2 reviews existing literatures of ASLT using conventional microphone arrays. The time-difference-of-arrival (TDOA) and steered-response-power beamformer (SRP beamformer) based algorithms for source localization will firstly be reviewed. The particle filter (PF) based framework which considers the TDOA and SRP-beamformer measurements across several consecutive frames will then be introduced. Due to the fusion of measurements across consecutive frames and the incorporation of source motion model, the PF framework transforms the localization problem into a tracking problem and estimates the trajectory of the source across several time frames. This PF based framework serves as the basis and benchmark for several proposed algorithms in the following chapters.

Chapter 3 presents a proposed algorithm for single-source tracking in a reverberant and noisy environment. This chapter first shows that the performance of the existing sequential importance resampling PF (SIRPF) tracking framework with SRP beamformer measurement degrades significantly in the presence of rever-

beration and noise in room environment. This is because the SIRPF-SRP tracking algorithm uses SRP beamformer spatial spectrum as the approximation of the measurement likelihood, which may not be optimal in a reverberant and noisy environment. To this end, a new SIRPF based tracking algorithm is proposed in which the measurement likelihood function is derived based on a new regional SRP beamformer (RSRP) function. The performance of the proposed SIRPF-RSRP algorithm is then examined under various simulated conditions with different levels of reverberation and noise.

Chapter 4 presents a proposed algorithm for tracking a single speech-source in the presence of sound interference. The existing SIRPF-SRP is not capable of distinguishing between speech source and other sound interference. To address the problem, the harmonicity feature in a speech signal is exploited in order to track only a speech source that is insensitive to other interference. By exploiting this unique speech feature, the harmonicity based SRP beamformer (HSRP) function is derived and incorporated with SIRPF framework. Simulations are conducted to verify the performance of the proposed SIRPF-HSRP algorithm in the presence of various sound interferers with various signal-to-interference ratios (SIRs).

Chapter 5 presents a proposed algorithm for tracking alternating multiple sources. An interactive classroom scenario is considered where the multiple moving/stationary speakers are active in turns during a conversation. In addition, potential sound interferers, e.g., fan noise, or interfering whisperer, may also be present at a lower volume. The first challenge of this problem lies in the fact that the abrupt change in the positions of the desired sources may occur and it requires the system to estimate the active source position rapidly. The second challenge is the presence of background noise, interference and reverberation and the algorithm has to be designed to achieve reasonable performance even for the low SNR, SIR and/or SRR frames. To achieve this, a swarm intelligence based PF (SWIPF) is proposed

which incorporates the particle swarm intelligence with PF. Both simulation and experiment are conducted to verify the performance of the proposed algorithm.

Part II

Chapter 6 serves as a review for DOA estimation using AVS. In this chapter, different signal models for AVS will firstly be formulated. These models will consider both single-source and multi-source scenarios. Various state-of-the-art DOA estimation algorithms using single AVS will then be introduced. The single-source DOA estimation algorithms serve as the basis and will be discussed first. Several multi-source DOA estimation algorithms will subsequently be used as benchmark algorithms for the comparison with the proposed multi-source DOA estimation algorithms in the following chapter.

Chapter 7 presents a proposed algorithm for DOA estimation of multiple sources using single acoustic vector sensor. A scenario where multiple speakers are simultaneously active in a noisy and reverberant environment is considered. The challenges are the overlapping of the simultaneously active source signals, as well as the presence of noise and reverberation. In this chapter, a multi-source DOA estimation algorithm is proposed. The DOA estimation is performed by identifying the low-reverberant-single-source (LRSS) zones in the time-frequency (TF) domain of the AVS received signals. These LRSS zones are expected to contain only the dominant source signal component and are less affected by reverberation and noise. Simulation results show that the proposed algorithm achieve lower errors than existing algorithms in a reverberant and noisy environment.

Chapter 8 finally summarizes the thesis and proposes several directions for future research.

1.3 Contributions of the thesis

Contributions made by the author are mainly described in Chapter 3 to Chapter 5 in Part I and Chapter 7 in Part II.

In Chapter 3 and Chapter 4, the author's contributions are on the topic of single-source tracking. It is well known that the performance of a single-source tracking algorithm degrades when reverberation and noise exist, or in the presence of sound interference. The SIRPF-RSRP and SIRPF-HSRP algorithms are proposed, respectively, in these two chapters for addressing these issues. Both of these algorithms exploit the SIRPF tracking framework which has been commonly adopted for single-source tracking. The commonality of these two algorithms also lies in the fact that both algorithms are aimed at deriving suitable approximations of the measurement likelihood for different considered environments.

For the proposed SIRPF-RSRP algorithm in Chapter 3, the underlying objective is to derive an approximation of the measurement likelihood for a reverberant and noisy environment using the RSRP beamformer function. In the conventional SIRPF-SRP algorithm, the measurement likelihood is derived based on the SRP beamformer function, which evaluates the power at discrete steered positions. This SRP beamformer function will, however, result in spurious peaks caused by reverberation and noise. This, in turn, affects the approximation of the measurement likelihood. The proposed tracking algorithm, on the other hand, exploits the RSRP beamformer function for approximating the measurement likelihood. The RSRP beamformer function integrates powers in the neighborhood region of each steered position, which results in a smoothing effect over the spatial spectrum. This smoothing effect has shown to be able to remove the spurious peaks caused by reverberation and noise. A more accurate approximation of measurement likelihood can therefore be derived and incorporated into SIRPF framework to achieve source tracking in a

reverberant and noisy environment.

For the proposed SIRPF-HSRP algorithm in Chapter 4, the objective is to derive an approximation of the measurement likelihood for the scenario where sound interference exist using the HSRP beamformer function. As opposed to the use of the non-discriminative SRP beamformer function in SIRPF-SRP algorithm, a HSRP beamformer function is derived which exploits the harmonicity nature of a typical speech signal. To achieve feature extraction, a multi-band excitation (MBE) fitting method is applied which estimates the harmonic bands belonging to the speech component from received signals. These extracted harmonic features are then used to derive the HSRP beamformer function. Due to the emphasis on only the speech harmonics, the HSRP beamformer function is only sensitive to the speech source and insensitive to other interference. The speech sensitive measurement likelihood can therefore be derived using the proposed HSRP beamformer function and incorporated into SIRPF framework for speech source tracking in the presence of interference.

In Chapter 5, the scenario where several speakers are active in turns during a conversation in a noisy and reverberant interactive classroom environment is considered. It is known that formulating the source dynamic model is challenging due to the unpredictable alternation between source positions. This implies that the direct application of the existing SIRPF framework will not achieve good performance since SIRPF relies significantly on the suitability of the source dynamic model. One example is that when a single-source model is used in SIRPF while the alternation occurs in practice, the particle sampling impoverishment problem will occur and this leads to the system not being able to quickly release the inactive source and switch to the newly active source. To address this problem, the extended Kalman particle filter (EKPF) uses an additional Kalman filter which takes into account the latest measurement information in order to compensate the model mismatch [12]. However, sampling of the particles is still sub-optimal if these measurements are erroneous

during frames affected by noise, interference and reverberation. In this chapter, a SWIPF tracking framework is proposed which jointly exploits both advantages of particle swarm intelligence and the particle filter. The PF framework is used as a sequential state estimator which is suitable for formulating the basic tracking problem. The limitation of PF, which lies in the particle sampling problem, is solved by the incorporation of interaction and memory mechanisms in particle swarm intelligence. The memory mechanism can be exploited to make the particles retain at their best-fit positions when measurements are corrupted by noise, interference and reverberation. Convergence of the particles is expected to be improved since, with inter-particle interaction, particles can be directed towards the newly active source region by sharing the information among themselves.

In Chapter 7, a multi-source DOA estimation algorithm using a single AVS is proposed. It is well-known that multi-source DOA estimation in a room environment is challenging due to reflections, background noise and overlapping of the source signals. Existing multi-source DOA estimation algorithms are derived based on the free-space model which may not be suitable for a reverberant environment. In addition, compared to the use of conventional microphone array, the advantage of using AVS for DOA estimation currently is still under exploration in the research community. This chapter starts from modelling the AVS received signals under reverberant condition. As opposed to the conventional microphone array, the unique structure of AVS, i.e., the co-location of the sensor elements can be exploited for deriving a DOA estimator which is robust against noise and reverberation. Specifically, in the proposed algorithm, the DOA estimation is achieved by first identifying the low-reverberant-single-source (LRSS) zones in the TF domain of the received signals, which are expected to contain the dominant source signal component with a high signal-to-reverberation ratio. By using only these LRSS zones for DOA estimation, the accuracy is expected to be improved.

1.4 Additional contribution

Apart from the proposed algorithms discussed above for ASLT application, the author also has contribution on the topic of noise reduction using single microphone for speech enhancement purpose [6].

Part I

Acoustic Source Localization and Tracking Using Microphone Arrays

Chapter 2

Literature Review

This chapter starts with the signal model for the problem of acoustic source localization and tracking (ASLT), followed by a brief review of the speech processing fundamentals. The time-difference-of-arrival (TDOA) and steered-response-power beamformer (SRP-beamformer) algorithms will then be reviewed. Given a stream of received signals, the location estimates derived from either TDOA or SRP-beamformer algorithms are obtained for a set of independent time frames without considering any relation between adjacent frames. The relationship between TDOA and SRP-beamformer algorithms will be further investigated, which will later be employed for the development of the proposed algorithm in Section 3.

In the following sections, focus will be shifted to the problem of tracking a single moving/stationary source. Given the TDOA or SRP-beamformer measurements, the source tracking problem can be viewed as a measurement fusion problem. It takes into account the temporal consistency of TDOA and SRP-beamformer measurements across successive frames such that the unreliable TDOA or SRP-beamformer measurements can be filtered out, leading to a more accurate position estimate. For this purpose, the particle filter tracking model, which belongs to a class of the well-known Bayesian filter framework, will be discussed. In the following,

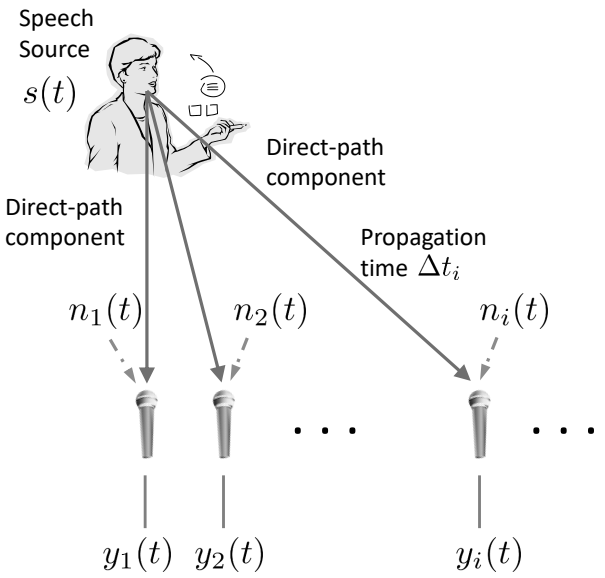


Figure 2.1: Free-space received signal model for omni-directional microphone array.

existing state-of-the-art algorithms for tracking of acoustic sources will be reviewed in detail.

2.1 Received signal model

2.1.1 Single-source free-space model

Consider a single-source free-space scenario where the sound source and microphones are placed in an ideal anechoic environment, as shown in Fig. 2.1. The signal $s(t)$ generated by the source is located at an unknown position $\mathbf{r}^{\text{src}} = [x,\ y]^{\mathsf{T}}$ where $(\cdot)^{\mathsf{T}}$ denotes matrix transpose, and N number of microphones distributed with known positions $\mathbf{r}_i^{\text{mic}}$ where $i = 1 \dots N$ denotes the microphone index. The source signal $s(t)$ radiates away from the source and the sound pressure level reduces as a function of distance from the source position. The signal captured by the ith microphone at time instant t can therefore be expressed as an attenuated and delayed version of

the original signal given by

$$y_i(t) = \alpha_i s(t - \Delta t_i) + n_i(t), \tag{2.1}$$

where $0 \leq \alpha_i \leq 1$ is the attenuation factor due to signal propagation, $n_i(t)$ is an additive noise at the ith microphone and Δt_i is the propagation time from the source to the ith microphone. Both the attenuation factor α_i and the propagation time Δt_i can, in theory, be used as localization cues since they are functions of the source position. The attenuation factor α_i is denoted by [13]

$$\alpha_i \propto \frac{1}{||\mathbf{r}_i^{\text{mic}} - \mathbf{r}^{\text{src}}||}, \tag{2.2}$$

where $||\cdot||$ denotes the Euclidean norm and the proportion sign indicates the attenuation is inversely proportional to source-sensor distance. The propagation time is denoted by

$$\Delta t_i = \frac{||\mathbf{r}_i^{\text{mic}} - \mathbf{r}^{\text{src}}||}{c}, \tag{2.3}$$

where c is the speed of sound. In a binaural microphone setup where only two microphones are used to mimic the human ears, α_i and Δt_i between the two microphones are generally known as the interaural level difference and interaural time difference cues, and the source direction can be estimated by exploiting these cues [14, 15]. However, it is noted that the attenuation factor α_i is subject to microphone gain, which in practise, may not be exactly calibrated to be normalized across many microphones. In addition, α_i is sensitive to noise [16] such that the localization accuracy may reduce when a non-stationary noise environment is encountered. Therefore, in a microphone array setup and by ignoring the signal scaling factor, α_i can be dropped and only the time-delay information Δt_i is considered for localizing the

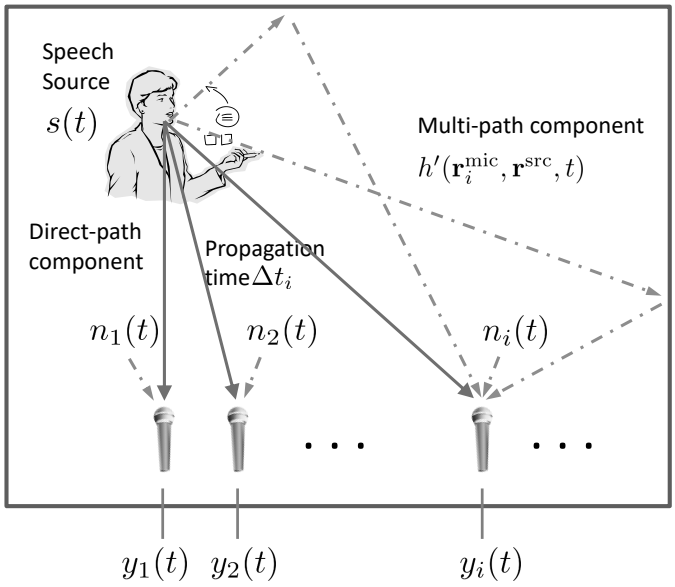


Figure 2.2: Reverberant received signal model for omni-directional microphone array.

source. Equation (2.1) can therefore be simplified as [17]

$$y_i(t) = s(t - \Delta t_i) + n_i(t). \tag{2.4}$$

Equation (2.4) is hereby referred as the *single-source free-space model*.

2.1.2 Single-source reverberant model

Consider a more realistic scenario where localization is to be performed in an enclosed environment, as shown in Fig. 2.2. A source signal $s(t)$ emanates from position $\mathbf{r}^{\text{src}}$. In this enclosed environment, microphones with known positions $\mathbf{r}_i^{\text{mic}}$ will capture not only the direct-path propagated signal but also any reflected signals from the room boundaries. To formulate both the direct-path component and the reflections in this case, the signal propagation from the source position to the ith microphone can be modelled as a linear time-invariant system [12]. Defining $h(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$ as the impulse response between the source and the ith microphone, and $*$ as linear

convolution, the microphone received signal can be expressed as

$$y_i(t) = s(t) * h(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t) + n_i(t), \tag{2.5}$$

where $n_i(t)$ denotes the additive noise. The room impulse response $h(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$ specifies all information for the propagation channel from the source position to the microphones and will vary for different source and microphone positions. In addition, $h(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$ can further be decomposed into direct- and multi-path components. Equation (2.5) can therefore be rewritten as [18]

$$y_i(t) = s(t - \Delta t_i) + s(t) * h'(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t) + n_i(t), \tag{2.6}$$

where the term $s(t - \Delta t_i)$ is the direct-path propagation component which has been defined in (2.4) and $h'(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$ in (2.5) denotes the remaining impulse response without the direct-path component. Equation (2.5) is referred as the *single-source reverberant model*.

Given the single-source free-space model in (2.4) and single-source reverberant model in (2.5), one must note that (2.4) is used only for analytical purpose or when the reverberation is negligible. For an enclosed environment being considered in this thesis, however, considerable amount of reverberation always exists. It is well-known that the multi-path component $h'(\mathbf{r}_i^{\text{mic}}, \mathbf{r}^{\text{src}}, t)$ and noise $n_i(t)$ in (2.6) severely degrade both localization and tracking performance. To exploit the time-delay information Δt_i at each microphones for localization, the TDOA and SRP beamformer based approaches will be discussed in Section 2.3.

2.2 Speech processing basics

The speech signal itself is non-stationary and its underlying statistics is time-varying. In speech signal processing, received signals are often processed in frames, and in each frame the speech signals are assumed to be stationary. By defining T as the frame length and no overlapping between consecutive frames, the kth frame of the received signal $y_i(t)$ can be denoted as a vector

$$\mathbf{y}_i(k) = [y_i(kT),\ y_i(kT+1),\ \ldots,\ y_i(kT+T-1)]^{\mathsf{T}}. \tag{2.7}$$

where $(\cdot)^{\mathsf{T}}$ denotes matrix transpose. In ASLT, the source is assumed stationary within each frame and so is the room impulse response [16]. The aim of ASLT is therefore to estimate the source position $\mathbf{r}_k^{\text{src}}$ at each frame k. In general, the speech signal is processed in the time-frequency domain using short-time Fourier transform (STFT). Consider the single-source free-space model in (2.4) for example, the time-frequency representation of the received signal $y_i(t)$ can be written as

$$\underline{y}_i(\omega, k) = \underline{s}(\omega, k)e^{\jmath\omega\Delta t_i} + \underline{n}_i(\omega, k), \tag{2.8}$$

where $\underline{y}_i(\omega, k)$ is the STFT coefficient of the received signal $y_i(t)$, $\underline{s}(\omega, k)$ is the STFT coefficient of the source signal $s(t)$ and $\underline{n}_i(\omega, k)$ is the STFT coefficient of the noise $n_i(t)$. The variable ω denotes the angular frequency.

2.3 Source Localization Algorithms

In this section, the time-difference-of-arrival (TDOA) and the steered response power (SRP) algorithms for source localization will be reviewed. The common feature between these two algorithms is the exploitation of time-delay information (phase-

delay if it is in frequency domain). In addition, both algorithms process the received signal frames independently and can therefore be used as the location measurements at individual time frames. The fusion of the measurements across successive frames will lead to the tracking problem and will be discussed in Section 2.4.

2.3.1 Time-difference-of-arrival based localization

The TDOA-based approach partitions the N number of microphones into M number of pairs and computation is based on each microphone pair independently. Any two microphones from the microphone array can form a pair. In a case where all the combinations of two microphones are considered, the full collection of all the pair combinations Υ^{full} can be expressed as

$$\Upsilon^{\text{full}} = \big\{(i,j)|i \leq N, j \leq N, i \leq j\big\}. \tag{2.9}$$

The total number of pairs is $M = \frac{N(N-1)}{2}$. In some works, with consideration of inter microphone distance and computational complexity, a subsect of the full collection can also be considered, e.g,

$$\Upsilon^{\text{sub}} = \big\{(i,j)|j - i = 1\big\}, \tag{2.10}$$

in which only adjacent microphones are considered as a pair.

The TDOA based approach is usually known as a dual-step approach since it first estimates the relative time delay for each microphone pair. The estimated TDOAs are then used to triangulate the source position. TDOA estimation from the received signals is generally one of the most challenging steps due to the presence of reverberation and background noise in the received signals as indicated in (2.6). There are a variety of techniques proposed for TDOA estimation, and an overview

of TDOA estimation techniques can be found in [18].

The generalized cross-correlation (GCC) method is one of the earliest TDOA estimation algorithms and has gained popularity since the landmark paper [19] published by Knapp and Carter in 1976. It was derived using the single-source free-space model in (2.4) for simplicity. With reference to (2.4), the TDOA between the i- and jth microphones is defined as

$$\begin{aligned}\tau_k^{(i,j)} &= \Delta t_j - \Delta t_i \\ &= \frac{1}{c}\big(||\mathbf{r}_k^{\text{src}} - \mathbf{r}_j^{\text{mic}}|| - ||\mathbf{r}_k^{\text{src}} - \mathbf{r}_i^{\text{mic}}||\big) \\ &= \mathcal{T}(\mathbf{r}_k^{\text{src}}).\end{aligned} \tag{2.11}$$

where $\mathbf{r}_k^{\text{src}}$ and $\mathbf{r}_i^{\text{mic}}$ denote the source and microphone positions, respectively, $||\cdot||$ denotes Euclidean distance, and c is the speed of sound. The function $\mathcal{T}(\cdot)$ denotes the TDOA function which implies that the source position is a non-linear function for each defined microphone pair. In order to estimate $\tau_k^{(i,j)}$ given the received signal frames $\mathbf{y}_i(k)$ and $\mathbf{y}_j(k)$, the GCC function between the ith and jth microphones can be computed by [19]

$$\Psi_k^{(i,j)}(\tau) = \int_{\Omega} w(\omega,k)\underline{y}_i(\omega,k)\underline{y}_j^*(\omega,k)e^{\jmath\omega\tau}d\omega, \tag{2.12}$$

where $\underline{y}_i(\omega,k)$ is the STFT coefficients of the ith microphone received signal, ω is the angular frequency, Ω is the frequency range that speech signal mainly exists, $(\cdot)^*$ denotes complex conjugation and $w(\omega,k)$ is a weighting function. The TDOA is then estimated by searching for the time delay corresponding to the highest cross-correlation value, i.e.,

$$\widehat{\tau}_k^{(i,j)} = \arg\max_{\tau} \Psi_k^{(i,j)}(\tau). \tag{2.13}$$

Figure 2.3 shows a typical example of the GCC function for one pair of

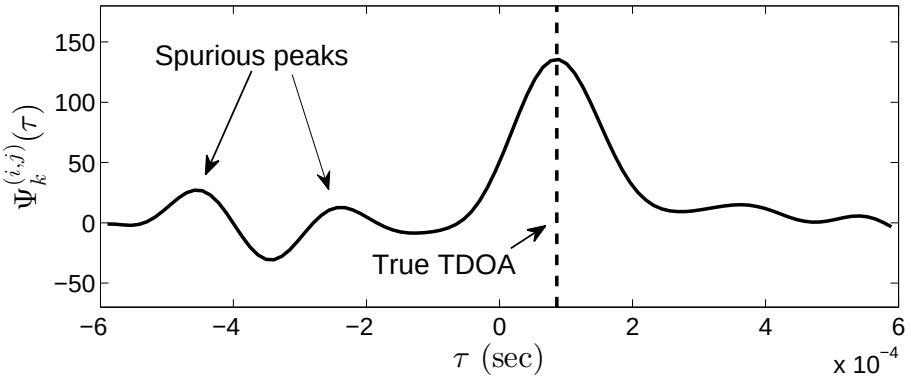


Figure 2.3: GCC-PHAT function computed for a pair of microphone received signals in a simulated environment with $T_{60} = 0.3$ s and SNR = 15 dB.

microphones computed using a simulated environment with a reverberation time $T_{60} = 0.3$ s and signal-to-noise ratio (SNR) of 15 dB. The actual TDOA, shown by the dash line, is computed given the true locations of the source and microphones. In can be observed that the GCC function, shown by the solid line, has a maximum corresponding to the actual TDOA, although some spurious peaks occur due to reverberation and noise. Therefore, by searching for the delay corresponding to the maximum of $\Psi_k^{(i,j)}(\tau)$ as indicated in (2.13), the TDOA can be estimated. It is worth noting that as reverberation and noise increase, the magnitudes of the spurious peaks may exceed that of the actual TDOA which renders source localization a challenging task.

There are a variety of member algorithms within the GCC family depending on how the weighting function $w(\omega, k)$ is defined. Commonly used weighting functions include the smoothed coherence transform [20], the maximum-likelihood processor [19] and the phase transform (PHAT) [19]. The well-known PHAT weighting function is defined as [19]

$$w^{\text{PHAT}}(\omega, k) = \frac{1}{|\underline{y}_i(\omega, k)\underline{y}_j(\omega, k)|}, \tag{2.14}$$

and by substituting it into (2.12), it can be noted that the PHAT weighting scheme

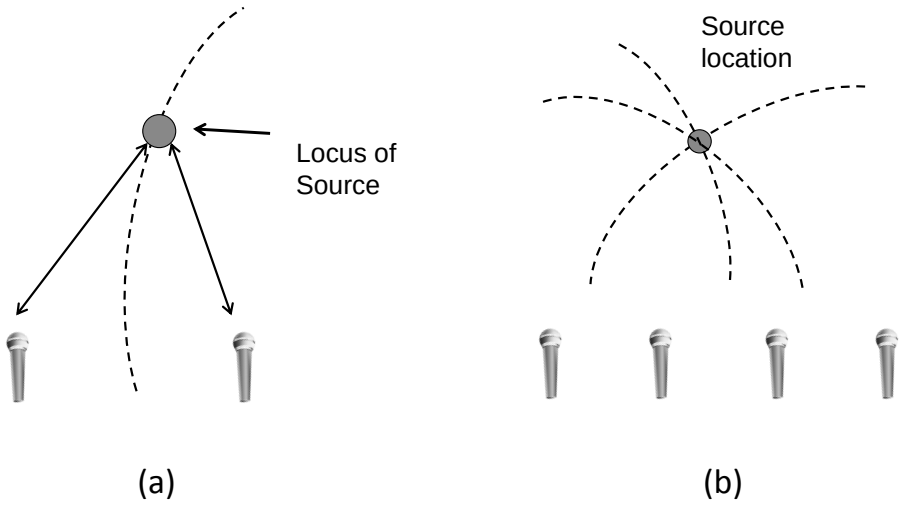


Figure 2.4: Illustration of 2D source position estimation using TDOAs. (a) TDOA between the two microphone signals defines a hyperbola. (b) Source location lies on the intersection of a set of hyperbolae.

removes the dependence of $\Psi(\tau)$ on source energies and emphasizes only on the phase of the cross-correlation spectrum containing time-delay information. Therefore, the PHAT weighting scheme exhibits, to some extent, robustness against reverberation and noise and has been widely used in recent literature [18].

After obtaining all the estimated TDOAs $\{\widehat{\tau}_k^{(i,j)} | (i,j) \in \Upsilon\}$ for a set of pairs Υ, the second step of position estimation involves solving a geometry problem. Note that for each pair of microphones, the range difference can be obtained from the estimated TDOA, leading to a hyperbola on which the source potentially lies as illustrated in Fig. 2.4 (a). Given a microphone array with a set of TDOA information, the source position $\mathbf{r}_k^{\text{src}}$, therefore, corresponds to the intersection of several hyperbolae as shown in Fig. 2.4 (b). In practice, due to the presence of errors in TDOA estimates, a single intersection may not exist. The source position is therefore often estimated by searching for the position that best fits the potential loci across all microphone pairs. Several algorithms proposed for the determination of the best fit can be found in [4,21]. The best-fit source position is often estimated by minimizing the cost function of fitting error. Suppose that the TDOA set $\{\widehat{\tau}_k^{(i,j)} | (i,j) \in \Upsilon\}$ is

estimated given an array of microphones, this cost function is defined as

$$\mathcal{E}(\mathbf{r}) = \sum_{(i,j)\in\Upsilon} \left[\widehat{\tau}_k^{(i,j)} - \mathcal{T}(\mathbf{r})\right]^2, \tag{2.15}$$

where the non-linear function $\mathcal{T}(\cdot)$ has been defined in (2.11) and $\mathbf{r} = [x \;\; y]^\mathsf{T}$ is the assumed source position. The iterative approach is often employed which updates $\mathbf{r}$ from an initial position until it achieves the source position associated with the minimum of cost function, i.e.,

$$\widehat{\mathbf{r}}_k^{\text{src}} = \arg\min_{\mathbf{r}} \mathcal{E}(\mathbf{r}). \tag{2.16}$$

An example of such iterative approach is the Levenberg-Marquardt algorithm [22,23].

It worth noting that in the source tracking framework, as will be discussed in Section 2.4, the TDOAs $\{\widehat{\tau}_k^{(i,j)} | (i,j) \in \Upsilon\}$ can be directly used as measurements. Therefore, the minimization steps for triangulating the source position in (2.15) and (2.16) are not required within the tracking framework.

2.3.2 SRP beamformer based localization

In contrast to the TDOA based approach, the SRP beamformer based approach is known as a one-step approach since it estimates the source location within a single step without the need for time-delay estimation. Initially, the beamformer is developed as a spatial filter which reinforces the source signal emanating from a given direction while suppressing the interference from other directions. The beamformer is subsequently used as a localization algorithm by steering the beamformer across the region of interest and computing the power for each steered location. As shown in Fig. 2.5, a spatial power spectrum can then be obtained in which the location with the highest power corresponds to the source position estimate. The family of

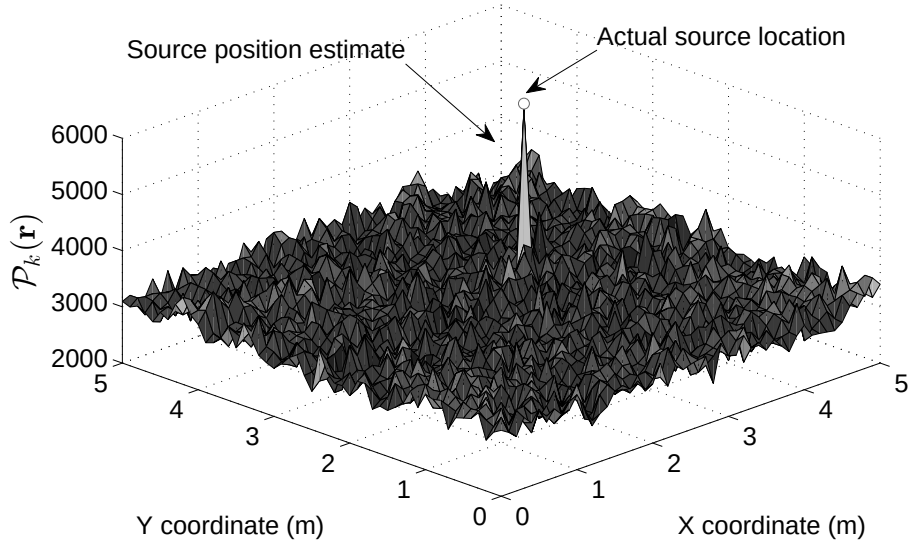


Figure 2.5: The spatial specturm of a SRP beamformer computed using eight microphones distributed along the room perimeters (two microphones are placed at each side of the room). The received signals are generated using simulation with $T_{60} = 0.3$ s and SNR = 15 dB.

beamformer includes, for example, steered response power (SRP) [24,25], minimum variance distortionless response [26], linearly constrained minimum variance [13,26].

The well-known SRP beamformer has gained popularity due to its simplicity in implementation. Considering N number of microphones, the SRP function defines the power for a steered location $\mathbf{r}$ given by

$$\mathcal{P}_k(\mathbf{r}) = \int_{\Omega} \left| \sum_{i=1}^{N} w(\omega, k) \underline{y}_i(\omega, k) e^{\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_i^{\text{mic}}\|}{c}} \right|^2 d\omega, \tag{2.17}$$

where $\mathbf{r} = [x \ y]^{\mathsf{T}}$ is the steered location in the region of interest, $w(\omega, k)$ is the weighting function similar to the one used in GCC function, $\underline{y}_i(\omega, k)$ is the STFT coefficient of the ith microphone received signal $y_i(t)$, c is the speed of sound and Ω is the frequency range in which speech signal mainly exists. In (2.17), the time delays $\|\mathbf{r} - \mathbf{r}_i^{\text{mic}}\|/c$ from the steered position $\mathbf{r}$ to all the microphones are first computed. The corresponding power of the steered position $\mathcal{P}_k(\mathbf{r})$ is then computed by phase-aligning the signal according to the signal delays. Similar to the GCC approach, the

PHAT weighting scheme is applied, i.e.,

$$w^{\text{PHAT}}(\omega, k) = \frac{1}{|\underline{y}_i(\omega, k)|}, \tag{2.18}$$

and by substituting it into (2.17), the signal power will be neglected and the SRP is computed based on the phase-delay information. Given the SRP definition in (2.17), the source position is estimated by steering the beamformer across the enclosure domain and searching for the position corresponding to the maximum power, i.e.,

$$\widehat{\mathbf{r}}_k^{\text{src}} = \arg\max_{\mathbf{r}\in\mathcal{D}} \mathcal{P}_k(\mathbf{r}). \tag{2.19}$$

where $\mathcal{D} = \{x, y | x_{\min} \leq x \leq x_{\max}, y_{\min} \leq y \leq y_{\max}\}$ is the considered enclosure domain.

Figure 2.5 shows an illustrative example of a SRP beamformer spatial spectrum using simulated data. The room dimension was 5 m $\times$ 5 m $\times$ 3 m and eight microphones were employed along the room perimeters (two microphones at each side). The environment parameters were set as $T_{60} = 0.3$ s and SNR = 15 dB. It can be observed that the SRP spectrum exhibits a maximum at the actual source location. The search process in (2.19) will therefore be able locate the source position. However, similar to the GCC function, performance of the SRP beamformer will also be degraded when the amount of reverberation and noise increases since, under such scenario, spurious peaks may occur.

It worth noting that the SRP beamformer based approach, in general, requires high computational complexity due to the search process over the enclosure domain of interest in (2.19). Such computational load is dependent on the resolution of the grid in which the beamformer is steered. Take a 5 m $\times$ 5 m enclosure for example with with a resolution of 0.1 m. This results in a 50×50 grid such that computation of the SRP function given by (2.17) for 2500 times is required. To reduce the

computational burden, a coarse-to-fine search has been proposed in [27] in which a low-resolution grid is used to localize a source before a refined search is performed. The use of particle filter, as will be discussed in Section 2.4, is another alternative to reduce the complexity since the SRP evaluation of (2.17) can be confined to a number of discrete particle positions (typically 100 to 300 particle positions at each frame) [28].

2.3.3 Relationship between TDOA and the SRP beamformer

In the previous sections the conventional TDOA and SRP beamformer based approaches have been reviewed. In this section, the implicit relationship between these two approaches will be reviewed [16]. Consider a given steered location $\mathbf{r} = [x, y]^{\mathsf{T}}$, with reference to (2.17) and (2.18), the SRP definition can be expanded as

$$\begin{aligned}
\mathcal{P}_k(\mathbf{r}) &= \int_{\Omega} \left| \sum_{i=1}^{N} \frac{\underline{y}_i(\omega, k)}{|\underline{y}_i(\omega, k)|} e^{\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_i^{\text{mic}}\|}{c}} \right|^2 d\omega \\
&= \int_{\Omega} \left(\sum_{i=1}^{N} \frac{\underline{y}_i(\omega, k)}{|\underline{y}_i(\omega, k)|} e^{\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_i^{\text{mic}}\|}{c}} \right) \left(\sum_{j=1}^{N} \frac{\underline{y}_j^*(\omega, k)}{|\underline{y}_j(\omega, k)|} e^{-\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_j^{\text{mic}}\|}{c}} \right) d\omega \\
&= \int_{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\underline{y}_i(\omega, k)\underline{y}_j^*(\omega, k)}{|\underline{y}_i(\omega, k)\underline{y}_j(\omega, k)|} e^{-\jmath\omega\mathcal{T}(\mathbf{r})} d\omega \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{\Omega} \frac{\underline{y}_i(\omega, k)\underline{y}_j^*(\omega, k)}{|\underline{y}_i(\omega, k)\underline{y}_j(\omega, k)|} e^{-\jmath\omega\mathcal{T}(\mathbf{r})} d\omega,
\end{aligned} \tag{2.20}$$

where $\mathcal{T}(\cdot)$ is the non-linear TDOA function defined by

$$\mathcal{T}(\mathbf{r}) = \frac{1}{c}\big(||\mathbf{r} - \mathbf{r}_j^{\text{mic}}|| - ||\mathbf{r} - \mathbf{r}_i^{\text{mic}}||\big). \tag{2.21}$$

In the second expression of (2.20), the indexes i and j are used separately for the purpose of linking with the TDOA definition in the following discussion. Note that

in the last expression of (2.20), microphone indexes between i and j are overlapped in the two summations. The overlapping component with $i = j$ can be defined as

$$\begin{aligned}\mathcal{P}_0 &= \sum_{i=1}^{N} \int_{\Omega} \frac{\underline{y}_i(\omega,k)\underline{y}_i^*(\omega,k)}{|\underline{y}_i(\omega,k)|^2} d\omega \\ &= N\Omega, \end{aligned} \tag{2.22}$$

which is a constant and not related to the time-delay information. Equation (2.20) can be rewritten as

$$\mathcal{P}_k(\mathbf{r}) = \mathcal{P}_0 + \sum_{\substack{i=1\ldots N,\\ j=1\ldots N,\\ i\neq j}} \int_{\Omega} \frac{\underline{y}_i(\omega,k)\underline{y}_j^*(\omega,k)}{|\underline{y}_i(\omega,k)\underline{y}_j(\omega,k)|} e^{-\jmath\omega\mathcal{T}(\mathbf{r})} d\omega. \tag{2.23}$$

On the other hand, with reference to (2.12) and (2.14), the GCC function $\Psi_k^{(i,j)}(\tau)$ can be written as a function of the steered location $\mathbf{r}$ as

$$\Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})) = \int_{\Omega} \frac{\underline{y}_i(\omega,k)\underline{y}_j^*(\omega,k)}{|\underline{y}_i(\omega,k)\underline{y}_j(\omega,k)|} e^{\jmath\omega\mathcal{T}(\mathbf{r})} d\omega. \tag{2.24}$$

In addition, A symmetric property can be observed for the cross-correlation given by

$$\Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})) = \Psi_k^{(j,i)}(-\mathcal{T}(\mathbf{r})). \tag{2.25}$$

Now considering the full set of microphone pairs defined as

$$\Upsilon^{\text{full}} = \big\{(i,j)|i \leq N, j \leq N, i \leq j\big\}, \tag{2.26}$$

and substituting (2.24) and (2.25) into (2.23), the SRP function can be rewritten as

$$\mathcal{P}_k(\mathbf{r}) = \mathcal{P}_0 + 2 \sum_{(i,j)\in\Upsilon^{\text{full}}} \Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})). \tag{2.27}$$

From (2.27), it can be observed that, in the absence of noise and reverberation, searching for a position $\mathbf{r}$ that maximizes the steered response power $\mathcal{P}_k(\mathbf{r})$ is equivalent to searching for a position that maximizes the sum of the GCC values across all microphone pairs corresponding to $\mathbf{r}$.

2.4 Acoustic source tracking

Conventional TDOA and SRP beamformer measurements estimate the source position for each data frame independently without considering the relationship between successive frames. However, it is well-known that these TDOA or SRP beamformer measurements exhibit temporal consistency across consecutive time frames particularly since the position of the source varies consistently in a moving-source scenario or when the source is stationary. Consider a simple case where a single source is moving with a constant speed, the estimated TDOA or SRP beamformer measurement at previous time frame $k-1$ should, to some extent, be consistent with measurements estimated at current time frame k. This temporal information should be utilized to filter out unreliable/erroneous TDOA or SRP beamformer measurements caused by noise and reverberation. To exploit such temporal consistency, the Bayesian filter theory is first introduced in the next section which forms the foundation of formulating the sequential state estimation problem. The particle filter, which belongs to the Bayesian filter family, will then be discussed for tracking of acoustic sources.

2.4.1 Bayesian filter model

The Bayesian filter theory considers the problem as illustrated in Fig. 2.6. One would like to estimate an unobservable state $\mathbf{x}_k$ sequentially at time frames $k = 1, 2, \ldots K$, given that K denotes the total number of frames. In addition, the state-transition

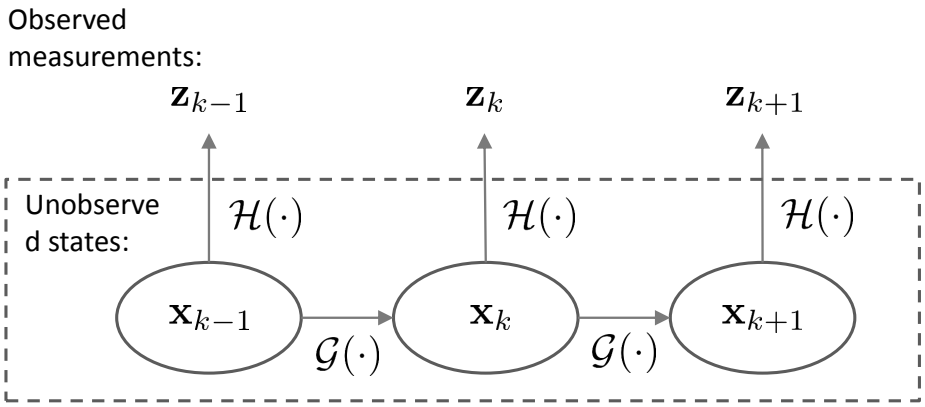


Figure 2.6: Illustration of the Bayesian filtering problem.

function (or sometimes known as the process function) $\mathcal{G}(\cdot)$ is, to some extent, assumed to be known with certain probabilistic uncertainties. At each time frame, although the direct access to the state $\mathbf{x}_k$ is not available, measurement $\mathbf{z}_k$, which is related to $\mathbf{x}_k$ is available subject to some measurement noise. The relationship between $\mathbf{z}_k$ and $\mathbf{x}_k$ is defined by a measurement function $\mathcal{H}(\cdot)$, which is assumed to be known. The above problem can be represented by the following state-space equations

$$\mathbf{x}_k = \mathcal{G}(\mathbf{x}_{k-1}, \mathbf{u}_k), \tag{2.28}$$

$$\mathbf{z}_k = \mathcal{H}(\mathbf{x}_k, \mathbf{w}_k), \tag{2.29}$$

where $\mathbf{u}_k$ is the process noise defining any uncertainty in the state-transition, and $\mathbf{w}_k$ is the measurement noise defining any errors in measurement.

The objective of Bayesian filtering is to estimate $\mathbf{x}_k$ iteratively at every time frame. These estimates are denoted by $\widehat{\mathbf{x}}_k$ throughout the thesis. To achieve good estimates, statistical approaches aim to estimate the posterior probability density function (pdf) $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ which describes the probability of $\mathbf{x}_k$ given the measurements $\mathbf{z}_{1:k}$ from the first frame up to the frame k. Sequentially, given $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$

at frame $k-1$, the state-transition probability $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ can be derived from a given process function $\mathcal{G}(\cdot)$ in (2.28). Similarly, the measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_{k-1})$ can be derived using $\mathcal{H}(\cdot)$ in (2.29). Therefore $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ for the current time frame index k can be computed by the following recursion

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1}, \tag{2.30}$$

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \propto p(\mathbf{x}_k|\mathbf{z}_{1:k-1})p(\mathbf{z}_k|\mathbf{x}_k). \tag{2.31}$$

Equation (2.30) is known as the *prediction step* since it predicts the pdf of $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ which only takes into account the state-transition probability $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. Equation (2.31) is known as the *update step* as it updates the pdf of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ by taking into account the current measurement $\mathbf{z}_k$. These two equations, however, are not implementable due to the integration over a continuous variable $\mathbf{x}_{k-1}$ in (2.30). Therefore, various algorithms have been proposed to approximate these prediction and update steps.

2.4.2 Particle filter basics

The Bayesian filtering problem in (2.30) and (2.31) can be solved by either the Kalman filter or Particle filter. Section 2.4.4 will elaborate the advantages and disadvantages of these two approaches for the ASLT problem. Here, the particle filter basics are first introduced. Particle filtering is an approximation technique used to represent the posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ via a set of particles of the state space with associated weights $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$, i.e.,

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \sum_{p=1}^{N_p} w_k^{(p)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(p)}), \tag{2.32}$$

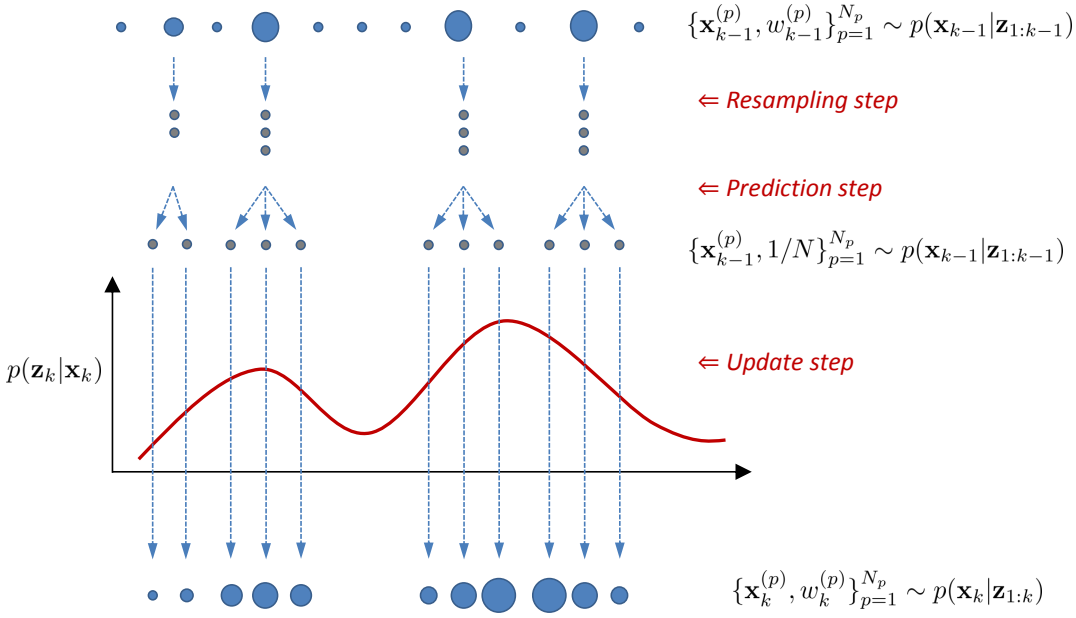


Figure 2.7: Illustration of one iteration in the particle filter. Given $\{(\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)})\}_{p=1}^{N_p}$ for approximation of $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ in the previous iteration, $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$ representing $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ for current iteration is obtained by the prediction, update and resampling steps. In the illustration, the circle denotes the particle and the size of the circle denotes its weight.

where the superscript $p = 1, \ldots, N_p$ denotes the particle index, N_p is the number of user-defined particles, $\mathbf{x}_k^{(p)}$ is the pth particle of state space, $w_k^{(p)}$ is its associated weight, and $\delta(\cdot)$ is the Dirac delta function. This approximation is often known as the Monte Carlo approximation since a number of samples (particles) are used to represent the pdf. The particle filter is therefore known as a sequential Monte Carlo solution and the iterations of (2.30) and (2.31) can now be discretized by the propagation and update steps of the approximated particle set $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$.

More specifically and as illustrated in Fig. 2.7, the Bayesian iteration in (2.30) and (2.31) can be solved as follows: suppose at time $k-1$, the set $\{(\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)})\}_{p=1}^{N_p}$ is an approximation of the posterior density $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$, the set $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$ representing $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is then obtained by a prediction step

$$\mathbf{x}_k^{(p)} \sim p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_{1:k}), \tag{2.33}$$

followed by an update step

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_{1:k})}. \tag{2.34}$$

Here $p^{(\mathrm{IS})}(\cdot)$ is the importance sampling density (otherwise known as the proposal density) from which the new particles for the following frame k are generated. After particle weight update and due to the proportionality involved in (2.34), a normalization process

$$w_k^{(p)} \Leftarrow \frac{w_k^{(p)}}{\sum_{i=1}^{N_p} w_k^{(i)}} \tag{2.35}$$

is often performed, where $\Leftarrow$ denotes assigning a new value to the variable. Finally, the obtained particle set $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$ is used to represent $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. Given $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$, the state estimate, at time frame index k, can then be computed using

$$\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \mathbf{x}_k^{(p)}. \tag{2.36}$$

Details of the above derivations can be found in [29].

One problem that may occur in PF is that after a few iterations, some of the particles may possess small weights and contribute insignificantly to the approximation of the posterior pdf. This is known as the *degeneracy problem* which results in a waste of computation [29]. To mitigate this problem, a resampling stage is often involved after each update step. Resampling is performed when the number of effective particles N_{eff} is less than a user-defined threshold N_{thr}, i.e., $N_{\mathrm{eff}} < N_{\mathrm{thr}}$, where $N_{\mathrm{eff}} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$. The resampling procedure is illustrated in Fig. 2.7. Given a fixed total number of particles, particles with large weights indicating higher significance in representing the pdf will be split into more particles, while the particles with small weights will be split into less number of particles or even be discarded. The particle weights are subsequently reset to a uniform value using $w_k^{(p)} = 1/N_p$.

Table 2.1: Summary of the Generic PF.

At time $k-1$, A set of particles $\{\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of posterior $p(\mathbf{x}_{k-1}|\mathbf{z}_{k-1})$.

For the kth frame:

1. *Particles propagation*: Draw the particles for current interaction according to the importance density,

$$\mathbf{x}_k^{(p)} \sim p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_{1:k})$$

2. *Update*: Each particle weight is updated by

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_{1:k})},$$

   followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)}(\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.

3. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\mathrm{eff}} < N_{\mathrm{thr}}$, where $N_{\mathrm{eff}} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$.

4. *Result*: the particle set $\{\mathbf{x}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$ is obtained for approximation of $p(\mathbf{x}_k|\mathbf{z}_k)$. The state estimate at the kth frame is $\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)}\mathbf{x}_k^{(p)}$.

Several resampling schemes can be found in [29]. The above discussion describes the well-known *generic PF*. A summary of the generic PF is listed in Table 2.1.

From the above discussion, a key step in PF is the particle sampling step in which particles are required to be sampled effectively for the approximation of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. In other words, one would expect

$$p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k}) \approx p(\mathbf{x}_k|\mathbf{z}_{1:k}) \tag{2.37}$$

such that the sampled particles from $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k})$ can be located in the area of significant $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. Unfortunately, such an optimal $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k})$ may not

often be estimated due to the non-linearity involved in (2.28) and (2.29). Therefore, a sub-optimal solution given by

$$p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_{1:k}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}) \tag{2.38}$$

is often used, where the state-transition density function is used as the IS density. This indicates that the particles are propagated depending only on the assumed state-transition model without taking into account the current measurement $\mathbf{z}_k$. By substituting (2.38) into (2.34), the weight update equation is simplified as

$$w_k^{(p)} \propto w_{k-1}^{(p)} p(\mathbf{z}_k|\mathbf{x}_k^{(p)}). \tag{2.39}$$

This is known as the *bootstrap PF* or *sequential importance resampling PF (SIRPF)* and it has been widely used in the area of ASLT due to its simplicity and efficiency [10, 11, 28, 30]. A summary of SIRPF is listed in Table 2.2.

It worth noting that although SIRPF is efficient for certain applications, the performance may degrade due to the sub-optimal choice of the IS density $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k}) = p(\mathbf{x}_k|\mathbf{x}_{k-1})$. In scenarios where a mismatch occurs between the assumed state-transition model and the actual state evolution, i.e., $p(\mathbf{x}_k|\mathbf{x}_{k-1}) \neq p(\mathbf{z}_k|\mathbf{x}_k)$, the *sampling impoverishment problem* may occur in which particles are sampled in the insignificant area of the posterior pdf. Other alternatives for approximating $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k})$ include the use of extended Kalman filter or unscented Kalman fiter [12, 31, 32]. In this thesis, a new solution will be proposed in Chapter 5 for the alternating source scenario.

Table 2.2: Summary of the sequential importance resampling PF.

At time $k-1$, A set of particles $\{\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of posterior $p(\mathbf{x}_{k-1}|\mathbf{z}_{k-1})$.

For the kth frame:

1. *Particles propagation*: Draw the particles using the prior propagation density
$$\mathbf{x}_k^{(p)} \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}).$$

2. *Update*: Each particle is then assigned a weight according to its likelihood
$$w_k^{(p)} \propto w_{k-1}^{(p)} p(\mathbf{z}_k|\mathbf{x}_k^{(p)}),$$
followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)}(\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.

3. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$.

4. *Result*: the particle set $\{\mathbf{x}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$ is obtained for approximation of $p(\mathbf{x}_k|\mathbf{z}_k)$. The state estimate at the kth frame is $\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \mathbf{x}_k^{(p)}$.

2.4.3 State-space formulation for acoustic source tracking

In ASLT, the state variable $\mathbf{x}_k$ is formulated as the location parameters of interest. The function $\mathcal{G}(\cdot)$ in (2.28) therefore defines the assumed human-motion model. The *random walk* and *Langevin process* source-dynamic models are often used to describe a realistic human motion for $\mathcal{G}(\cdot)$. The measurement variable $\mathbf{z}_k$, on the other hand, can be formulated as either SRP beamformer or TDOA measurement introduced in Section 2.3. The measurement function $\mathcal{H}(\cdot)$ in (2.29) defines the relationship between $\mathbf{z}_k$ and $\mathbf{x}_k$. Given that there are two models for formulating $\mathcal{G}(\cdot)$ and two models for formulating $\mathcal{H}(\cdot)$, four combinations in total can be derived as will be described in the following.

Random-walk process with SRP beamformer measurement: The

state vector is defined as $\mathbf{x}_k = [x_k,\ y_k]^\mathsf{T}$, where the two elements correspond to the x and y coordinates of the source location, respectively. Similarly, the measurement vector is defined as $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$, where $\widehat{x}_k$ and $\widehat{y}_k$ denote the source coordinates estimated from the maximum search in the spatial spectrum of the SRP beamformer in (2.19). The state-space equations in (2.28) and (2.29) can be represented by

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k, \tag{2.40}$$

$$\mathbf{z}_k = \mathbf{x}_k + \mathbf{w}_k, \tag{2.41}$$

where $\mathbf{u}_k$ is a 2×1 vector denoting the process noise and $\mathbf{w}_k$ is a 2×1 vector denoting measurement noise. In the random-walk process model, $\mathbf{u}_k$ is assumed to be sampled from a Gaussian distribution, i.e., $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}_{2\times 1}, \mathbf{\Sigma})$, with zero mean and covariance $\Sigma = \sigma_{\mathbf{u}}^2 \mathbf{I}_{2\times 2}$, where $\mathbf{I}_{2\times 2}$ denotes identity matrix. The parameter $\sigma_{\mathbf{u}}^2$ defines the level of uncertainty of the source motion in both x and y directions. The measurement noise $\mathbf{w}_k$ can be assumed either Gaussian distributed or non-Gaussian distributed depending on whether a Gaussian likelihood or pseudo likelihood is used to approximate the measurement likelihood. It worth noting that although the measurement $\mathbf{z}_k$ is defined as that obtained from the full-grid search across the entire SRP beamformer spatial spectrum, it is possible to avoid such a search process by using a pseudo measurement likelihood. Details of formulating the measurement noise statistics and the measurement likelihood will be discussed in Section 2.4.5.

Random-walk process with TDOA measurement: Given the state vector $\mathbf{x}_k = [x_k,\ y_k]^\mathsf{T}$, the measurement vector is defined as $\mathbf{z}_k = [\widehat{\tau}_k^{(1,2)},\ \ldots,\ \widehat{\tau}_k^{(i,j)},\ \ldots,\ \widehat{\tau}_k^{(N-1,N)}]^\mathsf{T}$ with elements corresponding to the TDOA estimates and $(i,j) \in \Upsilon$ where Υ is a pre-defined pair collection. The state-space

equations in (2.28) and (2.29) can therefore be represented by

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k, \tag{2.42}$$

$$\mathbf{z}_k = [\mathcal{T}^{(1,2)}(\mathbf{x}_k),\ \ldots,\ \mathcal{T}^{(i,j)}(\mathbf{x}_k),\ \ldots,\ \mathcal{T}^{(N-1,N)}(\mathbf{x}_k)]^\mathsf{T} + \mathbf{w}_k, \tag{2.43}$$

where $\mathcal{T}^{(i,j)}(\cdot)$ is the non-linear TDOA function similar to (2.11) but with superscript (i, j) for denoting the pair index, given by

$$\mathcal{T}^{(i,j)}(\mathbf{x}_k) = \frac{1}{c}\big(||\mathbf{x}_k - \mathbf{r}_j^{\text{mic}}|| - ||\mathbf{x}_k - \mathbf{r}_i^{\text{mic}}||\big). \tag{2.44}$$

Similar to (2.40) and (2.41), the 2×1 vector $\mathbf{u}_k$ denotes the process noise and is assumed to be Gaussian distributed as $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}_{2\times1}, \mathbf{\Sigma})$. The vector $\mathbf{w}_k$ with dimension $M \times 1$ denotes the measurement noise, and can be assumed either Gaussian distributed or non-Gaussian distributed depending on whether a Gaussian measurement likelihood or pseudo measurement likelihood is used (Details will be described in Section 2.4.5). Compared to the SRP beamformer measurement, it can be observed that non-linearity has been introduced in (2.43) when TDOAs are used as measurement.

Langevin process with SRP beamformer measurement: In contrast to the random-walk model, the Langevin-process model exploits the physical velocity of the source by formulating the state vector as $\mathbf{x}_k = [x_k,\ y_k,\ \dot{x}_k,\ \dot{y}_k]^\mathsf{T}$, where the first two elements x_k and y_k define the source position, and $\dot{x}_k$ and $\dot{y}_k$ denote the source velocity in x and y direction, respectively. Given the measurement vector $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ defined as the source location estimate from the maximum search in SRP beamformer spatial spectrum, the state-space equations in (2.28) and (2.29)

can be represented by

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k, \tag{2.45}$$

$$\mathbf{z}_k = \mathbf{C}\mathbf{x}_k + \mathbf{w}_k, \tag{2.46}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & a\Delta T & 0 \\ 0 & 1 & 0 & a\Delta T \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b\Delta T & 0 \\ 0 & b\Delta T \\ b & 0 \\ 0 & b \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{2.47}$$

In (2.45), the Langevin-process model assumes the process noise $\mathbf{u}_k$ as a 2×1 vector sampled from a Gaussian distribution, i.e., $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}_{2\times 1}, \mathbf{\Sigma})$ with zero mean and covariance $\mathbf{\Sigma} = \mathbf{I}_{2\times 2}$. The measurement noise $\mathbf{w}_k$ is a 2×1 vector sampled from either a Gaussian or non-Gaussian distribution. In (2.47), the variable ΔT is the time interval (in seconds) between consecutive frames, and a and b are the constants defined as $a = \exp(-\beta\Delta T)$, $b = \bar{v}\sqrt{1-a^2}$, where $\bar{v}$ is the steady-state velocity and β is the rate constant. In existing literatures, $\bar{v} = 0.8$ m/s, $\beta = 10$ Hz are often used [10, 28, 30, 33]. Compared to the random-walk process, the Langevin-process model is expected to achieve higher performance for a moving-source scenario with approximately constant velocity. This improved performance is due to the velocity components $\dot{x}_k$ and $\dot{y}_k$ formulated in the state transition. The performance of different source-dynamic models has been detailed in [34].

Langevin process with TDOA measurement: Given the state vector $\mathbf{x}_k = [x_k,\ y_k,\ \dot{x}_k,\ \dot{y}_k]^\mathsf{T}$ formulated in the Langevin process manner, and the measurement vector $\mathbf{z}_k = [\widehat{\tau}_k^{(1,2)},\ \ldots,\ \widehat{\tau}_k^{(i,j)},\ \ldots,\ \widehat{\tau}_k^{(N-1,N)}]^\mathsf{T}$ defined as the concatenation of estimated TDOAs, the state-space equations in (2.28) and (2.29) can be repre-

sented by

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k, \tag{2.48}$$

$$\mathbf{z}_k = [\mathcal{T}^{(1,2)}(\mathbf{C}\mathbf{x}_k),\ \ldots,\ \mathcal{T}^{(i,j)}(\mathbf{C}\mathbf{x}_k),\ \ldots,\ \mathcal{T}^{(N-1,N)}(\mathbf{C}\mathbf{x}_k)]^{\mathsf{T}} + \mathbf{w}_k, \tag{2.49}$$

where $\mathcal{T}^{(i,j)}(\cdot)$ is the non-linear TDOA function defined in (2.44) and matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ have been defined in (2.47). It worth noting that, similar to the case of (2.43), (2.49) again introduces non-linearity due to the TDOAs are used as measurement.

2.4.4 Motivation of using particle filter for acoustic source tracking

Essentially, the Bayesian filtering problem in (2.30) and (2.31) can be solved by either the Kalman filter or particle filter. For Kalman filter, the state-space functions $\mathcal{G}(\cdot)$ and $\mathcal{H}(\cdot)$ are required to be linear and that $\mathbf{u}_k$ and $\mathbf{w}_k$ are required to be Gaussian distributed. A closed-form solution can then be achieved where $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is derived as Gaussian distributed and the iterations of (2.30) and (2.31) are computed by the propagation and update steps of the mean and covariance of the Gaussian function. On the other hand, the particle filter essentially does not require any linearity and Gaussian assumptions [29,35]. Given the state-space formulations for AST discussed in Sec. 2.4.3, preference of Kalman filter or particle filter needs to be considered.

Table 2.3 shows a summary of advantages and disadvantages for using the Kalman filter and particle filter for each of state-space formulations discussed in Section 2.4.3. It can be observed that the Kalman filter requires the process noise $\mathbf{u}_k$ and measurement noise $\mathbf{w}_k$ to be Gaussian distributed. In addition, when the SRP beamformer is used to derive measurement $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^{\mathsf{T}}$, the computational complexity is high due to that estimation of source position using SRP beamformer involves

	Random-walk with SRP beamformer $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k$ $\mathbf{z}_k = \mathbf{x}_k + \mathbf{w}_k$	**Random-walk with TDOA** $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k$ $\mathbf{z}_k = \left[\mathcal{T}^{(1,2)}(\mathbf{x}_k), \ldots, \mathcal{T}^{(1,2)}(\mathbf{x}_k)\right]^{\mathrm{T}} + \mathbf{w}_k$
Kalman filter	Require $\mathbf{u}_k \sim \mathcal{N}(\cdot)$ Require $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ High complexity in obtaining $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$	Require $\mathbf{u}_k \sim \mathcal{N}(\cdot)$ Require $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ Additional linearization is needed for $\mathcal{T}(\cdot)$
Particle filter (Gaussian likelihood approach)	Assume $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ High complexity in obtaining $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$	Assume $\mathbf{w}_k \sim \mathcal{N}(\cdot)$
Particle filter (Pseudo-likelihood approach)	No Gaussianity assumption Low complexity as $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$ is not directly needed.	No Gaussianity assumption

	Langevin process with SRP beamformer $\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k$ $\mathbf{z}_k = \mathbf{C}\mathbf{x}_k + \mathbf{w}_k$	**Langevin process with TDOA** $\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k$ $\mathbf{z}_k = \left[\mathcal{T}^{(1,2)}(\mathbf{C}\mathbf{x}_k), \ldots, \mathcal{T}^{(1,2)}(\mathbf{C}\mathbf{x}_k)\right]^{\mathrm{T}} + \mathbf{w}_k$
Kalman filter	Require $\mathbf{u}_k \sim \mathcal{N}(\cdot)$ Require $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ High complexity in obtaining $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$	Require $\mathbf{u}_k \sim \mathcal{N}(\cdot)$ Require $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ Additional linearization is needed for $\mathcal{T}(\cdot)$
Particle filter (Gaussian likelihood approach)	Assume $\mathbf{w}_k \sim \mathcal{N}(\cdot)$ High complexity in obtaining $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$	Assume $\mathbf{w}_k \sim \mathcal{N}(\cdot)$
Particle filter (Pseudo-likelihood approach)	No Gaussianity assumption Low complexity as $\mathbf{z}_k = [\hat{x}_k, \hat{y}_k]^{\mathrm{T}}$ is not directly needed.	No Gaussianity assumption

Table 2.3: Comparison between Kalman filter and particle filter for different acoustic source tracking models.

computation of SRP functions over the entire surveillance area (see Sec. 2.3.2 for detailed explanation.) When the TDOA is used as measurement, non-linearity is introduced and additional linearization step is required for Kalman filter. On the other hand, although some of the formulations assume/satisfy linearity and Gaussianity, these two conditions essentially are not required for particle filter. When the SRP beamformer is directly used as the measurement $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^{\mathsf{T}}$, the computational complexity is high for the same reason. However, as will be discussed in the following section, the particle filter can avoid the need of obtaining $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^{\mathsf{T}}$ that involves high-complexity computation. This is achieved by using a pseudo-likelihood approach which approximates the measurement likelihood as the SRP function such that the number of evaluations of SRP function is only limited to the number of particles.

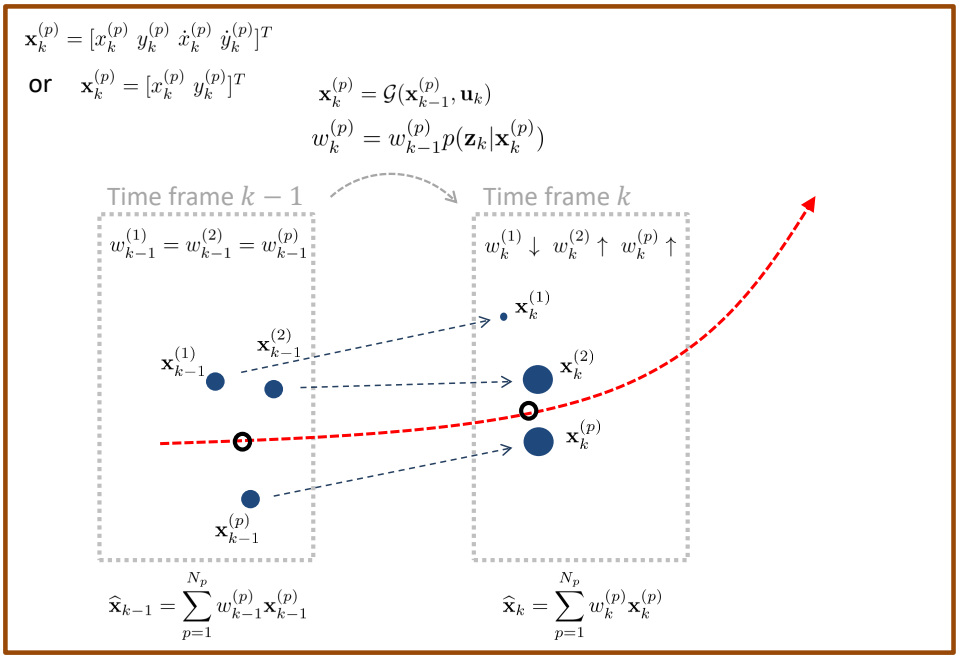


Figure 2.8: Illustration of one iteration of particle filter used for ASLT application. The dashed-curve arrow denotes the source trajectory. The solid circles denotes the particles and the size of the circle denotes the particle weight. The state estimate, as a weighted average of the particles, is denoted by the hollow circle.

Compared to Kalman filter, particle filter has advantages in low computational complexity when SRP beamformer is used as measurement, and absence of linearity requirement when TDOA is used as measurement. Therefore, although the Kalman filter has earlier been proposed (see e.g, [36]), the PF framework is deemed to be a better approach for the ASLT problem. The PF was first introduced to ASLT in [30] and has since gained great popularity as described in [10,11,28,30,31,33,37]. Throughout this thesis, only particle filter will be considered for acoustic source tracking.

2.4.5 Particle filter based acoustic source tracking

Among the PF family, the sequential importance resampling PF is most commonly used for ASLT [10, 11, 28, 30, 33]. Figure 2.8 shows a symbolic representation of SIRPF based ASLT for a single iteration. The actual source location trajectory is

indicated by the bold dash line. With reference to the SIRPF algorithm summarized in Table 2.2, the particles are first initialized in the neighborhood region of an estimated source location. At each time frame, the particles (denoted by dark circles) are then propagated according to the assumed source-motion model $\mathcal{G}(\cdot)$ (e.g., Langevin-process model or random-walk model) from time frame $k-1$ to k, as indicated by the thin dashed lines. The particles which are close to the actual source position/trajectory (indicated by the bold dash line) will be assigned a higher weight (indicated by the size of the dark circle) due to the high measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k^{(p)})$. Finally, the weighted average of the particles (shown by the lightly-shaded circle) is taken as the source position estimate.

In order to apply the particle filter tracking framework as discussed above, the formulation of measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ needs to be determined. Here, $p(\mathbf{z}_k|\mathbf{x}_k)$ needs to be formulated as a function of $\mathbf{x}_k$ for the update step in (2.34) or (2.39) for each particle $\mathbf{x}_k^{(p)}$. The likelihood function should reflect the fact that given a particle close to the true state, $\mathbf{x}_k^{(p)} \approx \mathbf{r}_k^{\text{src}}$ and a reliable measurement $\mathbf{z}_k$, $p(\mathbf{z}_k|\mathbf{x}_k^{(p)})$ achieves a high likelihood value for that particle. Since, for each SRP beamformer and TDOA based measurements, either a Gaussian or pseudo-likelihood can be used, four combinations can be derived in the following:

SRP beamformer with Gaussian likelihood: Given the measurement formulated as the position estimate from the full-grid maximum search in the SRP-beamformer spatial spectrum, $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^{\mathsf{T}}$, two hypotheses for the reliability of the measurement $\mathbf{z}_k$ at frame k are considered as

$$\mathcal{H}_k^0: \text{ SRP position estimate at frame } k \text{ being unreliable;}$$
$$\mathcal{H}_k^1: \text{ SRP position estimate at frame } k \text{ being reliable.}$$

Under the reliable hypothesis $\mathcal{H}_k^1$, the conditioned measurement likelihood can be

formulated as a Gaussian distribution with mean being the SRP location estimate given by

$$p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^1) = \mathcal{N}(\mathbf{r}_k; \mathbf{z}_k, \sigma_{\mathbf{z}}^2 \mathbf{I}_{2\times 2}), \tag{2.50}$$

where $\mathcal{N}(\cdot)$ denotes the Gaussian distribution function, $\mathbf{r}_k$ denotes the 2D position elements in the state $\mathbf{x}_k$, i.e.,

$$\mathbf{r}_k = \mathbf{x}_k, \quad \text{if} \quad \mathbf{x}_k = [x_k,\ y_k]^{\mathsf{T}}, \tag{2.51}$$

and

$$\mathbf{r}_k = \mathbf{C}\mathbf{x}_k, \quad \text{if} \quad \mathbf{x}_k = [x_k,\ y_k, \dot{x}_k,\ \dot{y}_k]^{\mathsf{T}}, \tag{2.52}$$

where $\mathbf{C}$ has been defined in (2.47), $\mathbf{z}_k$ denotes the position estimate from the full-grid maximum search in SRP beamformer spatial spectrum, $\sigma_{\mathbf{z}}^2 \mathbf{I}_{2\times 2}$ is the pre-defined covariance matrix. This formulation implies that the source position estimate from the SRP beamformer corresponds to the maximum likelihood estimate of the Gaussian distributed $p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^1)$. On the other hand, under the unreliable hypothesis $\mathcal{H}_k^0$, the conditioned measurement likelihood can be formulated as a uniform distribution over the enclosure domain as

$$p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^0) = \mathcal{U}_{\mathcal{D}}(\mathbf{r}_k), \tag{2.53}$$

where $\mathcal{U}(\cdot)$ denotes Gaussian distribution function, $\mathcal{D} = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max},\ y_{\min} \leq y_k \leq y_{\max}\}$ denotes the enclosure domain. Finally, the measurement likelihood can be expressed as

$$p(\mathbf{z}_k|\mathbf{x}_k) = p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^0)p(\mathcal{H}_k^0|\mathbf{x}_k) + p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^1)p(\mathcal{H}_k^1|\mathbf{x}_k), \tag{2.54}$$

where $p(\mathcal{H}_k^1|\mathbf{x}_k)$ denotes the prior probability for reliable hypothesis, and $p(\mathcal{H}_k^0|\mathbf{x}_k) =$

$1 - p(\mathcal{H}_k^1|\mathbf{x}_k)$. In practice, $p(\mathcal{H}_k^1|\mathbf{x}_k)$ can be derived as a monotonic function with SNR at a reference microphone or a binary decision depending on a voice activity decision [10].

The derivations above suffer from a major drawback - the SRP function in (2.17) has to be computed over the entire enclosure domain $\mathcal{D}$ in order to search for the maximum corresponding to the SRP beamformer measurement $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ as in (2.50); it requires high computational load for practical implementation.

SRP beamformer with pseudo-likelihood: The pseudo-likelihood approach circumvents the above drawback by assuming that the SRP function itself can be used as a pseudo-likelihood [10,28]. The SRP values corresponding to particle positions can, to some extent, represent the probability as it quantifies the energy originating from those positions. The pseudo-likelihood approach replaces (2.50) with

$$p(\mathbf{z}_k|\mathbf{x}_k, \mathcal{H}_k^1) = \{\mathcal{P}_k(\mathbf{r}_k)\}^{\gamma_\mathbf{z}}, \tag{2.55}$$

where $\mathcal{P}_k(\cdot)$ is the SRP function defined in (2.17), $\gamma_\mathbf{z}$ is a control parameter that regulates the SRP function for the approximation of the measurement likelihood [10] and $\gamma_\mathbf{z} = 2$ is often chosen [10,28]. As before, the variable $\mathbf{r}_k$ denotes the 2D position elements in the state $\mathbf{x}_k$. Finally, formulation of $p(\mathbf{z}_k|\mathbf{x}_k)$ is identical to the Gaussian approach given by (2.54).

It can be observed that (2.55) avoids the full-grid search in the spatial spectrum that is required in the SRP beamformer based localization algorithms, and this contributes to the popularity of PF for ASLT applications. The above technique confines the SRP evaluation to a small fixed number of particles (typically 100 to 200).

TDOA with Gaussian likelihood: Given the measurement formulated as TDOA estimates from M number of microphone pairs $\mathbf{z}_k =$

$[\widehat{\tau}_k^{(1,2)},\ \ldots,\ \widehat{\tau}_k^{(i,j)},\ \ldots,\ \widehat{\tau}_k^{(N-1,N)}]^{\mathsf{T}}$, Two hypotheses for the reliability of the measurement can be formulated as

$\mathcal{H}_k^0:\ \widehat{\tau}_k^{(i,j)}$ from the mth microphone pair at frame k being unreliable;

$\mathcal{H}_k^1:\ \widehat{\tau}_k^{(i,j)}$ from the mth microphone pair at frame k being reliable.

Under the reliable hypothesis $\mathcal{H}_k^1$, the conditioned measurement likelihood can be formulated as

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^1) = \mathcal{N}(\mathbf{r}_k; \widehat{\tau}_k^{(i,j)}, \sigma_\tau^2), \tag{2.56}$$

where $\mathcal{N}(\cdot)$ denotes Gaussian distribution function, $\mathbf{r}_k$ denotes the 2D position elements in the state $\mathbf{x}_k$, $\widehat{\tau}_k^{(i,j)}$ is the TDOA estimate obtained from the maximum of the GCC function as in (2.13). The above technique assumes that the maximum of the GCC function corresponds to the maximum of the Gaussian distributed $p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^1)$. On the other hand, under $\mathcal{H}_k^0$, the measurement likelihood can be formulated as a uniform distribution over the admissible TDOA range, i.e.,

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^0) = \frac{1}{2\tau_{\max}}, \tag{2.57}$$

where $\tau_{\max} = ||\mathbf{r}_j^{\text{mic}} - \mathbf{r}_i^{\text{mic}}||/c$. In addition, the prior probability $p(\mathcal{H}_k^1|\mathbf{x}_k)$ can be derived from the average SNR of each microphone pair, or the binary decision from a voice activity detection module [10], and $p(\mathcal{H}_k^0|\mathbf{x}_k) = 1 - p(\mathcal{H}_k^1|\mathbf{x}_k)$. With the above,

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k) = p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^0)p(\mathcal{H}_k^0|\mathbf{x}_k) + p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^1)p(\mathcal{H}_k^1|\mathbf{x}_k). \tag{2.58}$$

Finally, by noting that $\mathbf{z}_k = [\widehat{\tau}_k^{(1,2)},\ \ldots,\ \widehat{\tau}_k^{(i,j)},\ \ldots,\ \widehat{\tau}_k^{(N-1,N)}]^{\mathsf{T}}$ is a concatenation of

all the TDOA estimates with $(i,j) \in \Upsilon$, the measurement likelihood is given by

$$p(\mathbf{z}_k|\mathbf{x}_k) = \prod_{(i,j)\in\Upsilon} p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k). \tag{2.59}$$

Unlike Gaussian formulation for the SBF measurement, the Gaussian formulation for TDOA is widely used due to the relatively low complexity in the computation of $\widehat{\tau}_k^{(i,j)}$ [12,17,28,31]. In addition, by using the Gaussian approach, only the maximum of GCC function is taken into account for the formulation of measurement likelihood and the other spurious peaks will be omitted.

TDOA with pseudo-likelihood: The pseudo-likelihood approach assumes that the GCC function itself can also be used as a pseudo measurement likelihood since a high GCC value corresponds to a high probability of a given state being the actual source location. This approach replaces (2.56) by

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}_k^1) = \left\{ \Psi_k^{(i,j)}\big(\mathcal{T}(\mathbf{r}_k)\big) \right\}^{\gamma_\tau}, \tag{2.60}$$

where $\Psi_k^{(i,j)}(\cdot)$ is the GCC function defined in (2.12), γ_τ is the parameter for regulating the pseudo approximation. The non-linear TDOA function $\mathcal{T}(\cdot)$ is defined as

$$\mathcal{T}(\mathbf{r}_k) = \frac{1}{c}\Big(||\mathbf{r}_k - \mathbf{r}_j^{\text{mic}}|| - ||\mathbf{r}_k - \mathbf{r}_i^{\text{mic}}|| \Big), \tag{2.61}$$

where c is the speed of sound. Formulation of $p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k)$ and $p(\mathbf{z}_k|\mathbf{x}_k)$ is identical to what has been discussed in the Gaussian approach.

2.5 Existing algorithms and motivations of proposed algorithms

The SIRPF algorithm was first introduced in ASLT to track a single source in a reverberant and noisy environment in [28,30], and since then, it has drawn profound interest of the research community. A voice activity detector was subsequently integrated to mitigate degradation in performance caused by the non-stationarity of speech signals [10]. In SIRPF, the SNR is evaluated for each received signal frames and the prior probability $p(\mathcal{H}_k^1|\mathbf{x}_k)$ and $p(\mathcal{H}_k^0|\mathbf{x}_k)$ can be determined accordingly as a monotonic relationship of the SNR or a binary relationship via SNR-based thresholding. A similar approach can be found in [38] in which these two prior probabilities are determined by an additional estimation step of the source height. If the source height shows consistency with the previous estimates, it implies that the current measurement is reliable and hence $p(\mathcal{H}_k^1|\mathbf{x}_k)$ is high. In [33], both TDOA and SRP beamformer measurements were derived from an information theory perspective and fused into the SIRPF framework. In addition, the track-before-detect framework, which is capable of reducing the computational load, has been presented in [11]. In this algorithm, adjacent particles will share the same weight by performing the SBF measurement for only one of the particles, and this feature allows large number of particles to be involved in the algorithm. Although significant achievement has been shown in recent decades, tracking of a speech source in an adverse environment including background noise, reverberation and sound interference is still an open research topic. In this thesis, the contributions on solving these issues will be discussed in Chapter 3 and 4.

Besides single-source tracking, research has also been focused on addressing issues pertaining to multiple sources and alternating sources. For multi-source tracking, blind source separation techniques have been proposed to separate signals

corresponding to each source [39–41]. For the case of multiple sources where the number of sources is unknown and time-varying, the random finite set theory has been exploited to deal with varying dimension of the state vector [17, 41]. Furthermore, better particle sampling can be achieved by incorporating an existence grid that has been derived from a coarse search procedure via a beamformer [37]. For alternating-source tracking, where sources are active in turns and that only one source is active at any time instant, the extended Kalman filter [12, 31, 42] has been introduced to achieve fast convergence to the active source when an alternation occurs between any two speakers. In Chapter 5, the disadvantages of the extended Kalman filter will be discussed and a particle swarm intelligence based PF will be proposed for tracking of alternating speakers.

2.6 Chapter summary

Signal models for the ASLT problem are formulated. The TDOA and SRP beamformer based localization algorithms are then reviewed. For the tracking of a source position across a series of time frames, the particle filtering framework, which iteratively takes measurements from either the TDOA or SRP beamformer for sequential state estimation has been introduced. Existing state-of-the-art algorithms and challenges faced by the ASLT problem have been discussed. These challenges serve as a motivation for the proposed algorithms in the following chapters.

Chapter 3

Single-source Tracking in the Presence of Noise and Reverberation

In Chapter 2, the problem of acoustic source localization and tracking (ASLT) has been illustrated and reviewed. The conventional sequential importance resampling (SIRPF) tracking framework with the steered-response-power beamformer (SRP beamformer) measurement has gained popularity in ASLT application. However, it is well-known that tracking accuracy reduces in the presence of noise and reverberation. This is due to the fact that the spatial spectrum of the SRP beamformer is severely degraded by reverberation and noise resulting in an inaccurate approximation of the measurement likelihood. This chapter proposes to apply a regional SRP (RSRP) beamformer, which takes into account a circular region centered on each steered position, in order to achieve a more accurate approximation of the

Part of this chapter has been published as K. Wu and A. W. H. Khong, "Acoustic source tracking in reverberant environment using regional steered response power measurement," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, 2013.

measurement likelihood in a noisy and reverberant environment. The RSRP beamformer measurement is then incorporated to the SIRPF tracking framework. The performance of the proposed SIRPF-RSRP framework will be verified under various noisy and reverberant environments.

3.1 Introduction

As discussed in Section 2.4.5, the key step in SIRPF based acoustic source tracking is to determine the measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ in order for an appropriate weight to be assigned to each particle. The conventional SIRPF-SRP framework often uses the SRP function as an approximation of $p(\mathbf{z}_k|\mathbf{x}_k)$. In an actual room environment where background noise and reverberation exist, the aforementioned SIRPF-SRP framework suffers from performance degradation in tracking accuracy. This is due to the fact that noise and reverberation may result in spurious peaks in the spatial spectrum of the SRP beamformer and these spurious peaks compromise the approximation of $p(\mathbf{z}_k|\mathbf{x}_k)$ significantly.

Figure 3.1 (a) shows a typical spatial spectrum of the SRP beamformer in an environment with reverberation time $T_{60} = 350$ ms and signal-to-noise ratio (SNR) of 10 dB. It can be observed that this spatial spectrum comprises a large number of spurious peaks in addition to the peak corresponding to the source location. According to (2.55), the particles that have propagated to those spurious positions will therefore achieve a high measurement likelihood and large particle weights. As a result, the state estimate deviates from the actual source location. The motivation in this chapter is therefore to propose a measurement function that outperforms the SRP beamformer in terms of approximating $p(\mathbf{z}_k|\mathbf{x}_k)$. The proposed algorithm employs a regional smoothing process on the SRP beamformer spatial spectrum such that fewer spurious peaks will be attained, as indicated by Fig. 3.1 (b).

To achieve the above, instead of evaluating the power for each discrete steered position in the conventional SRP beamformer, a RSRP beamformer is applied that takes into account a circular region centered around each steered position [43]. The RSRP beamformer function is then used to derive the measurement likelihood in the SIRPF tracking framework via a proper non-linear mapping. This is achieved by considering distribution of the RSRP values. Simulation results show that the proposed SIRPF-RSRP algorithm is more robust than the SIRPF-SRP algorithm [10] and RSRP beamformer based localization algorithm [43] in a noisy and reverberant environment.

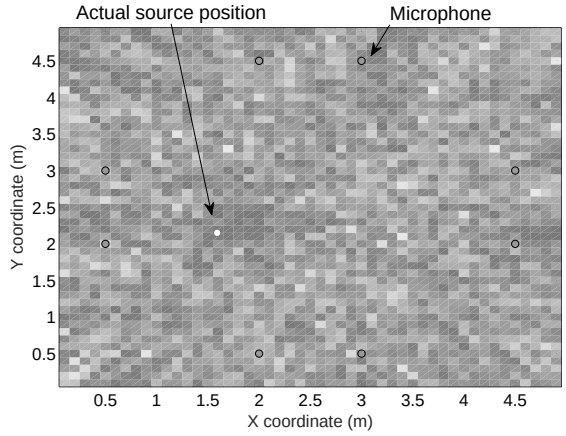
3.2 State-space formulation

Consider the state-space formulation in (2.45) and (2.46) in which the Langevin process and SRP beamformer measurement is considered as

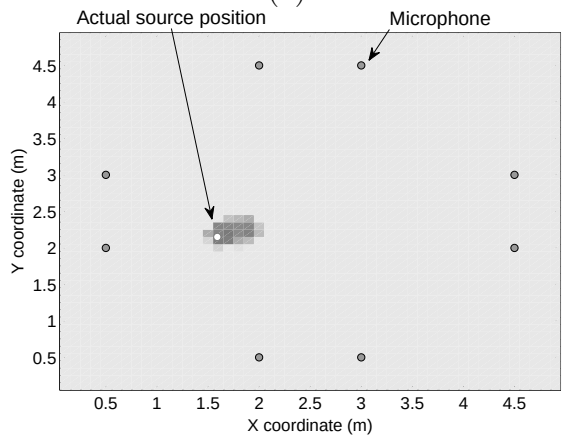
$$\begin{aligned}\mathbf{x}_k &= \mathcal{G}(\mathbf{x}_{k-1}, \mathbf{u}_k) \\ &= \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k, \qquad (3.1)\\ \mathbf{z}_k &= \mathcal{H}(\mathbf{x}_k, \mathbf{w}_k) \\ &= \mathbf{C}\mathbf{x}_k + \mathbf{w}_k, \qquad (3.2)\end{aligned}$$

where the state vector $\mathbf{x}_k = [x_k,\ y_k,\ \dot{x}_k,\ \dot{y}_k]^{\mathsf{T}}$ is defined as the concatenation of the source position and velocity in x and y directions, respectively, and the measurement vector $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^{\mathsf{T}}$ is defined as the source position estimate from the kth frame using the SRP beamformer. The variable $\mathbf{u}_k$ denotes the process noise while $\mathbf{w}_k$ denotes the measurement noise. Matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ have been defined in (2.47).

With reference to Table 2.3 and Sec. 2.4.4, due to the use of SRP beamformer measurement, applying Kalman filter will involve high computational complexity in



(a)



(b)

Figure 3.1: Spatial spectrums computed in a simulated environment with $T_{60} = 350$ ms and SNR = 10 dB. The higher value of SRP is indicated by dark color while the lower value is indicated by bright color. (a) Conventional SRP beamformer. (b) Proposed RSRP beamformer.

obtaining $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ where the SRP functions need to computed for the entire surveillance area. The use of particle filter can avoid the need of obtaining $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ that involves high-complexity computation by using a pseudo-likelihood approach as discussed in (2.55). For this reason, particle filter is preferable and will be considered.

3.3 Proposed SIRPF-RSRP algorithm

To mitigate the effect of reverberation and noise, A regional SRP beamformer [43] is proposed to be applied in the SIRPF framework. By integrating the power over a circular region centered on each steered position using this RSRP function, the spatial spectrum can be improved with fewer spurious peaks caused by noise and reverberation and hence more suitable for approximating $p(\mathbf{z}_k|\mathbf{x}_k)$.

3.3.1 RSRP beamformer measurement

The conventional SRP function has been defined in (2.17) as

$$\mathcal{P}_k(\mathbf{r}) = \int_{\Omega} \left| \sum_{i=1}^{N} w(\omega,k)\underline{y}_i(\omega,k) e^{\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_i^{\text{mic}}\|}{c}} \right|^2 d\omega. \tag{3.3}$$

In (3.3), the variable $\mathbf{r}$ is the steered position, i is the microphone index, N is the number of microphones, $\underline{y}_i(\omega,k)$ is the STFT coefficients of the ith microphone received signal, $w(\omega,k) = 1/|\underline{y}_i(\omega,k)|$ is the PHAT weighting function, $\mathbf{r}_i^{\text{mic}}$ is the ith microphone position, c is the speed of sound and Ω is the frequency range in which speech signal mainly exists.

Note that the conventional SRP beamformer in (3.3) is defined for a discrete steered position $\mathbf{r}$. Now consider a circular region $\mathbb{C}(\mathbf{r})$ centered at $\mathbf{r}$, as illustrated in Fig. 3.2, for the proposed RSRP beamformer formulation. This region is defined by

$$\mathbb{C}(\mathbf{r}) \triangleq \left\{ x, y \middle| \left|\left| [x,y]^{\mathsf{T}} - \mathbf{r} \right|\right|_2 \leq \rho \right\}, \tag{3.4}$$

where ρ denotes the radius of the circular region. The proposed RSRP function is therefore modified from the conventional SRP function in (3.3) by accumulating the

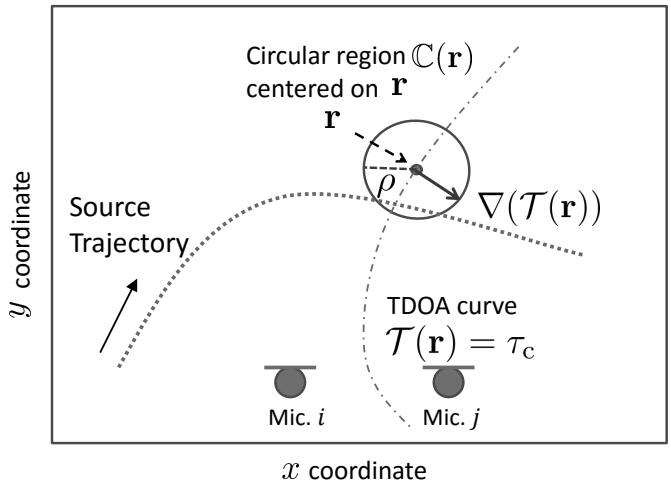


Figure 3.2: Illustration of computing regional steered response power for a circle region $\mathbb{C}(\mathbf{r})$ centered on a steered position $\mathbf{r}$.

power within $\mathbb{C}(\mathbf{r})$ as

$$\mathcal{P}_k^{\text{reg}}(\mathbf{r}) = \int_{\mathbf{r}' \in \mathbb{C}(\mathbf{r})} \mathcal{P}_k(\mathbf{r}') d\mathbf{r}', \tag{3.5}$$

where the superscript "reg" denotes for the "region". Such accumulation of power within consecutive regions across the whole room area can therefore be viewed as a smoothing process for the spatial spectrum.

With reference to (3.5), the integration is performed for continuous positions over a region $\mathbb{C}(\mathbf{r})$, which is generally untractable. To address this issue, recall that the conventional SRP function $\mathcal{P}_k(\mathbf{r})$ is related to the generalized cross-correlation (GCC) function as discussed in Section 2.3.3. This relationship is given by

$$\begin{aligned} \mathcal{P}_k(\mathbf{r}) &= \int_{\Omega} \left| \sum_{i=1}^{N} w(\omega, k) \underline{y}_i(\omega, k) e^{\jmath\omega \frac{\|\mathbf{r} - \mathbf{r}_i^{\text{mic}}\|}{c}} \right|^2 d\omega \\ &= \mathcal{P}_0 + 2 \sum_{(i,j) \in \Upsilon^{\text{full}}} \Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})), \end{aligned} \tag{3.6}$$

where $\Upsilon^{\text{full}} = \big\{(i,j) | i \leq N, j \leq N, i \leq j\big\}$ denotes the full microphone pair collec-

tion,

$$\Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})) = \int_{\Omega} \frac{\underline{y}_i(\omega,k)\underline{y}_j^*(\omega,k)}{|\underline{y}_i(\omega,k)\underline{y}_j(\omega,k)|} e^{\jmath\omega\mathcal{T}(\mathbf{r})} d\omega \tag{3.7}$$

is the GCC value for the steered position $\mathbf{r}$, and

$$\mathcal{T}(\mathbf{r}) = \frac{1}{c}\big(||\mathbf{r}-\mathbf{r}_j^{\text{mic}}|| - ||\mathbf{r}-\mathbf{r}_i^{\text{mic}}||\big) \tag{3.8}$$

is the time-difference-of-arrival (TDOA) between the ith and jth microphones for the steered position $\mathbf{r}$. From (3.6), one can observe that the power $\mathcal{P}_k(\mathbf{r})$ for a steered position $\mathbf{r}$ can be computed either using the conventional SRP definition in (3.3) or using the summation of GCC values for position $\mathbf{r}$ with a fixed energy term $\mathcal{P}_0$. It is also worth noting that the fixed energy term $\mathcal{P}_0$ is independent of the steered position $\mathbf{r}$, and hence in the following, this fixed energy term will be removed and the steered response power in (3.6) will be redefined as

$$\mathcal{P}_k(\mathbf{r}) = 2 \sum_{(i,j)\in\Upsilon^{\text{full}}} \Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r})). \tag{3.9}$$

Substituting (3.9) into (3.5), one can obtain

$$\begin{aligned}\mathcal{P}_k^{\text{reg}}(\mathbf{r}) &= 2 \sum_{(i,j)\in\Upsilon^{\text{full}}} \left\{ \int_{\mathbf{r}'\in\mathbb{C}(\mathbf{r})} \Psi_k^{(i,j)}(\mathcal{T}(\mathbf{r}'))d\mathbf{r}' \right\} \\ &= 2 \sum_{(i,j)\in\Upsilon^{\text{full}}} \left\{ \int_{\tau^{\text{l}}(\mathbf{r})}^{\tau^{\text{u}}(\mathbf{r})} \Psi_k^{(i,j)}(\tau)d\tau \right\},\end{aligned} \tag{3.10}$$

where in the second expression, the variable of interest for the integral has been changed from steered position $\mathbf{r}'$ to TDOA τ. It has been shown in [43] that the GCC function $\Psi_k^{(i,j)}(\cdot)$ for the steered position $\mathbf{r}'$ within a region $\mathbf{r}' \in \mathbb{C}(\mathbf{r})$ takes only TDOA values $\tau \in [\tau^{\text{l}}(\mathbf{r}), \tau^{\text{u}}(\mathbf{r})]$, where the lower bound $\tau^{\text{l}}(\mathbf{r})$ and upper bound $\tau^{\text{u}}(\mathbf{r})$ are only determined by the boundary of $\mathbb{C}(\mathbf{r})$. As illustrated in Fig. 3.2, in order to compute these two bounds given the region of interest $\mathbb{C}(\mathbf{r})$, the TDOA gradient

along which the TDOA exhibits the highest rate of change needs to be computed. By taking the derivative of (3.8), the TDOA gradient $\nabla(\mathcal{T}(\mathbf{r}))$ at position $\mathbf{r}$ can be derived as

$$\nabla(\mathcal{T}(\mathbf{r})) = [\nabla_x(\mathcal{T}(\mathbf{r})), \nabla_y(\mathcal{T}(\mathbf{r}))], \tag{3.11}$$

where $\nabla_x(\cdot) = \partial(\cdot)/\partial x$ such that

$$\nabla_x(\mathcal{T}(\mathbf{r})) = \frac{1}{c}\left(\frac{x - x_j^{\text{mic}}}{\|\mathbf{r} - \mathbf{r}_j^{\text{mic}}\|} - \frac{x - x_i^{\text{mic}}}{\|\mathbf{r} - \mathbf{r}_i^{\text{mic}}\|}\right), \tag{3.12}$$

$$\nabla_y(\mathcal{T}(\mathbf{r})) = \frac{1}{c}\left(\frac{y - y_j^{\text{mic}}}{\|\mathbf{r} - \mathbf{r}_j^{\text{mic}}\|} - \frac{y - y_i^{\text{mic}}}{\|\mathbf{r} - \mathbf{r}_i^{\text{mic}}\|}\right). \tag{3.13}$$

In (3.12) and (3.13), x and y denote the two-dimensional components of $\mathbf{r}$ while x_i^{mic} and y_i^{mic} denote the two-dimensional components of the ith microphone location. The lower and upper bounds of the TDOA can be computed by considering the product of the gradient magnitude and the distance along the gradient, i.e.,

$$\tau^{\text{l}}(\mathbf{r}) = \mathcal{T}(\mathbf{r}) - \|\nabla(\mathcal{T}(\mathbf{r}))\|\rho, \tag{3.14}$$

$$\tau^{\text{u}}(\mathbf{r}) = \mathcal{T}(\mathbf{r}) + \|\nabla(\mathcal{T}(\mathbf{r}))\|\rho, \tag{3.15}$$

where ρ has been defined as the radius of the circular region as in (3.4). The obtained TDOA lower and upper bounds can now be used for the computation of the regional SRP in (3.10). In practice, the GCC function is evaluated with finite number of discrete TDOA samples. Equation (3.10) can therefore be computed using

$$\mathcal{P}_k^{\text{reg}}(\mathbf{r}) = 2 \sum_{(i,j)\in\Upsilon^{\text{full}}} \sum_{\tau=\tau^{\text{l}}(\mathbf{r})}^{\tau^{\text{u}}(\mathbf{r})} \Psi_k^{(i,j)}(\tau). \tag{3.16}$$

Note that compared to (3.5), computation of the RSRP function in (3.16) is tractable since it transforms the continuous integration in (3.5) to a summation of discrete GCC values. Figure 3.1 (b) shows the RSRP beamformer spatial spectrum computed

using (3.16). It can be observed that, compared to Fig. 3.1 (a), Fig. 3.1 (b) exhibits fewer spurious peaks due to the regional smoothing process. It indicates that (3.16) can be used to formulate a better measurement likelihood than the one used in conventional SIRPF-SRP approach.

3.3.2 Approximation of measurement likelihood

Due to the removal of the fixed energy term $\mathcal{P}_0$ in (3.6), the computed RSRP in (3.16) may become negative. Therefore it cannot be used directly as a measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ as in (2.55). In this section, a monotonic mapping function $\mathcal{M}(\cdot)$ is proposed which maps the RSRP values into $p(\mathbf{z}_k|\mathbf{x}_k)$ given by

$$p(\mathbf{z}_k|\mathbf{x}_k) = \mathcal{M}(\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r})) \tag{3.17}$$

such that a higher value of $\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r})$ would indicate a higher probability in $p(\mathbf{z}_k|\mathbf{x}_k)$ and vise versa. In order to develop a suitable mapping function, the distribution of RSRP values can be first analyzed. Substituting (3.7) into (3.16), one can obtain

$$\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r}) = \sum_{(i,j)\in\Upsilon} \sum_{\mathcal{T}(\mathbf{r})=\tau^{\mathrm{l}}(\mathbf{r})}^{\tau^{\mathrm{h}}(\mathbf{r})} \int_{\Omega} e^{-\jmath\omega\mathcal{T}(\mathbf{r}^{\mathrm{src}})+\jmath\omega\mathcal{T}(\mathbf{r})} d\omega, \tag{3.18}$$

where $\mathbf{r}^{\mathrm{src}}$ is the actual source position. Equation (3.18) is useful for analyzing the distribution of RSRP values. The whole surveillance area $\mathcal{D}$ is split into two segments described by the following two paragraphs.

Distribution of RSRP values in the neighborhood area of source position: The neighborhood area of source position is defined as positions with distance from the true source position being less than a pre-defined threshold d_{t}, i.e., $\|\mathbf{r} - \mathbf{r}^{\mathrm{src}}\| \leq d_{\mathrm{t}}$. In this chapter, $d_{\mathrm{t}} = 0.2$ m was used. For positions in this area, $\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r})$ in (3.18) achieves the maximum due to the compensation of phase delays of the received

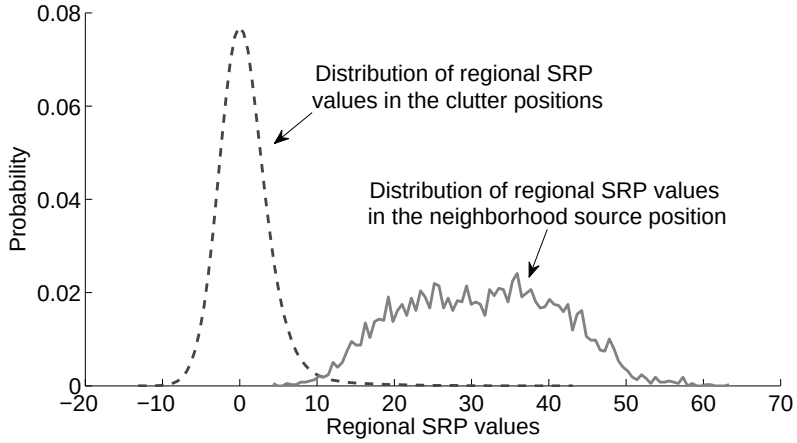


Figure 3.3: Distribution of the RSRP values. The distribution is computed for the clutter area and source neighbour area, respectively, in the entire surveillance area.

signals.

Distribution of RSRP values in the clutter area: The clutter area is defined as the positions which are at some distance away from the source position such that $\|\mathbf{r} - \mathbf{r}^{\text{src}}\| > d_{\text{t}}$. For these positions, due to mismatch in phase compensation, the phase is assumed to follow a uniform distribution [11], given by

$$\begin{aligned} O &= e^{-\jmath\omega\mathcal{T}(\mathbf{r}^{\text{src}})+\jmath\omega\mathcal{T}(\mathbf{r})} \\ &= e^{\jmath\theta}, \qquad \theta \sim \mathcal{U}_{[-\pi,\pi)}(\theta), \end{aligned} \tag{3.19}$$

where $\mathcal{U}_{[-\pi,\pi)}(\cdot)$ denotes uniform distribution over the angular range $[-\pi, \pi)$. In addition, due to the identically independent distributions of the phases and the sufficient number of summations for the phases, one can deduce, based on central-limit theorem, that the RSRP values for the clutter positions follow a Gaussian distribution, i.e.,

$$\mathcal{P}_k^{\text{reg}}(\mathbf{r}) \sim \mathcal{N}(0, \sigma_{\mathcal{P}}^2), \qquad \|\mathbf{r} - \mathbf{r}^{\text{src}}\| > d_{\text{t}}, \tag{3.20}$$

where $\sigma_{\mathcal{P}}^2$ is the variance of distribution of regional SRP values in the clutter positions.

Figure 3.3 shows the two distributions of the RSRP values in these two areas for a simulated room environment described in Sec. 3.4. The distribution of RSRP values in the neighborhood area of source position is indicated by the solid line, while the distribution of RSRP values in clutter area is indicated by the dashed line. The figure shows that the distribution of RSRP values in clutter area corresponds approximately to a zero-mean Gaussian distribution as expected. The variance $\sigma_{\mathcal{P}}^2$ depends on the TDOA summation boundary and number of microphone pairs used in (3.18). In the simulation, $\sigma_{\mathcal{P}}^2 = 25$ was observed when $M = 28$ and $\rho = 0.1$ m were used. On the other hand, the RSRP values corresponding to the neighborhood of source position are generally higher than the values corresponding to the clutter positions due to the phase compensation in (3.18). A threshold was therefore chosen to distinguish between these two distributions of RSRP values. In this work, an empirically determined threshold $\mathcal{P}_{\mathrm{t}} = 20$ was set in order to eliminate the effect of clutter positions as much as possible. This threshold should be modified accordingly if different M and ρ are used.

A normal cumulative distribution function (cdf) can be applied as the mapping given by

$$\mathcal{M}(\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r})) = \Phi(\mathcal{P}_k^{\mathrm{reg}}(\mathbf{r}), \mathcal{P}_{\mathrm{t}}, \sigma_{\Phi}^2), \tag{3.21}$$

where $\Phi(\cdot)$ is a normal cdf. As discussed, the threshold $\mathcal{P}_{\mathrm{t}} = 20$ is chosen so that the regional SRP values of clutter positions are mapped onto the lower end of $\Phi(\cdot)$, while those corresponding to the neighborhood of the source position are mapped onto the higher end of $\Phi(\cdot)$. The variable σ_{Φ}^2 is the variance of the normal cdf which determines its steepness. In this work, $\sigma_{\Phi}^2 = 12$ was chosen and performs well in the simulation. The likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ thus can be defined as

$$p(\mathbf{z}_k|\mathbf{x}_k) = \begin{cases} \mathcal{M}(\mathcal{P}_k^{\mathrm{reg}}(\mathbf{C}\mathbf{x}_k)), & \text{for voiced frame;} \\ \mathcal{U}_{\mathcal{D}}(\mathbf{C}\mathbf{x}_k), & \text{for unvoiced frame,} \end{cases} \tag{3.22}$$

Table 3.1: Summary of the sequential-importance-resampling PF.

At time $k-1$, A set of particles $\{\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of posterior $p(\mathbf{x}_{k-1}|\mathbf{z}_{k-1})$.

For the kth frame:

1. *Particles propagation*: Propagate each particle through the source dynamic model (3.1),
$$\mathbf{x}_k^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)}, \mathbf{u}_k).$$

2. *Update*: Each particle is then assigned a weight according to its likelihood
$$w_k^{(p)} \propto w_{k-1}^{(p)} p(\mathbf{z}_k|\mathbf{x}_k^{(p)}),$$
followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)}(\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.

3. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$.

4. *Result*: the particle set $\{\mathbf{x}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$ is obtained for approximation of $p(\mathbf{x}_k|\mathbf{z}_k)$. The state estimate at the kth frame is $\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \mathbf{x}_k^{(p)}$.

---

where $\mathbf{C}$ which takes the first two elements of the state vector $\mathbf{x}_k$ has been defined in (2.47), $\gamma_{\mathbf{z}} = 2$ is a control parameter to regulate the fusion of the SRP function to the likelihood [10,28], and $\mathcal{U}_{\mathcal{D}}(\cdot)$ is the uniform pdf over the considered enclosure domain $\mathcal{D} = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max}, y_{\min} \leq y_k \leq y_{\max}\}$. Note that in (3.22), the binary decision of voice activity is incorporated into the formulation of measurement likelihood, which is usually determined by evaluating whether the SNR of the kth frame signal is greater than a predefined threshold. The remaining procedures follow the SIRPF framework described in Table 3.1. The position estimate at each iteration $\widehat{\mathbf{r}}_k^{\text{src}}$ correspond to the first two elements of the state estimate $\widehat{\mathbf{x}}_k$, i.e., $\widehat{\mathbf{r}}_k^{\text{src}} = \mathbf{C}\widehat{\mathbf{x}}_k$.

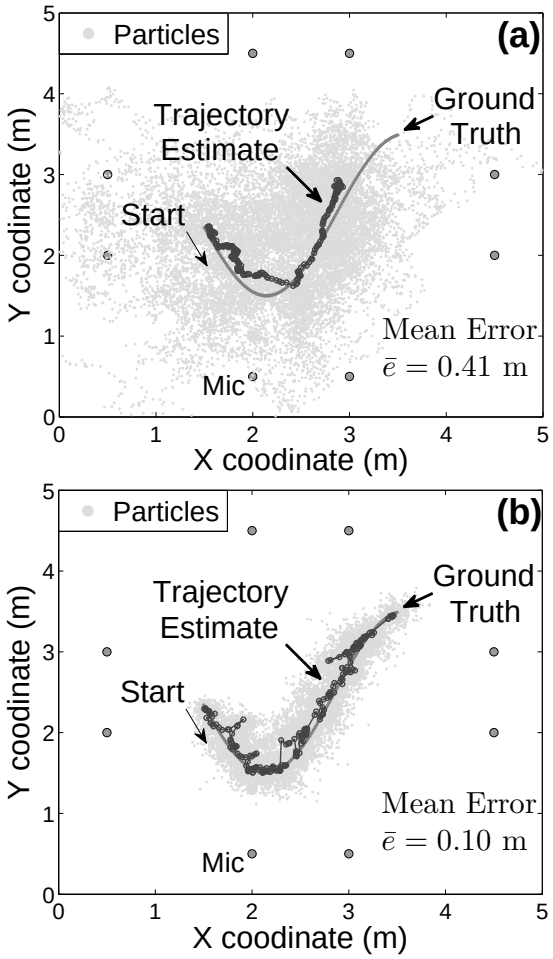


Figure 3.4: Comparison of tracking results with $T_{60} = 450$ ms and SNR = 10 dB. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-RSRP tracking algorithm.

3.4 Simulation results

Simulations were conducted in a room of dimension 5 m $\times$ 5 m $\times$ 2.5 m. Eight microphones were distributed 0.5 m away from the perimeter of the room (see Fig. 3.4.) A 13 s speech signal sampled at 16 kHz from the TIMIT database [44] was used as a source signal. The microphone signals were generated by the method of images [45]. White Gaussian noise (WGN) at different SNR was added to the microphone signals. The positions of the speech source were estimated using a frame size of 1024 samples with $N_p = 80$ particles. The radius of the circular region centered on each particle

was $\rho = 0.1$ m. The effective sample size threshold in SIRPF was $N_{\text{thr}} = 37.5$.

The proposed SIRPF-RSRP algorithm is compared with the conventional SIRPF-SRP tracking algorithm [10] where the simple binary voice/unvoice detector was implemented and the RSRP localization method without SIRPF framework [43]. The performance is quantified using $e_k = ||\widehat{\mathbf{r}}_k^{\text{src}} - \mathbf{r}_k^{\text{src}}||_2$, where $\widehat{\mathbf{r}}_k^{\text{src}}$ is the estimated position at the kth frame, and $\mathbf{r}_k^{\text{src}}$ is the true source position. The average tracking error $\bar{e} = \frac{1}{K}\sum_{k=1}^{K} e_k$ quantifies the performance across all audio frames, where K is the number of frames.

Figure 3.4 compares the tracking results of the two SIRPF based tracking algorithms when $T_{60} = 450$ ms. Figure 3.4 (a) shows that the performance of the conventional SIRPF-SRP algorithm [10] is significantly affected by room reverberation. The particles, indicated by the dotted points, are scattered around the surveilled region due to the poor performance of the conventional SRP measurements. The conventional SIRPF-SRP algorithm has an average tracking error of 0.41 m. Figure 3.4 (b) shows the performance of the proposed SIRPF-RSRP algorithm. The RSRP beamformer measurement results in well-propagated particles concentrated along the true source trajectory. The proposed SIRPF-RSRP algorithm achieves an averaged tracking error of 0.10 m, indicating that it outperforms the conventional SIRPF-SRP algorithm in this reverberant environment.

Figure 3.5 shows the average tracking error of the conventional SIRPF-SRP algorithm [10], the RSRP beamformer localization algorithm without SIRPF [43] and the proposed SIRPF-RSRP algorithm, across different reverberation times. Two cases of SNR = 10 and 3 dB were examined. The performance of these three algorithms reduces with reverberation time, as expected. The conventional SIRPF-SRP algorithm and the RSRP beamformer consistently exhibit higher tracking error than the proposed SIRPF-RSRP algorithm. The lower SNR condition further degrades the performance of conventional algorithms. Due to the improved robustness

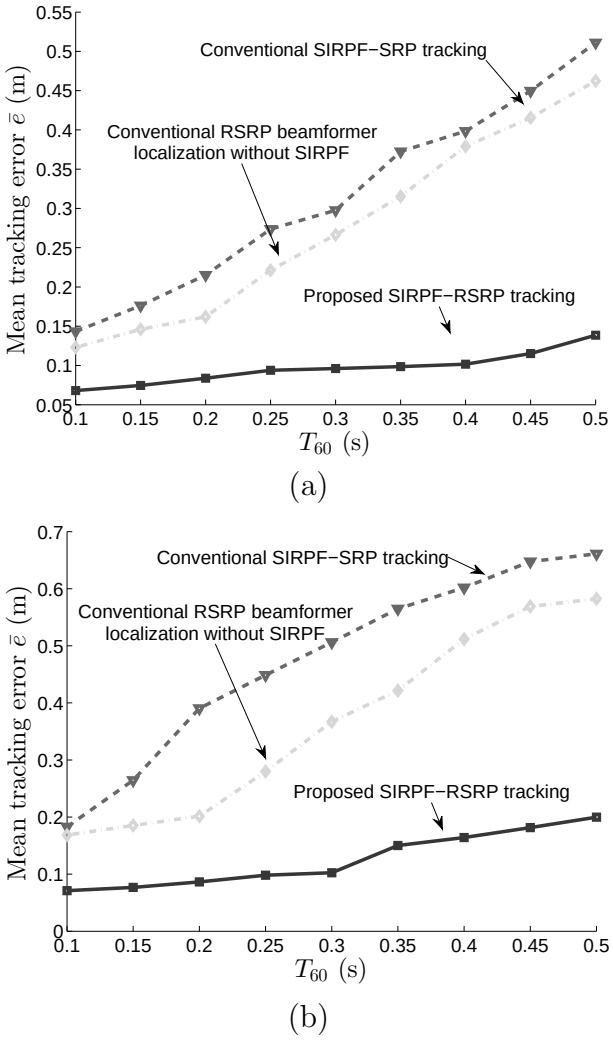


Figure 3.5: Variation of average tracking error with reverberation time for (a) SNR = 10 dB and (b) SNR = 3 dB.

against noise and reverberation in spatial spectrum, the RSRP-beamformer localization algorithm performs modestly better than the SIRPF-SRP algorithm, even though it does not exploit the temporal consistency of source positions. By incorporating the SIRPF framework and taking into account the temporal consistency of source positions, the proposed SIRPF-RSRP algorithm results in a mean error of less than 0.2 m, indicating that it outperforms both of the two conventional algorithms for the environments being examined. The improvement over the conventional algorithms becomes more significant at lower SNR and higher reverberant

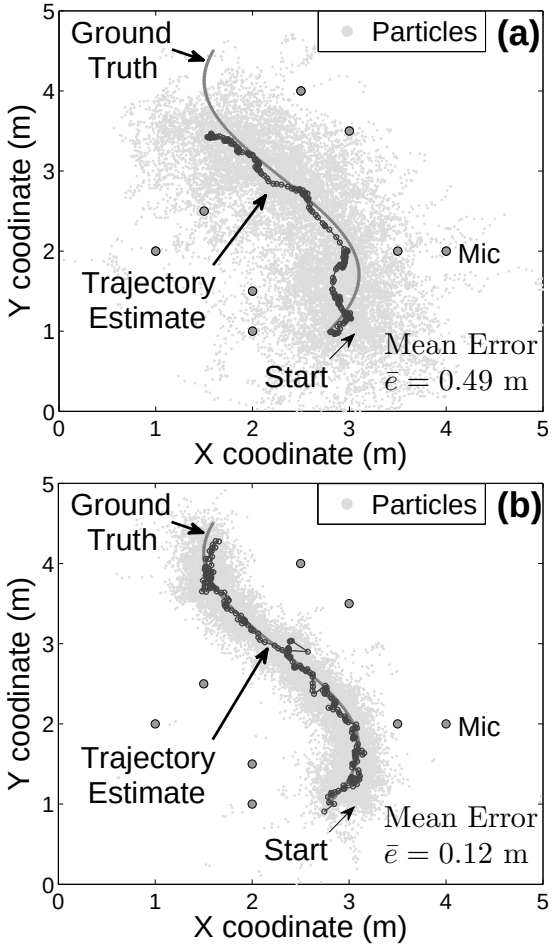


Figure 3.6: Comparison of tracking results with $T_{60} = 450$ ms and SNR = 10 dB using randomly distributed microphones. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-RSRP tracking algorithm.

condition.

To further examine the performance of the algorithm for different microphone array configurations, a scenario where microphones are randomly distributed as illustrated in Fig. 3.6 is also evaluated. The remaining parameters were the same as the previous simulations. The conventional SIRPF-SRP algorithm [10], shown in Fig. 3.6 (a), results in the particles scattered around the room enclosure and a mean tracking error of 0.49 m is exhibited. The proposed SIRPF-RSRP algorithm, shown in Fig. 3.6 (b), can, however, achieve lower mean tracking error of 0.12 m. This simulation indicates that the algorithm is not limited to the case where the

microphones have to be placed along the parameter of the room enclosure.

3.5 Chapter summary

This chapter proposes a SIRPF based acoustic source tracking framework by using a RSRP-beamformer for the approximation of the measurement likelihood. Instead of evaluating the power of discrete particle positions, the proposed SIRPF-RSRP algorithm takes into account a circular region centered on each particle by accumulating the power within each region to provide a more comprehensive likelihood evaluation. Simulation results show that the proposed SIRPF-RSRP algorithm achieves lower tracking error than the conventional algorithms in a noisy and reverberant environment.

Chapter 4

Single-source Tracking in the Presence of Sound Interference by Exploiting Speech Harmonicity

In Chapter 3, an algorithm for the tracking of a single source in the presence of reverberation and background noise has been proposed. In practice, the presence of sound interference (e.g. fan noise, air conditioner noise) in a typical room environment may also degrade the performance of the conventional sequential importance resampling PF (SIRPF) tracking framework with steered-response-power (SRP) beamformer measurement. In this chapter, a speech source tracking algorithm that is robust to interferers will be introduced. The proposed algorithm incorporates the harmonicity of speech signals. By exploiting this feature, the harmonicity based SRP (HSRP) beamformer can be derived and incorporated into the SIRPF tracking

Part of this chapter has been published as K. Wu and A. W. H. Khong, "Sound source localization and tracking for social robots," *Context Aware Human-Robot and Human-Agent Interaction*, pp 55-78, Springer, 2016, ©Springer International Publishing Switzerland 2016, and K. Wu, S. T. Goh and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Process. (ICASSP)*, 2013, ©[2013] IEEE.

framework. Performance of the proposed SIRPF-HSRP algorithm will be verified through comparison with the conventional SIRPF-SRP algorithm in the presence of various sound interferers.

4.1 Introduction

In general applications, a speech source is often considered as the desired source to achieve camera steering towards the active speaker, or to achieve signal enhancement given the estimated source location. However, in a typical room environment, an interference (e.g. fan noise, air conditioner noise, computer noise, or telephone ring noise) may often be present which degrades the performance of the tracking system. As a result, the tracking system localizes and tracks the interferers rather than the desired speech source. This is due to the fact that the existing tracking framework employs the SRP beamformer function as the measurement likelihood for particle weight update [10, 11, 28]. However, since the SRP beamformer is non-discriminative, the presence of interference will result in consistently high likelihood at the location(s) of the interferer(s) which, in turn, causes particles to converge at the wrong location away from the speech source.

Figure 4.1 (a) shows a typical spatial spectrum of the conventional SRP beamformer for an environment in which a speech source and a telephone ring interference co-exist with an signal-to-interference ratio (SIR) of 0 dB. It can be observed that the spatial spectrum of the conventional SRP beamformer has a significant peak in the interference position while the SRP for the source position is compromised. It can be foreseen that the measurement likelihood derived by this SRP beamformer function as in (2.55) cannot lead to an accurate approximation of the presence probability for the speech source.

In this chapter, a speech harmonicity based tracking algorithm is proposed

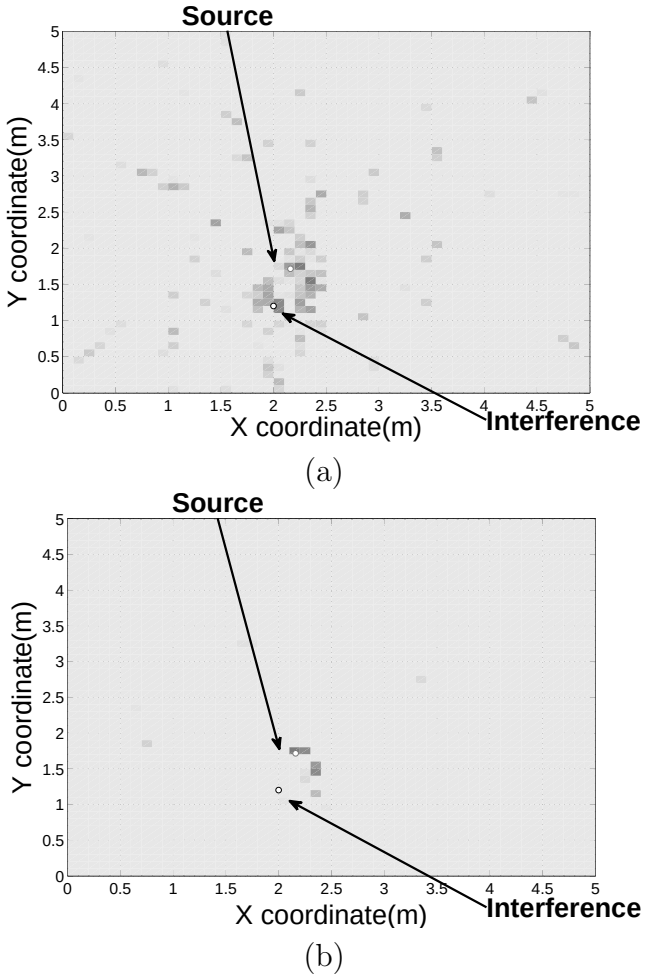


Figure 4.1: Spatial spectrums computed in a simulated environment with T_{60} = 250 ms and SNR = 15 dB in which a speech source and telephone ring source exist with SIR =0 dB. The higher value of SRP is indicated by dark color while the lower value is indicated by bright color. (a) Conventional SRP beamformer. (b) Proposed HSRP beamformer.

to deal with the effect of interference. Although the authors of [46, 47] proposed localization methods by jointly estimating the pitch frequency and source position, their primary aim is not to reject interference signals. In addition, conventional tracking algorithms do not consider any source feature [10, 11, 28]. The proposed algorithm, on the other hand, incorporates distinctive speech harmonics and a SIRPF framework to achieve robust speech source tracking in the presence of interference. First, a signal enhancement algorithm is employed to enhance the signal from a prior

estimated source location. The enhanced signal is then used to extract the speech harmonic information. A HSRP beamformer function is then derived (the spatial spectrum is shown in Fig. 4.1 (b) in which the dominant peak correspond to the speech source) by considering the fact that speech energy is concentrated on the harmonic bands, while the interference energy may be distributed over different frequency regions. Finally, a new particle weight update scheme is derived based on the HSRP beamformer function to achieve speech-sensitive source tracking. Simulations are conducted to compare the tracking performance between the proposed SIRPF-HSRP and the conventional SIRPF-SRP algorithms in the presence of interference, noise and reverberation.

The organization of this chapter is as follows: the state-space problem formulation is stated in Section 4.2. The speech spectrogram is then compared with some typical sound interference in Section 4.3 and the speech harmonic feature will be illustrated. Details of the proposed SIRPF-HSRP algorithm will be introduced in Section 4.4. In Section 4.5, simulations are conducted to evaluate the performance of the proposed algorithm in the presence of interference, noise and reverberation.

4.2 State-space formulation

Consider the single-source state-space formulation in (2.45) and (2.46) in which the Langevin process and SRP beamformer measurement are given by

$$\begin{aligned}\mathbf{x}_k &= \mathcal{G}(\mathbf{x}_{k-1}, \mathbf{u}_k) \\ &= \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k, \end{aligned} \tag{4.1}$$

$$\begin{aligned}\mathbf{z}_k &= \mathcal{H}(\mathbf{x}_k, \mathbf{w}_k) \\ &= \mathbf{C}\mathbf{x}_k + \mathbf{w}_k, \end{aligned} \tag{4.2}$$

where the state vector $\mathbf{x}_k = [x_k,\ y_k,\ \dot{x}_k,\ \dot{y}_k]^\mathsf{T}$ is defined as the concatenation of the source position and velocity in x and y directions, respectively and the measurement vector $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ is defined as the source position estimate from the kth frame using the SRP beamformer. The variable $\mathbf{u}_k$ denotes the process noise while $\mathbf{w}_k$ denotes the measurement noise. The matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ have been defined in (2.47).

With reference to Table 2.3 and Sec. 2.4.4, to avoid the need of obtaining $\mathbf{z}_k = [\widehat{x}_k,\ \widehat{y}_k]^\mathsf{T}$ that involves high-complexity computation, particle filter will be considered as the tracking framework.

4.3 Speech harmonic structure

Figure 4.2 shows the spectrogram of a typical speech signal obtained from the TIMIT database [44] and that corresponding to different sound interferers obtained from the NOISEX-92 database [48]. The speech spectrogram, as shown in Fig. 4.2 (a), indicates that several harmonics (dark curves) corresponding to multiple integers of a pith frequency are present. The pitch frequency represents the frequency of the vocal cord vibration which normally ranges from 100 to 300 Hz depending on whether it is male or female voice [49]. This spectrogram indicates that speech energy is dominant on these harmonics. Figure 4.2 (b) shows the spectrogram of a recorded fan noise where the energy is concentrated below 2 kHz. The spectrogram of a recorded power drill noise, shown in Fig. 4.2 (c), indicates similar energy distribution in the low frequency range although high energy spectral lines appear at approximately 1.5, 2 and 2.2 kHz. These dominant frequencies may be caused by the mechanical rotation or vibration. It is useful to note that no regular harmonic structure is exhibited in these two types of sound. In terms of the telephone ring sound, shown in Fig. 4.2 (d), a regular harmonic structure is caused by the presence of a single

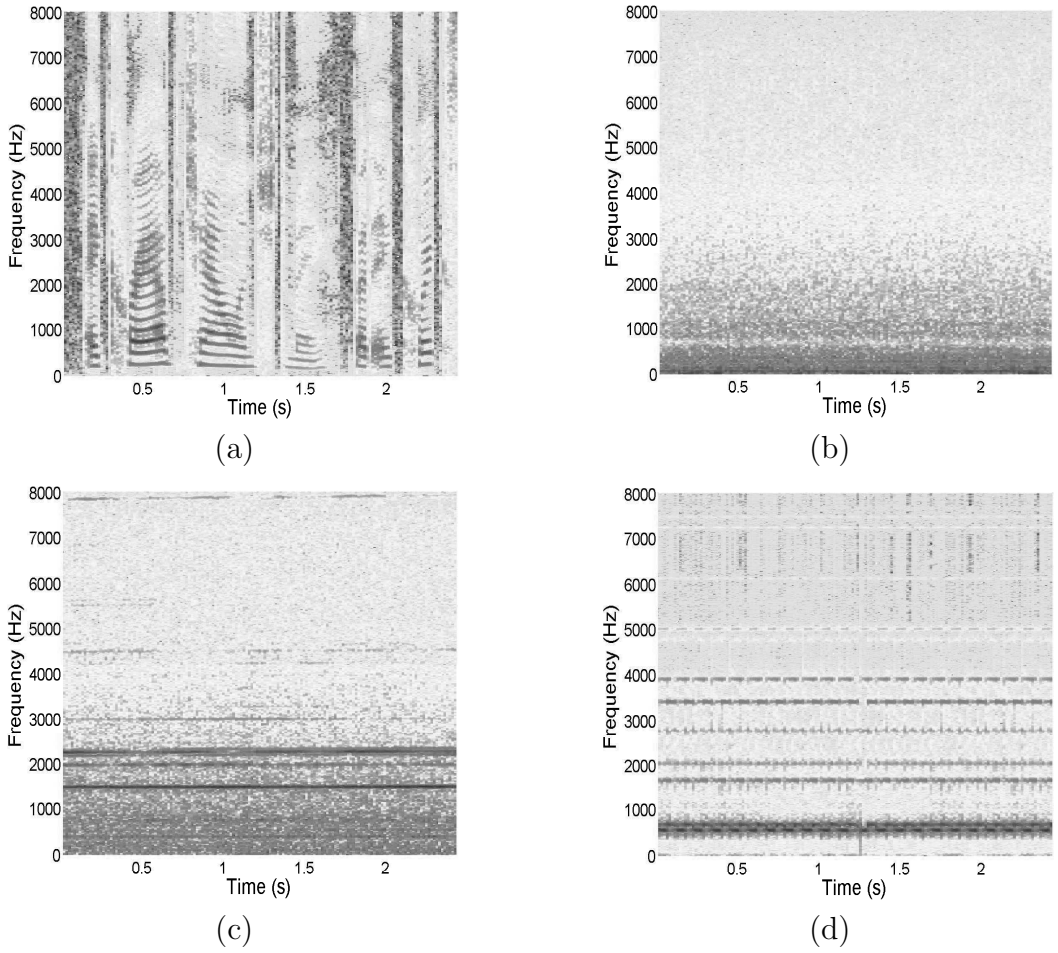


Figure 4.2: Spectrograms of different signals. (a) Speech signal spectrogram. (b) Fan noise spectrogram. (c) Power drill noise spectrogram. (d) Telephone ring noise spectrogram.

tone. However, the harmonics differ from that of the speech signal due to a difference in pitch frequency.

In the following, an assumption can be made that the sound interference does *not* share the same harmonic bands as speech due to different pitch frequency, or that the interference does not possess any harmonic structure. The key objective of the proposed method is to estimate these harmonic bands corresponding to the speech components and impose an emphasis on these harmonic bands since they provide high signal-to-interference ratio. Other frequency regions will not be utilized for

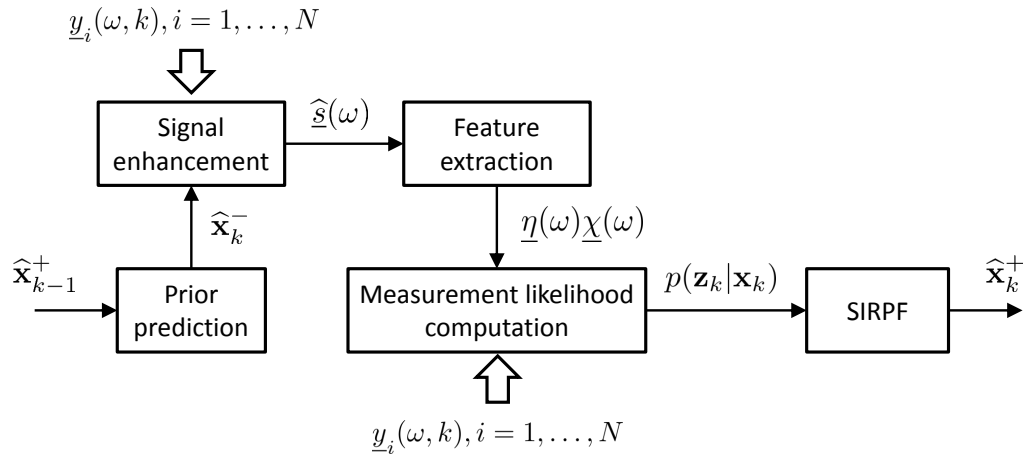


Figure 4.3: System diagram of the proposed SIRPF-HSRP algorithm.

tracking since these frequencies are contaminated by the sound interferers.

4.4 Proposed SIRPF-HSRP algorithm

To mitigate any degradation in performance due to interference, the proposed algorithm exploits speech harmonicity for deriving the HSRP beamformer function as shown in Fig. 4.1 (b) such that the corresponding measurement likelihood is predominantly weighted by the speech signal as opposed to the interference. The overall framework of the proposed algorithm is shown in Fig. 4.3 and the processing steps are as follow: (1) a prior source position is estimated using the assumed source dynamic model; (2) a speech enhancement module is then applied to enhance the source signal from that prior estimated position in order to extract speech feature; (3) the reliable harmonic bands are estimated using the enhanced signal in the following step; (4) a new HSRP beamformer measurement is then derived for the approximation of the measurement likelihood which emphasizes the high SIR harmonic bands while discarding other frequency regions; (5) the feature-oriented measurement likelihood is then used in SIRPF tracking framework.

4.4.1 Prior prediction

In general, a clean source signal is often required in order to extract the corresponding speech features. However, due to the presence of interference and background noise, obtaining such a clean source signal is challenging. To improve the performance of feature extraction, the algorithm incorporates a speech signal enhancement stage consisting of prior source position prediction and a beamformer. Considering the Langevin source dynamic model introduced in Section 4.2, for time frame index k. The prior source state can be estimated, using (4.1) as

$$\widehat{\mathbf{x}}_k^- = \mathcal{G}(\widehat{\mathbf{x}}_{k-1}^+, \mathbf{u}_k), \tag{4.3}$$

given the state estimate at previous frame. Here, $\widehat{\mathbf{x}}_{k-1}^+$ is the posterior state estimate at time frame index $k-1$. The prior source location estimate

$$\begin{aligned} \widehat{\mathbf{r}}_k^{\text{src}-} &= [\widehat{x}_k^- \quad \widehat{y}_k^-]^\mathsf{T} \\ &= \mathbf{C}\widehat{\mathbf{x}}_{k-1}^+, \end{aligned} \tag{4.4}$$

corresponds to the first two elements in $\widehat{\mathbf{x}}_{k-1}^+$ and $\mathbf{C}$ has been defined in (2.47). It is worth noting that this prior estimate is based only on the assumed source motion. Its objective is to facilitate the speech enhancement process as will be described in Section 4.4.2 to enhance the signal from this preliminary estimated source position. The feature-directed measurement, as will be described in subsequent subsections, will further refine the state estimate.

4.4.2 Feature extraction

After obtaining a prior estimate of the source position at each iteration, a signal enhancement algorithm can be employed to enhance the signal from that particular

position. It is worth noting that the beamformer was used as a localization technique in Section 2.3.2. However, beamforming was initially used for enhancing the signal from a known source position and suppressing the background noise and interference from other directions [26]. Therefore, various beamfomers can be applied to enhance the speech signal after a prior source location has been estimated. In this work, the delay-and-sum beamformer [50] is considered due to its simplicity although other forms of beamformer such as presented in [51, 52] may be used to enhance the speech signal. The enhanced speech signal, in the frequency domain at current frame k using the delay-and-sum beamformer steered to $\widehat{\mathbf{r}}_k^{\mathrm{src}-}$, is given by

$$\underline{\widehat{s}}(\omega) = \sum_{i=1}^{N} \mathcal{W}(\widehat{\mathbf{r}}_k^{\mathrm{src}-}) \underline{y}_i(\omega, k) e^{j\omega \|\widehat{\mathbf{r}}_k^{\mathrm{src}-} - \mathbf{r}_i^{\mathrm{mic}}\|_2 / c}, \tag{4.5}$$

where the time frame index k in $\underline{\widehat{s}}(\omega)$ has been omitted temporarily since it is implicit that $\underline{\widehat{s}}(\omega)$ is computed for the current frame and will be used only across frequency domain later. The variable i denotes the microphone index, N is the number of microphones, $\underline{y}_i(\omega, k)$ is the frequency-domain received signal from the ith microphone at kth frame. The variable ω is the angular frequency, c is the speed of sound, $\|\widehat{\mathbf{r}}_k^{\mathrm{src}-} - \mathbf{r}_i^{\mathrm{m}}\|_2$ is the distance from the prior estimated source position $\widehat{\mathbf{r}}_k^{\mathrm{src}-}$ to the ith microphone position $\mathbf{r}_i^{\mathrm{mic}}$, $\|\cdot\|$ is the Euclidean distance and $\mathcal{W}(\widehat{\mathbf{r}}_k^{\mathrm{src}-})$ is a monotonic function that weighs the ith microphone signal according to the source-sensor distance. In the simulations, the function $\mathcal{W}(\widehat{\mathbf{r}}_k^{\mathrm{src}-}) = \frac{1}{\|\widehat{\mathbf{r}}_k^{\mathrm{src}-} - \mathbf{r}_i^{\mathrm{m}}\|_2}$ has been found to perform well since it emphasizes the signal from the microphone that is closer to the source.

Figure 4.4 shows the signal enhancement result for a 6 s speech signal when a power drill interference is present at SIR = 5 dB and white Gaussian noise with signal-to-noise (SNR) ratio of 15 dB. These results were generated using the method of images [45] with T_{60} = 200 ms and eight microphones are placed 0.5 m away

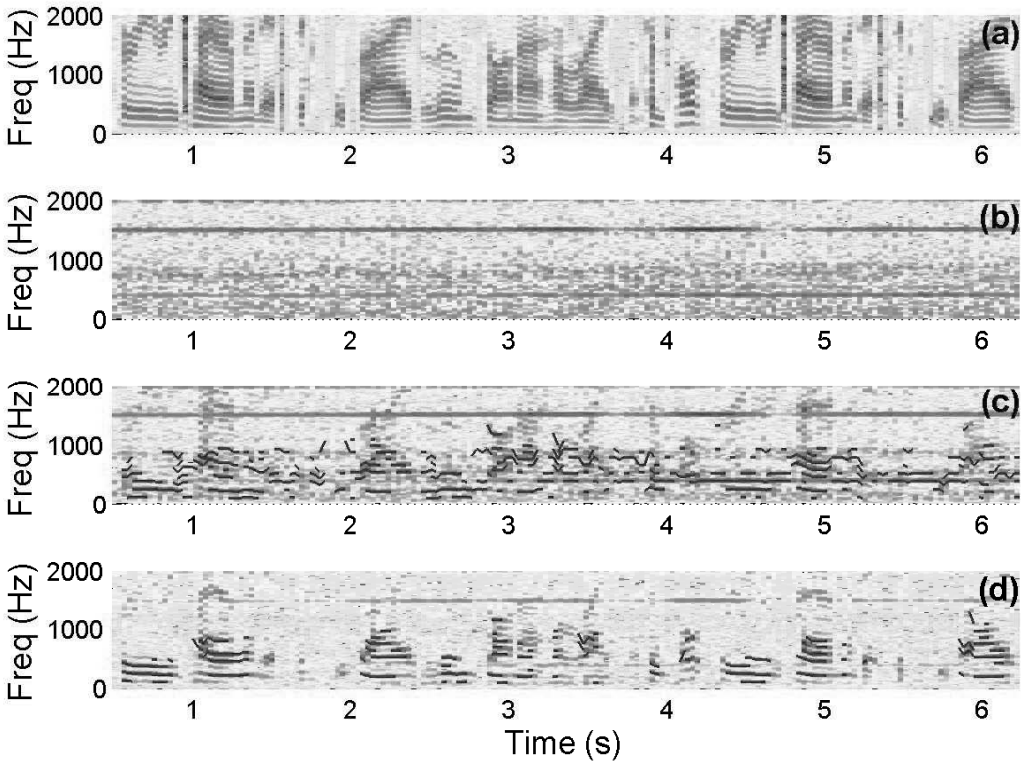


Figure 4.4: Spectrograms and selected harmonic bands indicated in dark lines. (a) Clean speech. (b) Power-drill interference. (c) Reference microphone received signal and its selected harmonic bands (in dark.) (d) Beamformer enhanced signal and its selected harmonic bands (in dark.)

from the room perimeter (see Fig. 4.7.) Figure 4.4 (a) shows the spectrogram of the original (clean) speech signal where a clear harmonic structure can be found. Figure 4.4 (b) shows the power drill interference spectrogram where no harmonic structure is present. In general, the source signal received by a single reference microphone is often distorted, especially when the interferer is close to the microphone, as shown in Figure 4.4 (c). Extraction of speech harmonics from this received signal is therefore challenging. The beamformer enhanced signal, as shown in Figure 4.4 (d), is indeed clearer than the single microphone received signal. The speech harmonics are dominant across the entire spectrogram although certain interference energy leakage is visible. The beamformer enhanced signal will be used for feature extraction in the next step.

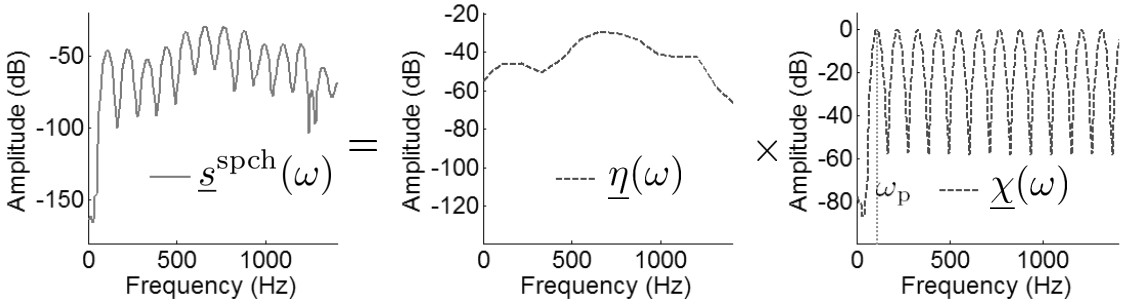


Figure 4.5: MBE model for a speech signal. The voice frame in frequency domain $\underline{s}^{\mathrm{spch}}(\omega)$ can be modeled as an a product of spectrum envelop $\underline{\eta}(\omega)$ and excitation spectrum $\underline{\chi}(\omega)$.

To extract the speech harmonics from a noisy spectrum, although many pitch estimation algorithms exist such as [53,54], use of them will only result in estimation of a pitch frequency. In this work, the multi-band excitation (MBE) fit method [55, 56] is considered because it is a one-step frequency-domain fitting approach which not only estimates the pitch frequency but also infers which harmonic frequency bands are reliable. The computed fitting error can be directly used for selecting reliable harmonic bands later as in (4.12). As indicated in Fig. 4.5, the MBE model defines a voiced frame in the frequency domain as the product of spectrum envelop $\underline{\eta}(\omega)$ and excitation spectrum $\underline{\chi}(\omega)$ given by [55]

$$\underline{s}^{\mathrm{spch}}(\omega) = \underline{\eta}(\omega)\underline{\chi}(\omega). \tag{4.6}$$

Here, $\underline{\chi}(\omega)$ is a function of a pitch frequency ω_{p} given by

$$\underline{\chi}(\omega) = \sum_{q=1}^{Q} \underline{\varpi}(\omega - q\omega_{\mathrm{p}}), \tag{4.7}$$

where q is the harmonic index, Q is the number of harmonics, and $\underline{\varpi}(\omega)$ is the Fourier transform of the Hamming window.

Now consider extracting the harmonic information from the beamformer en-

hanced signal $\widehat{\underline{s}}(\omega)$ via MBE model fitting. According to the MBE model, two variables ω_{p} and $\underline{\eta}(\omega)$ exhibit good representation of the voiced speech signal. These variables are aimed to be estimated via minimization of the fitting error between $\widehat{\underline{s}}(\omega)$ and the MBE modeled signal $\underline{\eta}(\omega)\underline{\chi}(\omega)$ given by

$$\mathcal{E}(\omega_{\mathrm{p}}) = \int_0^{2\pi} |\widehat{\underline{s}}(\omega) - \underline{\eta}(\omega)\underline{\chi}(\omega)|^2 d\omega, \tag{4.8}$$

where $\widehat{\underline{s}}(\omega)$ has been defined in (4.5).

In practice, in order to solve the non-linear minimization problem in (4.8), the whole spectrum is decomposed into Q harmonic bands. The qth harmonic band ranges in the interval $[a_q,\ b_q]$, where the lower and upper limits are defined as $a_q = \lceil (q-0.5)\omega_{\mathrm{p}} \rfloor$ and $b_q = \lceil (q+0.5)\omega_{\mathrm{p}} \rfloor$, respectively, and $\lceil \cdot \rfloor$ denotes the selection of the nearest frequency bin. The spectrum envelop $\underline{\eta}(\omega)$ is therefore decoupled into complex amplitude $\underline{\eta}_q$ for each harmonic band q, so that the fitting error for each harmonic band is

$$\varepsilon_q(\omega_{\mathrm{p}}) = \int_{a_q}^{b_q} |\widehat{\underline{s}}(\omega) - \underline{\eta}_q\underline{\chi}(\omega)|^2 d\omega, \tag{4.9}$$

and the total error in (4.8) becomes

$$\mathcal{E}(\omega_{\mathrm{p}}) = \sum_{q=1}^{Q} \varepsilon_q(\omega_{\mathrm{p}}). \tag{4.10}$$

Note that there is a subtle difference between (4.10) and (4.8). In (4.10) only the Q harmonic bands of interest is used in summation, while in (4.8) the whole spectrum is integrated. Now, the variable $\underline{\eta}_q(k)$ can be obtained by considering the derivative of (4.9) to be zero giving

$$\underline{\eta}_q = \frac{\displaystyle\int_{a_q}^{b_q} \widehat{\underline{s}}(\omega)\underline{\chi}^*(\omega) d\omega}{\displaystyle\int_{a_q}^{b_q} |\underline{\chi}(\omega)|^2 d\omega}, \tag{4.11}$$

where $(\cdot)^*$ denotes conjugate operation. The pitch frequency ω_{p} can be estimated by the following steps: each fitting error $\varepsilon_q(\omega_{\mathrm{p}})$ is evaluated using the optimal value of $\underline{\eta}_q$ obtained in (4.11). The error function in (4.10) is then computed with respect to all pitch frequencies ω_{p} of interest. Finally, the global minimum of $\mathcal{E}(\omega_{\mathrm{p}})$ is determined and the corresponding ω_{p} is selected as the estimated $\widehat{\omega}_{\mathrm{p}}$ due to speech.

4.4.3 HSRP beamformer and measurement likelihood

To obtain the feature-directed particle weight update, it is necessary to determine the most reliable harmonic bands and select those harmonic bands for the computation of the likelihood. Two criteria are proposed to determine the reliability of the harmonic bands: (1) the normalized fitting error and (2) the normalized harmonic energy.

First, the normalized fitting error [56] is defined, for each harmonic, as the effectiveness of a given frequency band to be fitted the speech harmonic model. It is computed as

$$\bar{\varepsilon}_q = \frac{\varepsilon_q(\widehat{\omega}_{\mathrm{p}})}{\int_{a_q}^{b_q} |\underline{\widehat{s}}(\omega)|^2 d\omega}, \tag{4.12}$$

where the fitting error $\varepsilon_q(\widehat{\omega}_{\mathrm{p}})$ is computed by substituting the estimated pitch frequency $\widehat{\omega}_{\mathrm{p}}$ into (4.9). The fitting error is normalized by the energy of each corresponding harmonic band.

In the second step, the normalized harmonic energy, defined by the ratio of energy distributed on that harmonic over the total energy, i.e.,

$$h_q = \frac{\int_{a_q}^{b_q} \underline{\eta}_q \underline{\chi}(\omega) d\omega}{\sum_{q=1}^{Q} \int_{a_q}^{b_q} \underline{\eta}_q \underline{\chi}(\omega) d\omega}. \tag{4.13}$$

is computed. Since the energy of the speech signal is expected to be concentrated in a structure that exhibits harmonicity, harmonic bands with low $\bar{\varepsilon}_q$ and high h_q

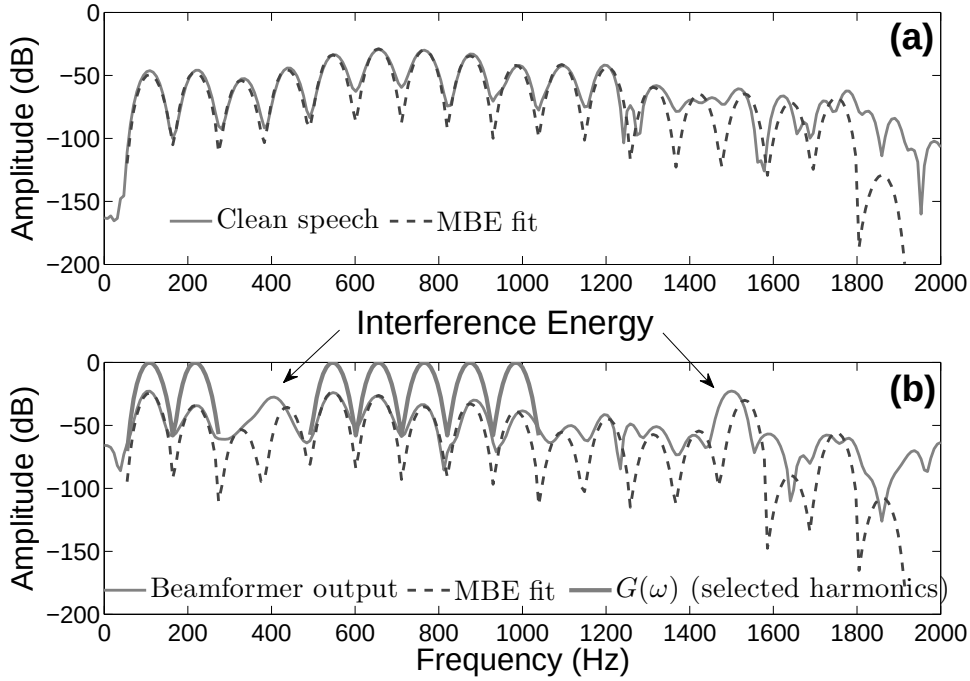


Figure 4.6: MBE fitting result. (a) Clean speech and MBE fit. (b) Beamformer output, MBE fit and $G(\omega)$ in the presence of a power drill signal.

are more likely to retain most of the speech components, while other regions are expected to contain the interference signal. Therefore, two thresholds $\bar{\varepsilon}_{\text{thr}}$ and h_{thr} are used for selecting the reliable (speech) harmonic bands such that

$$G_q(\omega) = \begin{cases} |\varpi(\omega - q\widehat{\omega}_{\text{p}})|, \text{ if } \bar{\varepsilon}_q \leq \bar{\varepsilon}_{\text{thr}} \ \& \ h_q \geq h_{\text{thr}}, \ \omega \in [a_q, b_q]; \\ 0, \text{ otherwise}, \end{cases} \tag{4.14}$$

$$G(\omega) = \sum_{q=1}^{Q} G_q(\omega). \tag{4.15}$$

Equation (4.14) indicates that only harmonic bands which satisfy the thresholds are selected while the other frequency bands are discarded. Equation (4.15) indicates that the selection process is carried out across the entire frequency bands of interest. The sum of the selected harmonic bands are denoted as $G(\omega)$.

Figure 4.6 shows extraction results of the speech harmonics using a frame of

32 ms. Figure 4.6 (a) shows the MBE fitting result, computed using (4.9)-(4.11), for the case of a clean speech where no interferer is present. One can observe that the MBE approximation, shown by the dotted line, is capable of estimating the harmonics of the clean speech. Figure 4.6 (b) shows result for the case where a power-drill signal is added into the speech signal at an SIR=5 dB. The spectrum of the beamformer output $\underline{\widehat{s}}(\omega)$, shown by the solid line, therefore consists of spectral components corresponding to the power drill at 400 and 1500 Hz and the speech signal. Comparing Figs. 4.6 (a) and (b), one can observe that the MBE fit shown in Fig. 4.6 (b) is able to estimate the speech harmonics with reasonable accuracy albeit with some estimation errors. The estimated reliable speech harmonic bands are shown with $G(\omega)$ and are denoted by the bold lines (which has been normalized to 0 dB for clarity.) The selected harmonics over all the frames are shown in Fig. 4.4 (d) where a 6 s speech in the presence of power-drill interference is considered. One can observe that employing the beamformer and MBE fit, speech harmonic bands can be estimated as indicated by the dark lines of the spectrogram.

The aforementioned procedures from (4.3) to (4.15) can iteratively be computed for every incoming signal frame $\underline{y}_i(\omega, k)$ such that the obtained $G(\omega)$ can be denoted as $G(\omega, k)$ by adding the notation of frame index k. Given the extracted harmonicity-based feature $G(\omega, k)$, the proposed HSRP beamformer function $\mathcal{P}_k^{\text{Ham}}(\mathbf{r})$ is defined in a way similar to the conventional SRP function as

$$\mathcal{P}_k^{\text{Ham}}(\mathbf{r}) = \int_{\Omega} \left| \sum_{i=1}^{N} w(\omega, k) \underline{y}_i(\omega, k) e^{\jmath\omega \frac{\|\mathbf{r}-\mathbf{r}_i^{\text{mic}}\|}{c}} \right|^2 d\omega, \tag{4.16}$$

in which the weighting function $w(\omega, k)$ is modified by adding the selection of the harmonic bands $G(\omega, k)$ as

$$w(\omega, k) = \frac{G(\omega, k)}{|\underline{y}_i(\omega, k)|}. \tag{4.17}$$

In (4.16), the parameter Ω is defined as the frequency range over which the HSRP

Table 4.1: Summary of the proposed SIRPF-HSRP algorithm.

At time $k-1$, given that a set of particles $\{\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of posterior $p(\mathbf{x}_{k-1}|\mathbf{z}_{k-1})$, the posterior state estimate is $\widehat{\mathbf{x}}_{k-1}^{+} = \sum_{p=1}^{N_p} w_{k-1}^{(p)} \mathbf{x}_{k-1}^{(p)}$.

For the kth frame:

1. *Prior prediction*: Propagate the previous state estimate through (4.3) to obtain prior estimate of the current state $\widehat{\mathbf{x}}_k^{-}$.

2. *Feature extraction*: Apply beamformer according to (4.4) (4.5) to enhance the signal from the prior estimated position $\widehat{\mathbf{r}}_k^{-}$, and extract speech features using (4.9)-(4.11).

3. *Particles propagation*: Propagate each particle through the source dynamic model (4.1),
$$\mathbf{x}_k^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)}, \mathbf{u}_k).$$

4. *Posterior weights update*: Obtain the feature directed particle likelihood using (4.12)-(4.18) and each particle is then assigned a weight according to its likelihood
$$w_k^{(p)} = w_{k-1}^{(p)} p(\mathbf{z}_k|\mathbf{x}_k^{(p)}),$$
followed by normalization $w_k^{(p)} \Leftarrow w_k^{(p)} (\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$. The posterior state estimate is $\widehat{\mathbf{x}}_k^{+} = \sum_{p=1}^{N_p} w_k^{(p)} \mathbf{x}_k^{(p)}$.

5. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p} (w_k^{(p)})^2)^{-1}$.

---

function is evaluated. Similar to the pseudo-likelihood method [10, 28], the HSRP function is used to define the measurement likelihood in the SIRPF framework,

$$p(\mathbf{z}_k|\mathbf{x}_k) = \begin{cases} \left\{\mathcal{P}_k^{\text{Ham}}(\mathbf{C}\mathbf{x}_k)\right\}^{\gamma_{\mathbf{z}}}, & \text{for voiced frame;} \\ \mathcal{U}_D(\mathbf{C}\mathbf{x}_k), & \text{for unvoiced frame,} \end{cases} \tag{4.18}$$

where $\gamma_{\mathbf{z}} = 2$ is a control parameter to regulate the HSRP function for the likelihood approximation [10], and $\mathcal{U}_D(\cdot)$ is the uniform pdf over the considered enclosure

domain $D = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max}, y_{\min} \leq y_k \leq y_{\max}\}$. The likelihood function is then used to update the particle weights. The proposed SIRPF-HSRP is summarized in Table 4.1. The obtained source position estimate corresponds to the first two elements of the posterior state estimate denoted by

$$\mathbf{r}_k^{\text{src}+} = \mathbf{C}\widehat{\mathbf{x}}_k, \tag{4.19}$$

where $\widehat{\mathbf{x}}_k$ is the posterior state estimate and $\mathbf{C}$ has been defined in (2.47).

4.5 Simulation results

Simulations were conducted using synthetic impulse responses generated by the method of images [45]. The dimension of the room was 5 m $\times$ 5 m $\times$ 2.5 m, and the reverberation time T_{60} was varied between 200 and 300 ms. Eight microphones were distributed 0.5 m away from the perimeter of the room as shown in Fig. 4.7. An 8 s male speech signal sampled at 16 kHz from the TIMIT database [44] was used as a source signal. A power drill (PD) signal and a recorded telephone ring (TR) signal obtained from the NOISEX-92 database [48] were used as interferers. White Gaussian noise of 15 dB SNR was added to the microphone signals. The speed of source was approximately set at 0.6 m/s. The positions of speech source were estimated using a frame size of 512 samples with $N_p = 100$ particles. The remaining parameters include an effective sample size threshold $N_{\text{thr}} = 37.5$, harmonic-band thresholds $\bar{\varepsilon}_{\text{thr}} = 0.6$ and $h_{\text{thr}} = 0.03$. A total of 12 harmonic bands ($Q = 12$) was considered. The proposed SIRPF-HSRP algorithm is compared with the conventional SIRPF-SRP algorithm [10]. Both algorithms were evaluated using $0 \leq \Omega \leq 2$ kHz from which, for the proposed algorithm, speech pitch frequency was estimated from 100 to 300 Hz using (4.9)-(4.11). In this chapter, the performance is quantified using

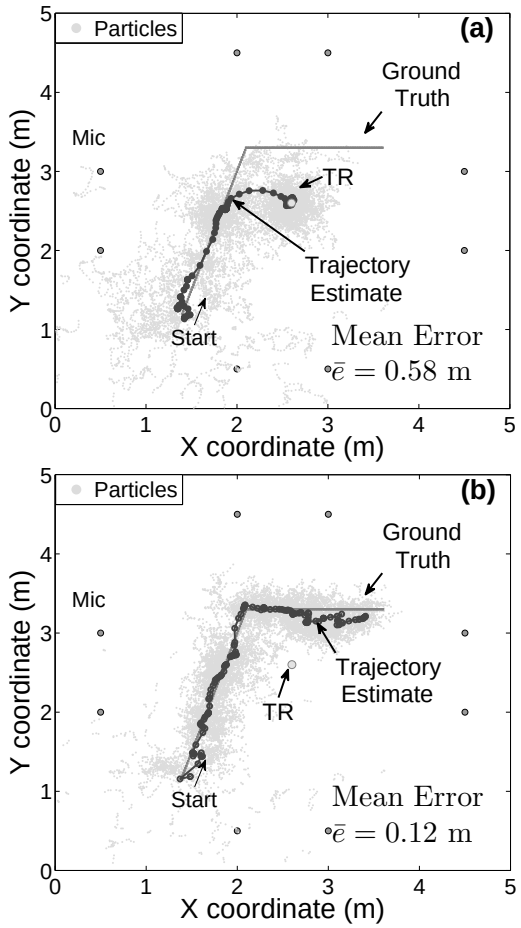


Figure 4.7: Comparison of tracking results when telephone ring (TR) is present at SIR = -3 dB, $T_{60} = 250$ ms. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-HSRP tracking algorithm.

the averaged tracking error across all audio frames, i.e.,

$$\bar{e} = \frac{1}{K} \sum_{k=1}^{K} ||\widehat{\mathbf{r}}_k^{\text{src}+} - \mathbf{r}_k^{\text{src}}||_2, \tag{4.20}$$

where $\widehat{\mathbf{r}}_k^{\text{src}+}$ is the posterior estimated position at kth frame, $\mathbf{r}_k^{\text{src}}$ is the true source position, $||\cdot||_2$ is the Euclidean norm and K is the number of frames.

Figure 4.7 compares the tracking result for $T_{60} = 250$ ms in the presence of a telephone ring at -3 dB SIR. Figure 4.7 (a) shows that the tracking performance of the conventional SIRPF-SRP approach is adversely affected by the interferer. Due

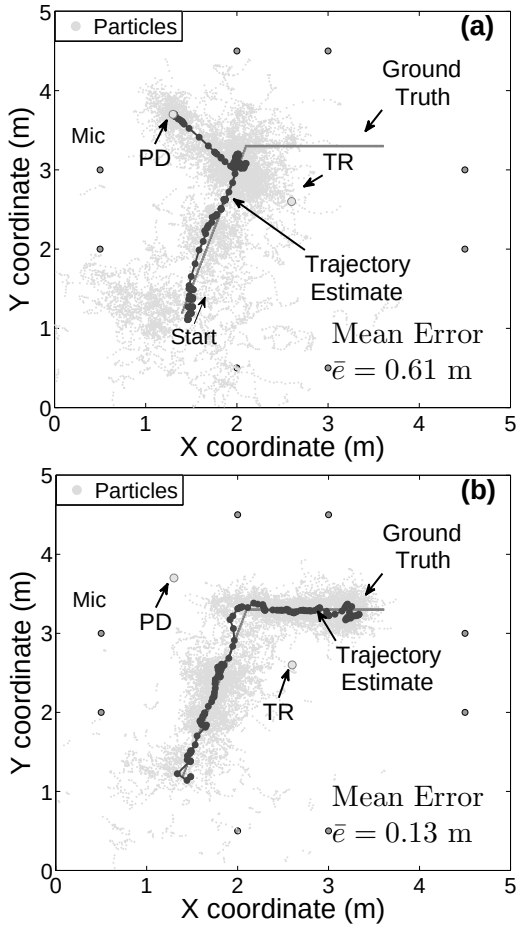


Figure 4.8: Comparison of tracking results when both power drill (PD) and telephone ring (TR) are present at SIR = 3 dB, 0 dB, respectively, T_{60} = 250 ms. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-HSRP tracking algorithm.

to the high measurement likelihood of the conventional approach in region within the neighborhood of the interferer, the particles will be 'trapped' once they are propagated there, which, in this case corresponded to the region near the telephone ring. The conventional SIRPF-SRP algorithm has an average error of 0.58 m indicating that it does not converge to the speech source trajectory. On the other hand, Fig. 4.7 (b) shows the tracking performance of the proposed SIRPF-HSRP algorithm. This result shows that the proposed algorithm is only modestly affected by the presence of the telephone ring achieving an average error of 0.12 m. In addition, it is worth noting from this comparison that while the SIRPF-HSRP algorithm

Table 4.2: Comparison of mean tracking error $\bar{e}$ between the conventional SIRPF-SRP tracking algorithm and the proposed SIRPF-HSRP tracking algorithm.

	Conventional SIRPF-SRP		Proposed SIRPF-HSRP	
	$T_{60} = 0.2$ s	$T_{60} = 0.3$ s	$T_{60} = 0.2$ s	$T_{60} = 0.3$ s
PD (SIR = 3 dB)	0.56 m	0.59 m	0.11 m	0.15 m
TR (SIR = 0 dB)	0.51 m	0.59 m	0.09 m	0.13 m
TR (SIR = -3 dB)	0.53 m	0.64 m	0.10 m	0.15 m
PD+TR (SIR = 3, 0 dB)	0.57 m	0.68 m	0.12 m	0.16 m
PD+TR (SIR = 3, -3 dB)	0.65 m	0.69 m	0.15 m	0.18 m
PD+TR (SIR = 3, -6 dB)	1.08 m	1.01 m	0.20 m	0.75 m

focuses only on the voiced frames of the speech signal as can be seen from (4.6), its performance is higher than that of conventional SIRPF-SRP algorithm.

Figure 4.8 shows the tracking result when both power drill and telephone ring are present at 3 dB and 0 dB SIRs, respectively, with $T_{60} = 250$ ms. Again, Fig. 4.8 (a) shows the conventional SIRPF-SRP approach losing track of the speech source. The particles are 'trapped' at the region near the power drill, leading to the average error of 0.61 m. On the other hand, the proposed SIRPF-HSRP algorithm, shown in Fig. 4.8 (b), retains its robustness with an average error of 0.13 m.

Table 4.2 shows the average tracking error for various test conditions. The source trajectory and interference positions remain the same as previous setup. These results show that the proposed SIRPF-HSRP algorithm can achieve better accuracy than the conventional SIRPF-SRP algorithm. For instance, in the presence of power drill at 3 dB SIR, the conventional algorithm exhibits a large tracking error of 0.56 m when $T_{60} = 0.2$ s. The proposed algorithm achieves an error of 0.11 m which translates to an 80% reduction of error over the conventional algorithm. Furthermore, the proposed algorithm maintains its robustness in localization and tracking in the presence of two interferers while the conventional approach suffers from large tracking error under low SIR condition. However, it is also observed that the performance of the proposed algorithm degrades modestly when reverberation

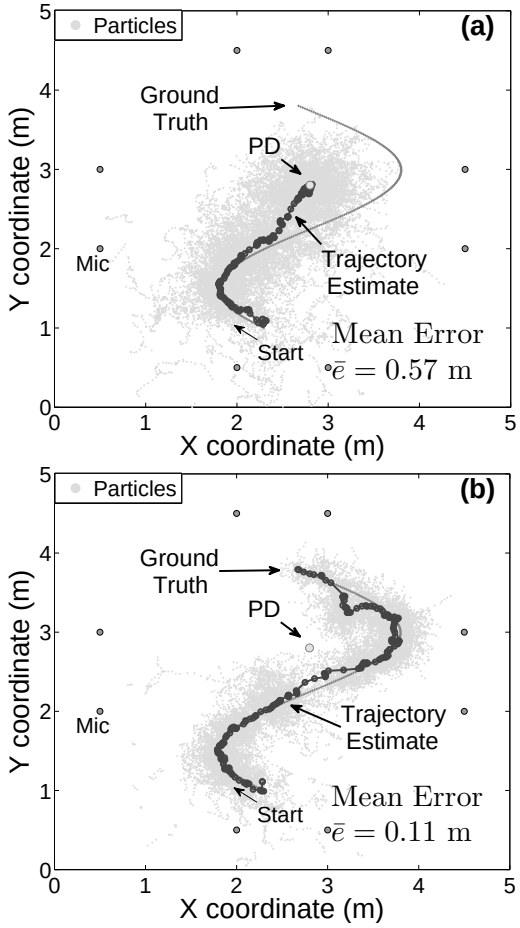


Figure 4.9: Comparison of tracking results when power drill (PD) is present at SIR = 3 dB, $T_{60} = 200$ ms. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-HSRP tracking algorithm.

time is increased. The proposed algorithm may fail under adverse environment as indicated for the case of $T_{60} = 0.3$ s when both the power drill and the phone ring are present at SIR of 3 and -6 dB.

Various source trajectories and interference configurations were also examined as shown in Figs. 4.9 and Figs. 4.10. As before, these results show that the conventional SIRPF-SRP approach is likely to be affected by interferers while the proposed SIRPF-HSRP approach retains its robustness; the particles are propagated closely along the source trajectory.

Figure 4.11 shows the performance of both algorithms under different rever-

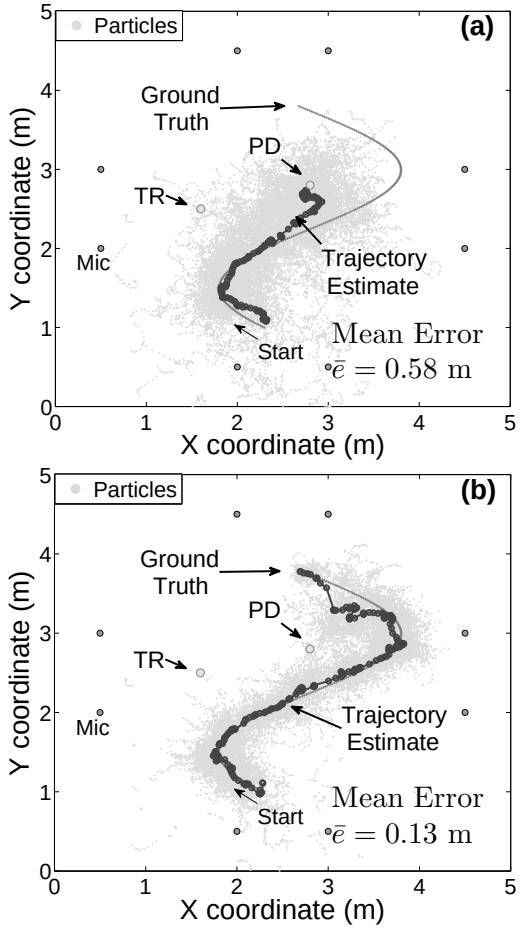


Figure 4.10: Comparison of tracking results when both power drill (PD) and telephone ring (TR) is present at SIR = 3 dB, 0 dB, respectively, T_{60} = 200 ms. (a) Conventional SIRPF-SRP tracking algorithm [10]. (b) Proposed SIRPF-HSRP tracking algorithm.

beration conditions. Fig. 4.11 (a) shows the results when a power drill is present at an SIR = 0 dB. The conventional SIRPF-SRP tracking algorithm, indicated by the dashed line, results in consistently high tracking errors of more than 1 m. The proposed SIRPF-HSRP algorithm, shown by the solid line, results in errors of less than 0.3 m when T_{60} is below 0.35 s. However, the performance deteriorates rather significantly when T_{60} is beyond 0.4 s. This performance degradation is mainly due to the increased reverberation time which causes error in harmonic feature extraction. In an adverse environment, the reverberation effect would cause distortion of the harmonic structure of the received speech signal, especially in the high frequency

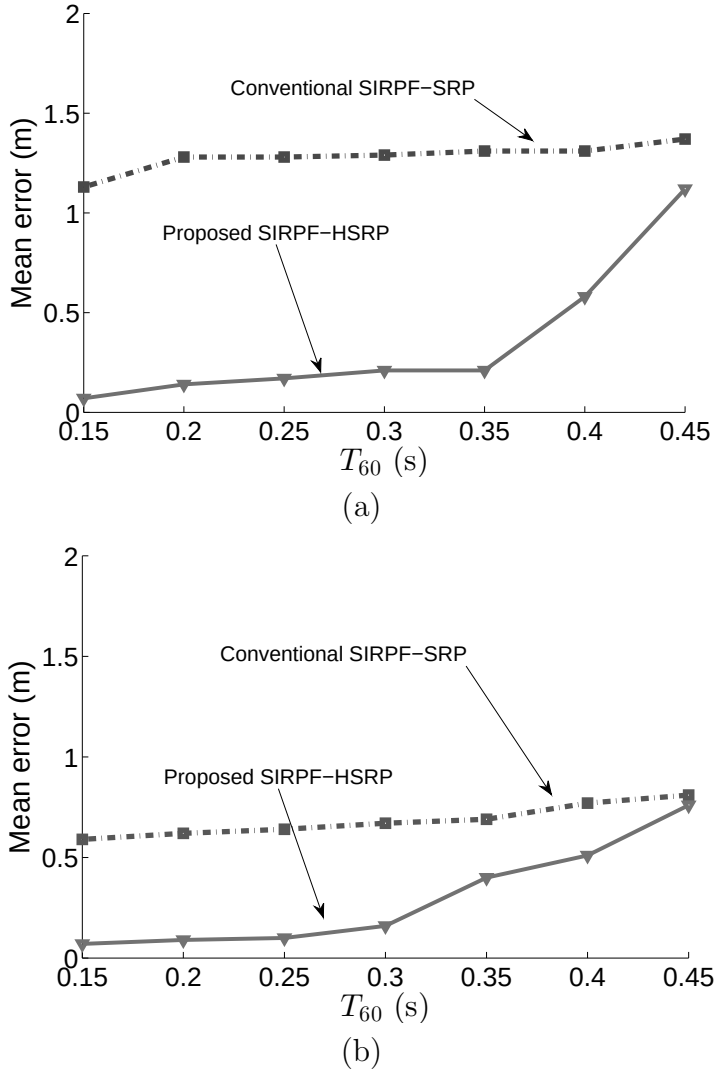


Figure 4.11: Comparison of mean tracking error versus different reverberation time T_{60}. (a) Power drill is present at SIR = 0 dB. (b) Telephone ring is present at SIR = -5dB.

range. This would result a degradation of MBE fitting method on detecting the reliable harmonic bands corresponding to speech components. As a consequence, the algorithm may fail to distinguish speech against interference in its tracking result and result in high tracking error in a very reverberant environment. Similar result can be observed from Fig. 4.11 (b) where the telephone ring is present at SIR = -5 dB. The conventional tracking algorithm consistently results in high tracking errors of more than 0.5 m, while the proposed algorithm deteriorates when T_{60} is higher

than 0.3 s.

4.6 Chapter summary

In this chapter, the SIRPF-HSRP speech tracking algorithm, which is robust to sound interference is proposed. This algorithm is capable of estimating the speech harmonic bands to achieve tracking of a speech source. By only emphasizing the harmonic bands, a better speech-sensitive measurement likelihood can be achieved resulting in a better weight update for the particles. Simulation results presented in this chapter show that the proposed SIRPF-HSRP algorithm can achieve a lower tracking error than the conventional SIRPF-SRP algorithm in the presence of multiple interferers.

Chapter 5

Alternating Source Tracking

In Chapter 3 and 4, two algorithms have been proposed for single-source tracking in the presence of noise, reverberation and sound interference. Both of them utilize the conventional sequential-importance-resampling particle filter (SIRPF) framework but incorporating new measurement likelihood approximations. Apart from single-source tracking, this chapter considers the problem of tracking alternating acoustic sources in which multiple sources are active alternately. This may occur, for example, in an interactive classroom environment where the instructor is moving around and the occupants (instructor and students) speak alternatively during an interaction session. For such a scenario, the tracking performance of the SIRPF framework may degrade significantly due to its incapability of detecting any sudden transition between sources when an alternation occurs. In this chapter, a swarm intelligence based PF (SWIPF) framework is proposed which jointly exploits the advantages of particle swarm optimization (PSO) and PF for tracking of alternating sources. The performance of the proposed SWIPF algorithm will be verified via comparison with state-of-the art algorithms for both simulated and actual room

Part of this chapter has been published as K. Wu, V. G. Reju, A. W. H. Khong and S. T. Goh, "Swarm Intelligence Based Particle Filter for Alternating Speaker Localization and Tracking Using Microphone Arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, 2017.

environment.

5.1 Introduction

This chapter considers the problem of localizing and tracking the position of a dominant source which alternates frequently between multiple people. This may occur, for example, in an interactive classroom environment where the instructor is moving around and the occupants (instructor and students) speak alternatively during an interaction session. In addition, acoustic interference such as fan noise, or signals from other source(s), may also be present. In such a scenario, the dominant talker needs to be tracked in the presence of interferers and noise. Apart from the well-studied multi-source scenario where multiple simultaneously active sources are to be tracked [17,37,39,41,57–60], the considered alternating source scenario presents two different challenges: (i) the rapid change in positions of interest, where the algorithm is required to switch from the previous speaker to the active (desired) speaker rapidly, and (ii) the presence of background noise, interference and reverberation, where tracking performance is to be maintained for frames with low signal-to-noise ratio (SNR), signal-to-interference ratio (SIR) and/or signal-to-reverberation ratio (SRR).

Given the above challenges, direct application of the SIRPF [10,28] for tracking of alternating sources will not achieve good performance. This is because SIRPF only employs a prior propagation density for particle sampling, and its performance is highly dependent on the predefined source-dynamic model. The SIRPF algorithms based on the single-source dynamic model [10, 28] may result in a lag in detection of the newly active source when an alternation occurs. The multi-source tracking algorithms [17,37,39,41,57–60] are also not suitable for the considered scenario since these algorithms may wrongly take the interferer as the desired source. For exam-

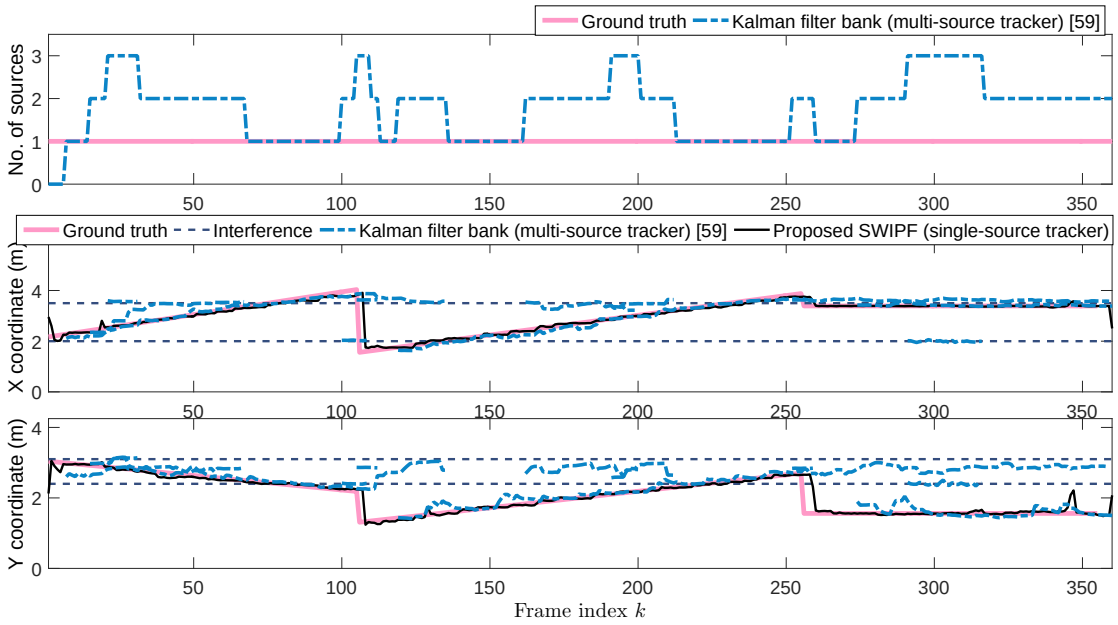


Figure 5.1: A single trial result for illustrating the problem of over estimation of number of the sources using the Kalman filter bank algorithm [59]. The result is generated using simulated data with $T_{60} = 300$ ms and SNR = 15 dB. Three talkers are active in turns and two interferers are present with SIR = 6 dB.

ple, the algorithm proposed in [59] employs a bank of Kalman filters where each filter can be initialized to track one of the talkers if he/she keeps active. When this algorithm is directly applied for the alternating-source scenario, the Kalman filters will be wrongly initialized which, in turn, results in tracking any undesired active interferers. Similarly, the probability hypothesis density (PHD) filter has been proposed where the birth/death mechanism was introduced to deal with the case of addition/removal of active source [17, 41]. When this algorithm is directly applied for alternating source scenario, the birth state will be wrongly initialized and associated with the undesired interferers. Note that without additional post-processing and given only coordinates (of source and interferers) as outputs, existing algorithms are unable to discriminate between the source and interferers. In addition, these multi-source tracking algorithms do not have any mechanism to quickly detect a new dominant talker once an alternation has occurred [12]. To verify the above, Fig. 5.1 shows a simulation result of the Kalman filter bank [59]. Three talkers are

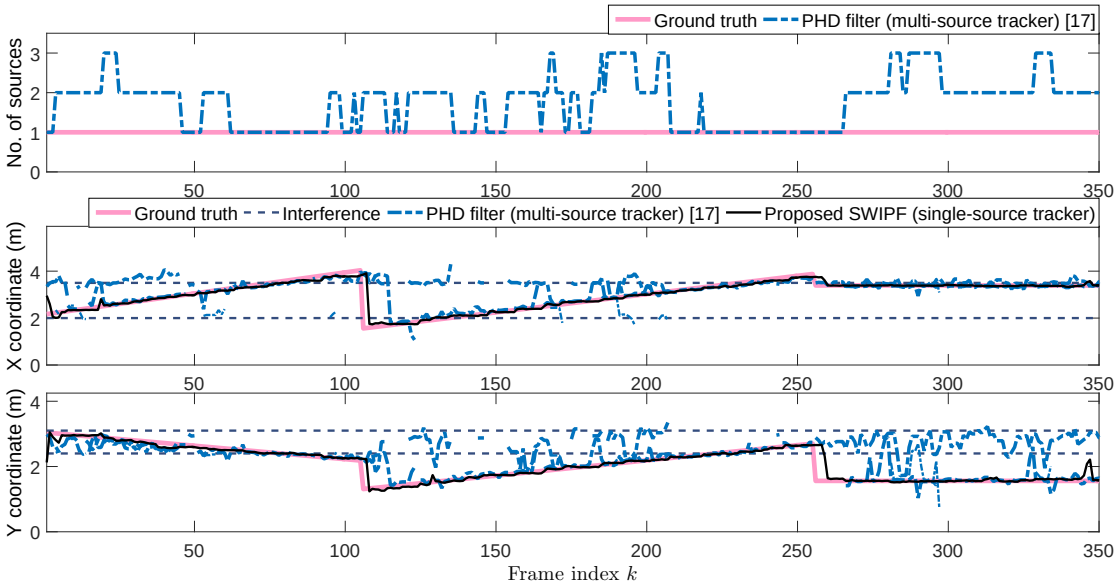


Figure 5.2: A single trial result for illustrating the problem of over estimation of number of the sources using the PHD filter algorithm. The result is generated using simulated data with $T_{60} = 300$ ms and SNR = 15 dB. Three talkers are active in turns and two interferers are present with SIR = 6 dB.

active in turns in the presence of two interferers. It can be observed that the Kalman filter bank [59] frequently tracks the interferers, resulting in an over-estimation of the number of sources. In addition, this algorithm exhibits a delay in detecting the second and third talker. A similar result is shown in Fig. 5.2 where the comparison is made with the PHD filter. The PHD filter frequently tracks the undesired interferers in addition to the desired source and exhibits a delay when alternation occurs.

To address the alternating-source problem, the extended Kalman PF (EKPF) was proposed where the importance sampling (IS) density is estimated by an extended Kalman filter using the latest measurements [12, 42]. Although the EKPF has shown to achieve fast convergence, it suffers from performance degradation in a noisy and reverberant environment. This is due to the sub-optimal particle sampling when erroneous TDOA measurements are used during the low SNR, SIR and/or SRR frames.

In this chapter, a swarm intelligence based PF (SWIPF) is proposed. Inspired by the foraging behavior of a bird flock, particle swarm optimization (PSO) was originally proposed to optimize a static fitness function [61, 62], which has been applied for localizing and tracking of sources in video [63, 64] and acoustics [65, 66]. While PF incorporates a source-dynamic model for predicting the particles, making it attractive for recursive state estimation in AST, its performance is limited by the approximation of IS density and particle sampling. In addition, particles propagate independently without any interaction between themselves. On the contrary, particles of the PSO are endowed with the ability to remember their best-fit positions and interact within the population. This swarm intelligence capability found in PSO therefore is exploited for tracking alternating sources. In essence, the memory mechanism allows particles to remain at their previous best-fit positions when TDOA measurement of the current time frame is erroneous. Convergence of the particles is expected to be improved since, with inter-particle interaction, particles can now be directed towards the active source region by sharing the fitness information among themselves.

As opposed to [66–69] where either PSO or PF is partially utilized, the proposed framework jointly exploits advantages of both PF and PSO. In [66], prediction of particles is replaced by the swarm move, which excludes the capability of incorporating any source-dynamic model for the alternating source scenario. While [68, 69] exploit PSO for particle sampling, swarm intelligence is confined within each PF iteration such that memory of the best-fit information cannot be inherited across time frames to compensate the effect of any erroneous measurement. On the contrary, in the unified framework of the proposed SWIPF, PF is used for sequential state estimation, which, in turn, allows us to incorporate the proposed alternating source-dynamic model for predicting the particles effectively. The interaction mechanism from PSO is then exploited for particle convergence. Due to a newly introduced

fitness decay mechanism, the memory mechanism from PSO is exploited across time frames such that memory of the previous best-fit information can be utilized if a source-dynamic model mismatch occurs or if the current signal frame is corrupted.

This chapter is organized as follows: Section 5.2 discusses the problem formulation and Section 5.3 reviews the PSO framework. In Section 5.4, the alternating-source dynamic model is proposed and the measurement likelihood is formulated. The proposed SWIPF algorithm is discussed in Section 5.5. In Section 5.6, the performance of the proposed algorithm is compared with existing techniques via simulation and experiment while Section 5.7 concludes this chapter.

5.2 State-space formulation

Consider the problem of estimating the position of a dominant source which alternates frequently between multiple talkers across a time period. The state vector is defined as $\mathbf{x}_k = [x_k,\ y_k]^{\mathsf{T}}$ for representing the two-dimensional coordinate of the source of interest at frame k which may suddenly change if an alternation occurs. In addition, instead of using the SRP beamformer as the measurement in Chapter 3 and 4, the TDOA estimate $\widehat{\tau}_k^{(i,j)}$ is now used as measurement. As has been discussed in Chapter 2, a possible way to obtain the TDOA estimate $\widehat{\tau}_k^{(i,j)}$ is by searching for the maximum of the generalized cross-correlation (GCC) function given by

$$\widehat{\tau}_k^{(i,j)} = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} \Psi_k^{(i,j)}(\tau), \tag{5.1}$$

where $\tau_{\max} = ||\mathbf{r}_j - \mathbf{r}_i||/c$ is the maximum admissible TDOA value [18] and the GCC function

$$\Psi_k^{(i,j)}(\tau) = \int_{\Omega} w(\omega, k) \underline{y}_i(\omega, k) \underline{y}_j^*(\omega, k) e^{\jmath\omega\tau} d\omega \tag{5.2}$$

has been defined in (2.12). With N number of microphones, the measurement vector is defined by concatenating TDOA estimates for a collection of microphone pairs as $\mathbf{z}_k = [\widehat{\tau}_k^{(1,2)}, \ldots, \widehat{\tau}_k^{(i,j)}, \ldots, \widehat{\tau}_k^{(N-1,N)}]^\mathsf{T}$. As illustrated in Fig. 5.9, this work considers four microphone arrays with total of 16 microphones and each array consists of 3 pair of adjacent microphones. The pair collection can therefore be defined as $\Upsilon = \{(i,j)|j-i=1, i=1,2,3,5,6,7,9,10,11,13,14,15\}$.

The state-space model can be formulated by

$$\mathbf{x}_k = \mathcal{G}(\mathbf{x}_{k-1}, \mathbf{u}_k), \tag{5.3}$$

$$\mathbf{z}_k = \mathcal{H}(\mathbf{x}_k, \mathbf{w}_k). \tag{5.4}$$

In (5.3), the state-transition function $\mathcal{G}(\cdot)$ defines the state evolution across time frames and will be specified by a new alternating-source dynamic model in Section 5.4.1. The measurement function $\mathcal{H}(\cdot)$ in (5.4) introduces non-linearity such that

$$\begin{aligned}\mathbf{z}_k &= [\widehat{\tau}_k^{(1,2)}, \ldots, \widehat{\tau}_k^{(i,j)}, \ldots, \widehat{\tau}_k^{(N-1,N)}]^\mathsf{T} \\ &= [\mathcal{T}^{(1,2)}(\mathbf{x}_k), \ldots, \mathcal{T}^{(i,j)}(\mathbf{x}_k), \ldots, \mathcal{T}^{(N-1,N)}(\mathbf{x}_k)]^\mathsf{T} + \mathbf{w}_k,\end{aligned} \tag{5.5}$$

where $\mathcal{T}^{(i,j)}(\cdot)$ is the non-linear function defined as

$$\mathcal{T}^{(i,j)}(\mathbf{x}_k) = \frac{1}{c}\big(||\mathbf{x}_k - \mathbf{r}_j^{\text{mic}}|| - ||\mathbf{x}_k - \mathbf{r}_i^{\text{mic}}||\big). \tag{5.6}$$

The variable $\mathbf{u}_k$ denotes process noise and $\mathbf{w}_k$ denote measurement error.

The generic particle filter [29, 70], as has been discussed in Table 2.1 of Section 2.4.2, determines a solution for solving the state-space equations in (5.3) and (5.4) and obtaining $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ iteratively. In the prediction stage, particles of the previous time frame $\mathbf{x}_{k-1}^{(p)}$ are propagated to the current time frame by sampling

the IS density described by

$$\mathbf{x}_k^{(p)} \sim p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k), \tag{5.7}$$

where the superscript (IS) denotes for the importance sampling density. In the update stage, each particle weight is updated according to

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)}. \tag{5.8}$$

In general, the generic PF requires resampling to mitigate the problem of degeneration [29]. Note that the generic PF discussed here is different from the SIRPF due to the use of the optimal IS density $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$. It will be proved in this chapter that it is indeed the use of the optimal IS density that contributes to the high performance of the proposed algorithm over the SIRPF tracking framework.

5.3 Basic concept of particle swarm optimization

PSO derives an optimal solution of a fitness function $\mathcal{F}(\mathbf{x})$ within a given search space [61, 62]. In contrast to PF where each particle moves independently in the particle propagation step, PSO allows each particle to retain memory of its best-fit position and communicate with other particles. Given $\mathcal{F}(\mathbf{x})$ to be maximized, PSO is initialized using a group of randomly distributed particles $\{\mathbf{x}_0^{(p)}\}_{p=1}^{N_p}$. Two important factors are introduced: the previous best position $\mathbf{x}_{\mathrm{pb}}^{(p)}$ that each particle has found so far, and the global best position $\mathbf{x}_{\mathrm{gb}}$ found within the entire population.

Particle movement can then be described as [61]

$$\mathbf{v}_k^{(p)} = \chi\Big[\mathbf{v}_{k-1}^{(p)} + \varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\text{pb}}^{(p)} - \mathbf{x}_{k-1}^{(p)}) + \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\text{gb}} - \mathbf{x}_{k-1}^{(p)})\Big], \tag{5.9}$$

$$\mathbf{x}_k^{(p)} = \mathbf{x}_{k-1}^{(p)} + \mathbf{v}_k^{(p)}, \tag{5.10}$$

where $\mathbf{v}_k^{(p)}$ denotes the particle velocity, φ_1 and φ_2 are the acceleration parameters, $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$ are vectors with elements uniformly sampled within the range $[0, 1]$ and $\odot$ denotes element-wise product. Defining $\varphi = \varphi_1 + \varphi_2$ and $\varphi > 4$, a constriction parameter $\chi = \frac{2}{|2-\varphi-\sqrt{\varphi^2-4\varphi}|}$ is used to generate a damping effect on each particle's oscillation. Such a constriction can prevent misconvergence — a phenomenon where the particles increasingly oscillate until out-of-bound error occurs [61].

In (5.9) and (5.10), the movement of each particle mainly depends on the individual cognitive knowledge $\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\text{pb}}^{(p)} - \mathbf{x}_{k-1}^{(p)})$ and the social influence $\varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\text{gb}} - \mathbf{x}_{k-1}^{(p)})$ which drive the particle to $\mathbf{x}_{\text{pb}}^{(p)}$ and $\mathbf{x}_{\text{gb}}$, respectively, using oscillated forces introduced by $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. The variables φ_1 and φ_2 define the relative weights on these two forces. In general, $\varphi_1 = \varphi_2$ is assumed if no prior knowledge of the relative weights is available [61]. After (5.9) and (5.10), $\mathbf{x}_{\text{pb}}^{(p)}$ and $\mathbf{x}_{\text{gb}}$ are updated by

$$\mathbf{x}_{\text{pb}}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_k^{(p)}, & \text{if } \mathcal{F}(\mathbf{x}_k^{(p)}) \geq f_{\text{pb}}^{(p)}; \\ \mathbf{x}_{\text{pb}}^{(p)}, & \text{otherwise,} \end{cases} \tag{5.11}$$

$$\mathbf{x}_{\text{gb}} = \arg\max_{\mathbf{x}_{\text{pb}}^{(p)}} \mathcal{F}(\mathbf{x}_{\text{pb}}^{(p)}), \tag{5.12}$$

where $\mathcal{F}(\mathbf{x}_k^{(p)})$ denotes fitness value evaluated at the propagated particle and $f_{\text{pb}}^{(p)} = \mathcal{F}(\mathbf{x}_{\text{pb}}^{(p)})$ is the fitness value of the previous best-fit position.

The PSO presented above is well suited for a static optimization problem where $\mathcal{F}(\mathbf{x})$ and its optimal solution are time invariant. Application of the PSO

for AST, however, requires further modification since the AST involves a dynamic search of the source position which varies across time.

5.4 Proposed Models for Alternating-source Tracking

In this section, a new source-dynamic model $\mathcal{G}(\cdot)$ that is suitable for the alternating-source scenario and the corresponding $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is derived. The measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ will then be formulated by taking into account the reliability of $\widehat{\tau}_k^{(i,j)}$ due to noise, interference and/or reverberation. These two sources of information will then be used in SWIPF described in Section 5.5.

5.4.1 Alternating Source-dynamic Model

The state-transition function $\mathcal{G}(\cdot)$ in (5.3) determines the state evolution across time frames. The Langevin and random walk processes have commonly been used to model this state transition for single-source tracking application [10,28]. While these processes have also been extended for tracking alternating sources by simply increasing the process noise to account for alternation uncertainty [12, 31, 42], the underlying continuous moving-source assumption may not be valid for the alternating-source scenario.

The function $\mathcal{G}(\cdot)$ that is being proposed better suits the alternating-source scenario by introducing two hypotheses:

$\mathcal{S}0_k$: An alternation has not occurred at the kth frame;

$\mathcal{S}1_k$: An alternation has occurred at the kth frame.

Under hypothesis $\mathcal{S}0_k$, indicating that the source location is consistent with the previous frames, either a Langevin process model or a random walk model can be used to describe the continuous talker motion. In this work, due to computational efficiency, the random walk model, which models perturbation of the source location within the neighborhood region of the last iteration, is used. Therefore, the state is predicted using

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k, \quad \text{if} \quad \mathcal{S}0_k, \tag{5.13}$$

where $\mathbf{u}_k \sim \mathcal{N}(\cdot; \mathbf{0}_{2\times 1}, \mathbf{\Sigma}_{2\times 2})$, $\Sigma_{2\times 2} = \sigma_{\mathbf{u}}^2 \mathbf{I}_{2\times 2}$ is the covariance of the Gaussian distribution and $\sigma_{\mathbf{u}}^2$ defines the variance of the human motion both in x and y directions. Given (5.13), the conditioned state-transition probability can be written as

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}0_k) = \mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k-1}, \mathbf{\Sigma}_{2\times 2}). \tag{5.14}$$

On the other hand, under hypothesis $\mathcal{S}1_k$, an alternation is expected to have occurred. This leads the state being re-initialized in a uniformly distributed manner within the enclosed domain as

$$\mathbf{x}_k = \mathcal{U}_{\mathcal{D}}(\mathbf{x}_k), \quad \text{if} \quad \mathcal{S}1_k, \tag{5.15}$$

where the function $\mathcal{U}_{\mathcal{D}}(\cdot)$ denotes a multivariate uniform distribution over the enclosure domain $\mathcal{D}$. In this work, a rectangular shoe-box room is considered such that $\mathcal{D} = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max}, y_{\min} \leq y_k \leq y_{\max}\}$, where the variables $x_{\min}$, $x_{\max}$ and $y_{\min}$, $y_{\max}$ denote boundaries in the x and y directions, respectively. The corresponding conditioned state-transition probability can be written as

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}1_k) = \frac{1}{(x_{\max} - x_{\min})(y_{\max} - y_{\min})}. \tag{5.16}$$

Additional simulation shows that the use of Langevin process will not bring significant performance improvement compared to the random walk model since the effect of Langevin process can be reduced by the swarm update in (5.30).

Apart from $\mathcal{U}_{\mathcal{D}}(\cdot)$ in (5.15), other state distributions may also be used to reflect the occurrence of any new active source, e.g., a lower probability near the room boundaries. However, determination of this pdf is environment dependent and beyond the scope of this chapter.

Given the two hypotheses, the state-transition probability in (5.8) can then be computed by

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}0_k)p(\mathcal{S}0_k|\mathbf{x}_{k-1}) + p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}1_k)p(\mathcal{S}1_k|\mathbf{x}_{k-1}). \quad (5.17)$$

Here $p(\mathcal{S}1_k|\mathbf{x}_{k-1})$ and $p(\mathcal{S}0_k|\mathbf{x}_{k-1})$ are the prior probabilities of the hypothesis $\mathcal{S}1_k$ and $\mathcal{S}0_k$, respectively. The following defines $p(\mathcal{S}1_k|\mathbf{x}_{k-1}) = P_{\text{alt}}$ and $p(\mathcal{S}0_k|\mathbf{x}_{k-1}) = 1 - P_{\text{alt}}$, where P_{alt} is a predefined probability value for alternation.

In the above derivation, the source is assumed to alternate at every frame with an empirical probability of P_{alt}. This assumption may be violated when the existing source is continuously active. Hence, direct incorporation of the proposed $\mathcal{G}(\cdot)$ into the SIRPF will still lead to model mismatch. This model mismatch problem will be addressed using the proposed memory mechanism within SWIPF described in Section 5.5.

5.4.2 Measurement Likelihood

The GCC function $\Psi_k^{(i,j)}(\tau)$ defined in (5.2) is used to obtain $\widehat{\tau}_k^{(i,j)}$, which will then be used to formulate $p(\mathbf{z}_k|\mathbf{x}_k)$. As shown in the upper plot of Fig. 5.3, the peak corresponding to the true TDOA may be lower than the spurious peaks caused by interference, noise and reverberation. Therefore, as opposed to using only the maximum of $\Psi_k^{(i,j)}(\tau)$ described in (5.1), multiple peaks in $\Psi_k^{(i,j)}(\tau)$ are considered in order to increase the including the true peak. In this work, for each $\Psi_k^{(i,j)}(\tau)$, peaks that are higher than 0.7 of the maximum peak will be used as TDOA candidates.

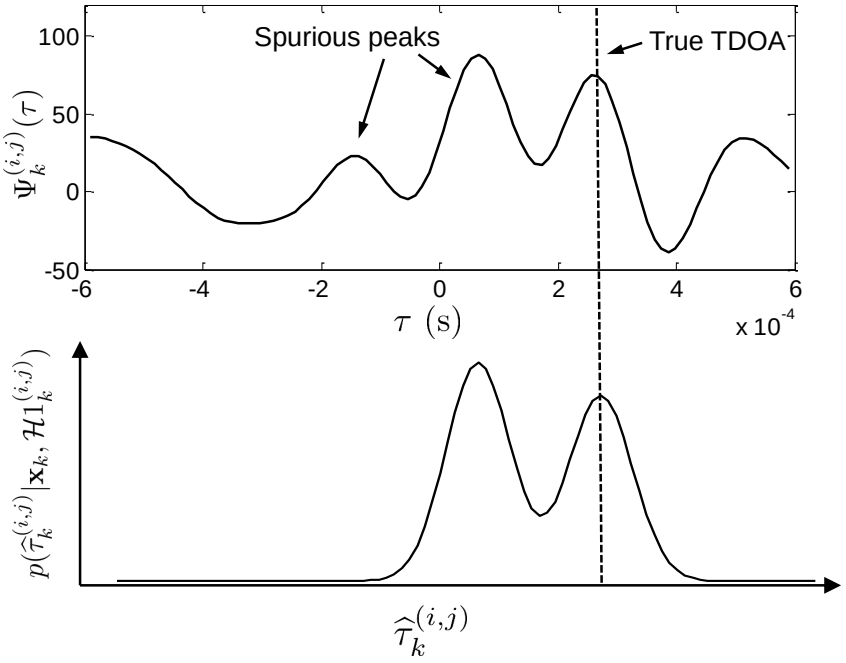


Figure 5.3: Upper plot: GCC function computed using (5.2) for actual recorded signals with the same room-source setup described in Sec. 5.6 at an estimated SNR = 15 dB and $T_{60} = 0.35$ s; Lower plot: conditioned measurement likelihood $p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}1_k^{(i,j)})$ computed as a mixture of Gaussian pdfs as in (5.19) using the peaks in upper plot.

These candidates are denoted by

$$\{\widehat{\tau}_k^{(i,j),\ell}\}_{\ell=1}^{L}, \tag{5.18}$$

where ℓ is the peak index and L is the number of peaks for that pair.

Furthermore, due to variation in speech energy, the received signal power may vary across different microphone pairs and time frames. Two hypotheses are introduced to describe the reliability of estimated TDOA:

$$\mathcal{H}0_k^{(i,j)}: \text{ TDOA estimate } \widehat{\tau}_k^{(i,j)} \text{ being unreliable;}$$
$$\mathcal{H}1_k^{(i,j)}: \text{ TDOA estimate } \widehat{\tau}_k^{(i,j)} \text{ being reliable.}$$

Under hypothesis $\mathcal{H}1_k^{(i,j)}$, one of the elements in $\{\widehat{\tau}_k^{(i,j),\ell}\}_{\ell=1}^{L}$ will correspond to the true TDOA. Given that it is uncertain as to which $\widehat{\tau}_k^{(i,j),\ell}$ corresponds to the true

TDOA, it is proposed to include all the elements in $\{\widehat{\tau}_k^{(i,j),\ell}\}_{\ell=1}^{L}$ such that the conditioned measurement likelihood is formulated as a mixture of Gaussian pdfs given by [28]

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}1_k^{(i,j)}) = \sum_{\ell=1}^{L} \bar{a}^{\ell}\, \mathcal{N}(\widehat{\tau}_k^{(i,j),\ell}; \mathcal{T}^{(i,j)}(\mathbf{x}_k), \sigma_{\tau}^2), \tag{5.19}$$

where $\bar{a}^{\ell} = a^{\ell}/\sum_{\ell=1}^{L} a^{\ell}$ is the normalized amplitude of each TDOA candidate given that a^{ℓ} is the amplitude of the ℓth peak of $\Psi_k^{(i,j)}(\tau)$, $\mathcal{T}^{(i,j)}(\cdot)$ is the nonlinear function defined in (5.6), and σ_{τ}^2 is the variance of the Gaussian component. The normalization after (5.19) ensures that the integral of the derived pdf equals to unity. It also implies that, as opposed to using only the maximum within the set $\{\widehat{\tau}_k^{(i,j),\ell}\}_{\ell=1}^{L}$, amplitude-based weightings are being applied on each candidate candidate $\widehat{\tau}_k^{(i,j),\ell}$. These weights do not compromise the fact that $\widehat{\tau}_k^{(i,j),\ell}$ with a higher amplitude, in higher probability that it corresponds to the true TDOA of the desired dominant source. The lower plot of Fig. 5.3 shows the Fig. 5.3 shows the derived $p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}1_k^{(i,j)})$ computed using the two peaks of Although the maximum of the pdf may not correspond to the true TDOA, a high probability is derived for the true TDOA. It is worth noting that a high density caused by reverberation, noise and interference will not be consistent spatially (across microphone pairs) and temporally (across time frames) due to inconsistency of the spurious peaks in $\Psi_k^{(i,j)}(\tau)$ [31]. These biased effects can therefore be minimized since the proposed algorithm employs multiple microphone pairs and incorporates tracking and memory mechanism during successive frames.

For the case of hypothesis $\mathcal{H}0_k^{(i,j)}$, the conditioned measurement likelihood can be modeled as a uniform distribution within the maximum admissible TDOA range, i.e.,

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}0_k^{(i,j)}) = \frac{1}{2\tau_{\max}}, \tag{5.20}$$

where $\tau_{\max}$ has been defined after (5.1). Therefore, the measurement likelihood of

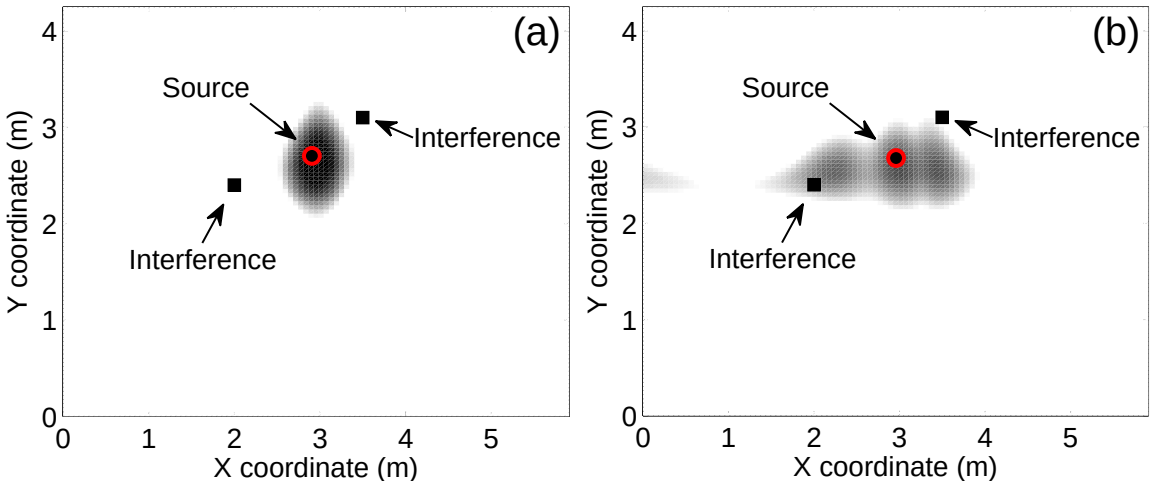


Figure 5.4: A simulation example of $p(\mathbf{z}_k|\mathbf{x}_k)$ in the log-domain as a function of $\mathbf{x}_k$ for a room environment with $T_{60} = 300$ ms and SNR = 15 dB. The desired talker and two interferers are present with the SIR = 10 dB for each interferer. (a) at the 45th frame (high source energy frame); (b) at the 47th frame (low source energy frame).

the mth microphone pair is

$$p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k) = p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}0_k^{(i,j)})p(\mathcal{H}0_k^{(i,j)}|\mathbf{x}_k) + p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k, \mathcal{H}1_k^{(i,j)})p(\mathcal{H}1_k^{(i,j)}|\mathbf{x}_k), \tag{5.21}$$

where $p(\mathcal{H}0_k^{(i,j)}|\mathbf{x}_k)$ and $p(\mathcal{H}1_k^{(i,j)}|\mathbf{x}_k)$ are the prior hypothesis probabilities. These probabilities can be determined using the instantaneous power ratio

$$\lambda_k^{(i,j)} = 10\log_{10}\left\{\frac{P_k^{(i,j)}}{P_{\text{noise}}^{(i,j)}}\right\}, \tag{5.22}$$

where $P_k^{(i,j)}$ is the instantaneous power of the microphone pair (i,j) at frame k computed by averaging over the two channels in that pair. The variable $P_{\text{noise}}^{(i,j)}$ denotes the noise power at the same microphone pair evaluated during non-speech periods obtained by the voice-activity detection algorithm [10]. By assuming that noise is stationary across a few frames during the tracking process, one can derive $P_k^{(i,j)} \geq P_{\text{noise}}^{(i,j)}$ giving $\lambda_k^{(i,j)} \geq 0$. Note that a higher value of $\lambda_k^{(i,j)}$ implies close source-sensor distance, absence of occlusion between source and sensors or high

signal power. This translates to a higher probability that $\widehat{\tau}_k^{(i,j)}$ is more reliable. The proposed algorithm therefore adopts, similar to [10], a monotonic mapping function $p(\mathcal{H}1_k^{(i,j)}|\mathbf{x}_k) = \frac{2}{\pi}\arctan(\lambda_k^{(i,j)})$ for the pdf formulation, and $p(\mathcal{H}0_k^{(i,j)}|\mathbf{x}_k) = 1 - p(\mathcal{H}1_k^{(i,j)}|\mathbf{x}_k)$.

With reference to the definition of $\mathbf{z}_k$ in (5.5) and assuming that $\widehat{\tau}_k^{(i,j)}$ from each microphone pair is independent [17,42], the overall measurement likelihood can be written as

$$p(\mathbf{z}_k|\mathbf{x}_k) = \prod_{(i,j)\in\Upsilon} p(\widehat{\tau}_k^{(i,j)}|\mathbf{x}_k). \tag{5.23}$$

where Υ denotes the microphone pair collection which has been defined in Section 5.2.

Figure 5.4 shows examples of the computed $p(\mathbf{z}_k|\mathbf{x}_k)$ as function of $\mathbf{x}_k$. It can be observed that a high density (denoted by dark color) has been achieved at the desired source position in Fig. 5.4 (a) due to high SIR, SNR and/or SRR at that frame, while during the low source energy frames in Fig. 5.4 (b), the density is comparatively significant for the interferers. The pdf $p(\mathbf{z}_k|\mathbf{x}_k)$ therefore serves as a weighting function in (5.8) only during signal frames with high SIR, SNR and/or SRR. For other frames, the proposed algorithm exploits the memory mechanism as described in the following section.

5.5 Proposed Swarm Intelligence Based PF

Performance of the PF is mainly determined by the sampling of particles. Conventional SIRPF approximates the IS density as $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k) \approx p(\mathbf{x}_k|\mathbf{x}_{k-1})$ [10,28]. However, this sub-optimal IS density may result in the sampling impoverishment problem. As illustrated in Fig. 5.5, when a mismatch between the assumed state-transition model and actual source motion occurs, $p(\mathbf{x}_k|\mathbf{x}_{k-1}) \neq p(\mathbf{x}_k|\mathbf{z}_{1:k})$ and most

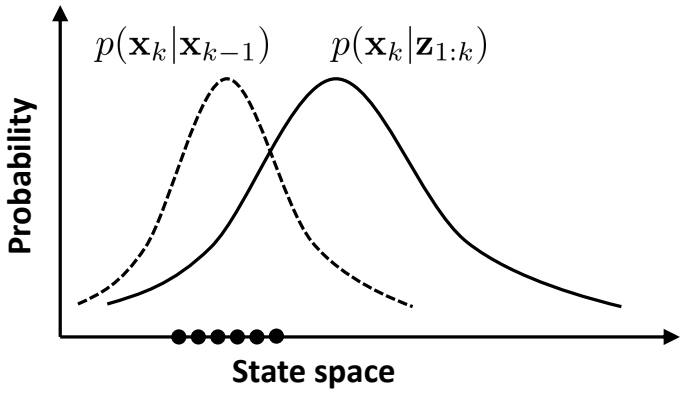


Figure 5.5: An illustration of the sampling impoverishment problem in the PF.

of the particles that are sampled from $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ contributes insignificantly for the approximation of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. In the alternating-source tracking scenario, this implies that if a single-source model (e.g. the Langevin process model [10]) is used, the particles may wrongly be sampled near the state space of the previous talker position when another talker has become active. While this lag effect can be reduced by applying the proposed alternating source-dynamic model $\mathcal{G}(\cdot)$, model mismatch may still occur where alternation is wrongly assumed for a continuously active source. As a result, particles may wrongly be sampled to detect any non-existent "new" source. The EKPF addresses the model mismatch problem by approximating $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ using the latest measurement $\mathbf{z}_k$ [12]. However, particle sampling may still be compromised if the latest measurement $\mathbf{z}_k$ is erroneous. For the talker tracking scenario, this problem occurs in frames with low SNR, SIR, and/or SRR.

To address the above problems, the proposed SWIPF exploits the advantages of PF and PSO jointly for optimal particle sampling. In the following, without loss of generality, the proposed SWIPF is first introduced in the context of a generic state estimation problem. The solution of SWIPF for alternating source tracking will then be discussed.

5.5.1 Swarm Intelligence Based PF

As opposed to the conventional particle filter where only $\mathbf{x}_k^{(p)}$ is used for frame k, the proposed SWIPF incorporates swarm intelligence attributes that comprise of historical best-fit particle $\mathbf{x}_{\text{pb}}^{(p)}$ and its previous best-fit fitness $f_{\text{pb}}^{(p)}$ for memorizing historical information. This results in the full information set of a particle given by

$$\mathcal{X}_k^{(p)} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}. \tag{5.24}$$

Furthermore, a fitness function is to be formulated for directing the particles towards the optimal position in state space. By considering the measurement likelihood $p(\mathbf{z}_k|\mathbf{x})$ as a function of $\mathbf{x}$, the fitness function is proposed as

$$\mathcal{F}_k(\mathbf{x}) \triangleq \mathcal{M}\{p(\mathbf{z}_k|\mathbf{x})\}, \tag{5.25}$$

where $\mathcal{M}(\cdot)$ denotes a monotonic function. Unlike the static function $\mathcal{F}(\cdot)$ in (5.11) and (5.12), the subscript k in $\mathcal{F}_k(\cdot)$ implies that the proposed fitness function varies across time frames for a dynamic state tracking problem for which, examples have been shown in Fig. 5.4. The subscript k in $\mathbf{x}$, however, has been temporarily omitted to imply that this time-varying function is defined for a general space of $\mathbf{x}$. The definition of (5.25) is motivated by the fact that the maximum of $\mathcal{F}_k(\mathbf{x})$ corresponds to the maximum likelihood estimate of $\mathbf{x}$ from $p(\mathbf{z}_k|\mathbf{x})$ and the particles will converge to this state estimate. A schematic diagram of the proposed SWIPF is shown in Fig. 5.6 and the detailed steps are described in the following.

A) Particle Prediction

Given the full information set of a particle at previous frame $\mathcal{X}_{k-1}^{(p)} = \{\mathbf{x}_{k-1}^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}$, It is now to derive the prediction of $\mathcal{X}_{k|k-1}^{(p)} = \{\mathbf{x}_{k|k-1}^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}$

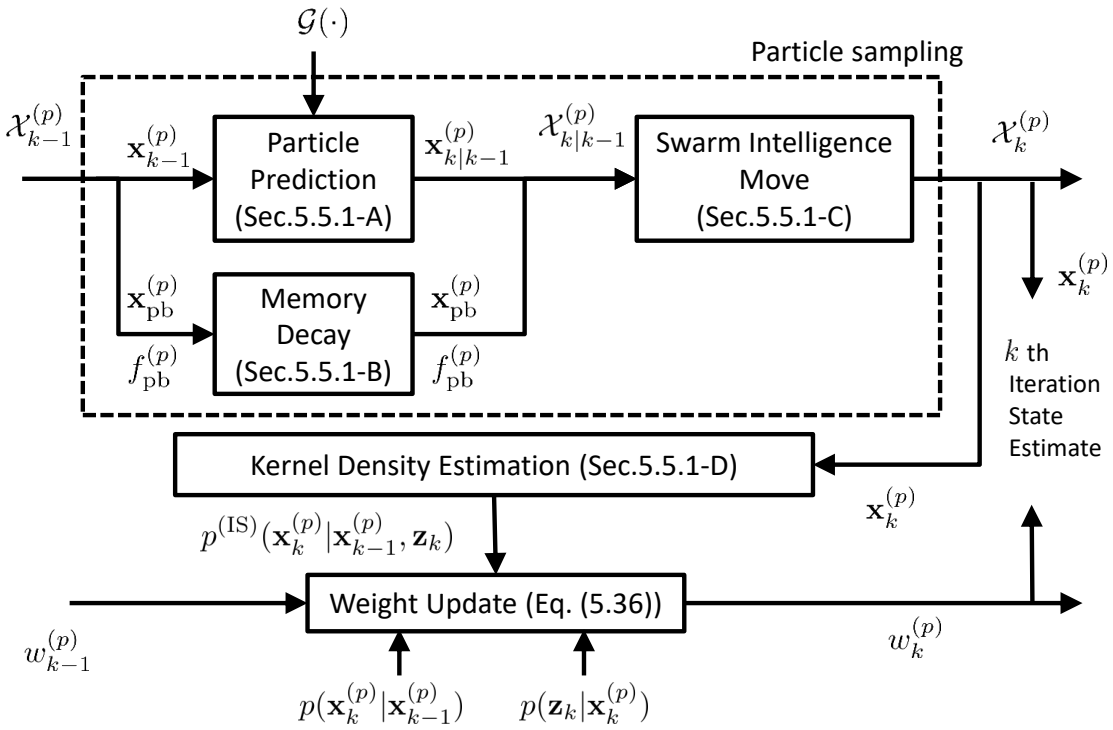


Figure 5.6: Schematic diagram of the proposed SWIPF algorithm for a single iteration.

where the subscript $k|k-1$ signifies prediction. Firstly, the state component $\mathbf{x}_{k|k-1}^{(p)}$ can be predicted using the assumed state-transition model $\mathcal{G}(\cdot)$ given by

$$\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)}). \tag{5.26}$$

Similar to conventional PF, this step propagates the particles in order to explore any potential state candidate according to knowledge of state-transition statistics.

B) Memory Decay Mechanism

To obtain $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and $f_{\mathrm{pb}}^{(p)}$ in $\mathcal{X}_{k|k-1}^{(p)}$, note that both of them represent the historical best-fit information of the particle, which is crucial to direct the particle towards the "memorised" best state estimate. Nevertheless, to deal with the time-varying nature of $\mathcal{F}_k(\cdot)$ where the past information has to be gradually reduced while the recent information should be emphasized, a linear decay of $f_{\mathrm{pb}}^{(p)}$ is first performed at

each iteration as

$$f_{\text{pb}}^{(p)} \Leftarrow f_{\text{pb}}^{(p)} - \Delta f, \tag{5.27}$$

where Δf denotes the fitness decay amount — a lower value of Δf indicates a slower memory decay such that historical information will be weighted more.

After fitness decay, comparison can be made between the latest fitness value of the predicted $\mathbf{x}_{k|k-1}^{(p)}$ evaluated using $\mathcal{F}_k(\cdot)$ and the reduced version of $f_{\text{pb}}^{(p)}$ from (5.27). If $\mathbf{x}_{k|k-1}^{(p)}$ is closer to the true state than $\mathbf{x}_{\text{pb}}^{(p)}$, and provided that $\mathcal{F}_k(\cdot)$ is reliable, one would have $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) > f_{\text{pb}}^{(p)}$. Hence, $\mathbf{x}_{\text{pb}}^{(p)}$ and $f_{\text{pb}}^{(p)}$ should be timely updated using the latest best-fit information. Otherwise, historical information should be preserved. The above can be described using

$$\mathbf{x}_{\text{pb}}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_{k|k-1}^{(p)}, & \text{if } \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \geq f_{\text{pb}}^{(p)}; \\ \mathbf{x}_{\text{pb}}^{(p)}, & \text{otherwise,} \end{cases} \tag{5.28}$$

$$f_{\text{pb}}^{(p)} \Leftarrow \begin{cases} \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}), & \text{if } \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \geq f_{\text{pb}}^{(p)}; \\ f_{\text{pb}}^{(p)}, & \text{otherwise.} \end{cases} \tag{5.29}$$

Here, the memory mechanism of particle swarm intelligence has been exploited by taking the time-varying nature of $\mathcal{F}_k(\cdot)$ into account. The historical information $f_{\text{pb}}^{(p)}$ and $\mathbf{x}_{\text{pb}}^{(p)}$ will be preserved if $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\text{pb}}^{(p)}$. This can occur if 1) $\mathbf{x}_{k|k-1}^{(p)}$ is wrongly predicted by $\mathcal{G}(\cdot)$, or if 2) the fitness function $\mathcal{F}_k(\cdot)$ is erroneous at the kth frame.

C) Swarm Intelligence Move for Optimal Particle Sampling

Given $\mathcal{X}_{k|k-1}^{(p)}$, the particle information for the kth frame $\mathcal{X}_k^{(p)} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}$ can be achieved by the swarm intelligence move. Similar to (5.9) and (5.10), $\mathbf{x}_k^{(p)}$ in

$\mathcal{X}_k^{(p)}$ can be obtained using

$$\begin{aligned}\mathbf{x}_k^{(p)} = \; & \mathbf{x}_{k|k-1}^{(p)} + \chi\big[\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\text{pb}}^{(p)} - \mathbf{x}_{k|k-1}^{(p)}) \\ & + \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\text{gb}} - \mathbf{x}_{k|k-1}^{(p)})\big],\end{aligned} \tag{5.30}$$

where $\mathbf{x}_{\text{gb}}$ is evaluated among all $\mathbf{x}_{\text{pb}}^{(p)}$, i.e.,

$$\mathbf{x}_{\text{gb}} = \arg\max_{\mathbf{x}_{\text{pb}}^{(p)}} f_{\text{pb}}^{(p)}. \tag{5.31}$$

It can be observed in (5.30) that the particle has been driven to $\mathbf{x}_k^{(p)}$ to explore for a better state estimate, according to a joint force contributed by its individual historical information $\mathbf{x}_{\text{pb}}^{(p)}$ and social interaction information $\mathbf{x}_{\text{gb}}$. Since a movement has been made for $\mathbf{x}_k^{(p)}$ from $\mathbf{x}_{k|k-1}^{(p)}$, the latest fitness value of $\mathbf{x}_k^{(p)}$ needs to be evaluated using $\mathcal{F}_k(\cdot)$ and the variables $\mathbf{x}_{\text{pb}}^{(p)}$, $f_{\text{pb}}^{(p)}$ in $\mathcal{X}_k^{(p)}$ need to be updated as

$$\mathbf{x}_{\text{pb}}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_k^{(p)}, & \text{if } \mathcal{F}_k(\mathbf{x}_k^{(p)}) \geq f_{\text{pb}}^{(p)}; \\ \mathbf{x}_{\text{pb}}^{(p)}, & \text{otherwise,} \end{cases} \tag{5.32}$$

$$f_{\text{pb}}^{(p)} \Leftarrow \begin{cases} \mathcal{F}_k(\mathbf{x}_k^{(p)}), & \text{if } \mathcal{F}_k(\mathbf{x}_k^{(p)}) \geq f_{\text{pb}}^{(p)}; \\ f_{\text{pb}}^{(p)}, & \text{otherwise.} \end{cases} \tag{5.33}$$

Finally, while $\mathbf{x}_k^{(p)}$ obtained in (5.30) represents the latest particle, the variable $\mathbf{x}_{\text{pb}}^{(p)}$ records the best-fit state estimate among $\mathbf{x}_{k-1}^{(p)}$, $\mathbf{x}_{k|k-1}^{(p)}$ and $\mathbf{x}_k^{(p)}$ in terms of closeness to the true state. It is therefore proposed to use $\mathbf{x}_{\text{pb}}^{(p)}$ as the sampled particle for the kth iteration by performing the following assignment

$$\mathbf{x}_k^{(p)} \Leftarrow \mathbf{x}_{\text{pb}}^{(p)}. \tag{5.34}$$

The swarm intelligence based particle sampling is summarized in Table 5.1.

Table 5.1: Swarm intelligence based particle sampling.

Input: $\{\mathcal{X}_{k-1}^{(p)}\}_{p=1}^{N_p} = \{\mathbf{x}_{k-1}^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}_{p=1}^{N_p}$

For each particle:

- Predict $\mathbf{x}_{k|k-1}^{(p)}$ using the state-transition model, i.e., $\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)})$.
- Apply fitness decay on $f_{\text{pb}}^{(p)}$ to de-emphasize the past information, i.e., $f_{\text{pb}}^{(p)} \Leftarrow f_{\text{pb}}^{(p)} - \Delta f$.
- Evaluate the fitness for the predicted $\mathbf{x}_{k|k-1}^{(p)}$ using $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)})$, compare with $f_{\text{pb}}^{(p)}$ and update $\mathbf{x}_{\text{pb}}^{(p)}$ and $f_{\text{pb}}^{(p)}$ using (5.28) and (5.29).

End

- Obtain global best particle using $\mathbf{x}_{\text{gb}} = \arg\max_{\mathbf{x}_{\text{pb}}^{(p)}} f_{\text{pb}}^{(p)}$.

For each particle:

- Apply swarm intelligence move to obtain $\mathbf{x}_k^{(p)}$ using

$$\mathbf{x}_k^{(p)} = \mathbf{x}_{k|k-1}^{(p)} + \chi\big[\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\text{pb}}^{(p)} - \mathbf{x}_{k|k-1}^{(p)}) + \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\text{gb}} - \mathbf{x}_{k|k-1}^{(p)})\big]$$

- Evaluate the fitness for $\mathbf{x}_k^{(p)}$ using $\mathcal{F}_k(\mathbf{x}_k^{(p)})$, compare with $f_{\text{pb}}^{(p)}$ and update $\mathbf{x}_{\text{pb}}^{(p)}$ and $f_{\text{pb}}^{(p)}$ using (5.32) and (5.33).
- Assign $\mathbf{x}_k^{(p)}$ by the best-fit particle using $\mathbf{x}_k^{(p)} \Leftarrow \mathbf{x}_{\text{pb}}^{(p)}$.

End

Output: $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}_{p=1}^{N_p}$.

Note that the above can be viewed as a hierarchical sampling for achieving $\mathbf{x}_k^{(p)} \sim p^{(\text{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ where $p^{(\text{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k) \approx p(\mathbf{x}_k|\mathbf{z}_{1:k})$. That is, given $\mathbf{x}_{k-1}^{(p)}$ for approximating $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$, sampling at the current frame has been achieved via prediction using $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ in (5.26), followed by an update step using the latest measurement $\mathbf{z}_k$ when evaluating $\mathcal{F}_k(\cdot)$ in swarm intelligence described in (5.28)-(5.34).

D) Weight Update

Similar to conventional PF, computation of $w_k^{(p)}$ is given by

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)}, \tag{5.35}$$

where $p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})$ can be computed by substituting $\mathbf{x}_k^{(p)}$ in $\mathcal{X}_k^{(p)}$ and $\mathbf{x}_{k-1}^{(p)}$ in $\mathcal{X}_{k-1}^{(p)}$ into (5.17). Furthermore, given the fitness definition in (5.25) and the availability of $f_{\mathrm{pb}}^{(p)}$, $p(\mathbf{z}_k|\mathbf{x}_k^{(p)}) = \mathcal{M}^{-1}(f_{\mathrm{pb}}^{(p)})$ where $\mathcal{M}^{-1}(\cdot)$ denotes the inverse function of $\mathcal{M}(\cdot)$ defined in (5.25).

For the computation of $p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$, since particles have been well-sampled as $\{\mathbf{x}_k^{(p)}\}_{p=1}^{N_p}$ by swarm intelligence, $p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ can be estimated from these sampled particles using kernel density estimation (KDE) [71, 72] as

$$p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k) = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathcal{K}_{\mathbf{H}}(\mathbf{x}_k^{(p)} - \mathbf{x}_k^{(i)}), \tag{5.36}$$

where $\mathcal{K}_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2}\mathcal{K}(\mathbf{H}^{-1/2}\mathbf{x})$ is a scaled kernel density function specified by the bandwidth matrix $\mathbf{H}$ while $\mathcal{K}(\cdot)$ is the basic kernel density function. The choice of $\mathcal{K}(\cdot)$ is not crucial to the accuracy of density estimation and $\mathcal{K}(\cdot) = \mathcal{N}(\cdot; \mathbf{0}, \mathbf{I})$ is commonly used [71]. In addition, $\mathbf{H}$ controls the the amount of smoothing in the estimated pdf. Here, the Silverman's rule of thumb [73] is considered, $\mathbf{H} = (N_p)^{-1/3}\mathrm{diag}([\sigma_x^2,\ \sigma_y^2])$, where σ_x and σ_y denote the variance of particles in x and y directions, respectively.

Table 5.2 summarizes the proposed SWIPF algorithm. Given the estimated $\{\mathcal{X}_k^{(p)},\ w_k^{(p)}\}_{p=1}^{N_p}$, the state estimate at the kth frame is then given by

$$\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \mathbf{x}_k^{(p)}. \tag{5.37}$$

Table 5.2: Summary of the SWIPF algorithm.

For all particles, initialize $\mathcal{X}_0^{(p)} = \{\mathbf{x}_0^{(p)},\ \mathbf{x}_{\text{pb}}^{(p)},\ f_{\text{pb}}^{(p)}\}$ where $\mathbf{x}_0^{(p)} \sim \mathcal{N}(\widehat{\mathbf{x}}_0, \mathbf{I}_{2\times 2})$, $\widehat{\mathbf{x}}_0$ is initial state, $\mathbf{I}_{2\times 2}$ is identity matrix, $\mathbf{x}_{\text{pb}}^{(p)} = \mathbf{x}_0^{(p)}$ and $f_{\text{pb}}^{(p)} = -\infty$. The weight is initialized as $w_k^{(p)} = 1/N_p$.

For the kth frame:
Input: $\{\mathcal{X}_{k-1}^{(p)},\ w_{k-1}^{(p)}\}_{p=1}^{N_p}$

1. *Particle sampling*: obtain optimally sampled $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ using Table. 5.1.

2. *IS density estimation*: compute $p^{(\text{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ using kernel density estimation as described in (5.36), i.e.,

$$p^{(\text{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k) = \frac{1}{N_p}\sum_{i=1}^{N_p} \mathcal{K}_{\mathbf{H}}(\mathbf{x}_k^{(p)} - \mathbf{x}_k^{(i)}).$$

3. *Weight update*: compute $p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})$ using (5.17), obtain $p(\mathbf{z}_k|\mathbf{x}_k^{(p)}) = \mathcal{M}^{-1}(f_{\text{pb}}^{(p)})$ and then update $w_k^{(p)}$ using

$$w_k^{(p)} \propto\ w_{k-1}^{(p)}\ \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\text{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)},$$

   followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)}(\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.

4. *Resampling*: resample $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ if the effective sample size is below a threshold, i.e., $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$.

Output: $\{\mathcal{X}_k^{(p)},\ w_k^{(p)}\}_{p=1}^{N_p}$
End

The proposed SWIPF incorporates the resampling step similar to conventional SIRPF. When the effective sample size is below a threshold, i.e., $N_{\text{eff}} < N_{\text{thr}}$, $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ is resampled. This step again improves the particle sampling since any ineffective particles will be replaced by other effective particles. While mathematical analysis to determine the biasness of the state estimate is beyond the scope of this chapter, Monte Carlo simulations suggest that the state estimate generally exhibits low error as will be shown in Sec. 5.6.

5.5.2 Application on Alternating Source Tracking

To apply the proposed SWIPF for the alternating source tracking scenario, the monotonic function $\mathcal{M}(\cdot)$ in (5.25) is to be defined. Since each $\widehat{\tau}_k^{(i,j)}$ in $\mathbf{z}_k$ spanning over a small TDOA range (in ms), the density value of $p(\mathbf{z}_k|\mathbf{x})$ in (5.23) potentially has a high permissible range. Therefore, a logarithmic function is used as $\mathcal{M}(\cdot)$ to achieve a lower permissible range. The fitness function in (5.25) can therefore be rewritten as

$$\mathcal{F}_k(\mathbf{x}) \triangleq \ln\{p(\mathbf{z}_k|\mathbf{x})\}, \quad \mathbf{x} \in \mathcal{D}. \tag{5.38}$$

In addition, for the considered talker tracking application, a fitness decay amount $\Delta f = \alpha(f_{\max} - f_{\mathrm{v}})$ is applied for (5.27) such that

$$f_{\mathrm{pb}}^{(p)} \Leftarrow f_{\mathrm{pb}}^{(p)} - \alpha(f_{\max} - f_{\mathrm{v}}), \tag{5.39}$$

where $f_{\max}$ is the maximum of the fitness values evaluated during an initial calibration period in which the desired source is active, f_{v} is the average fitness value during an initial period in which the desired source is silent, and α is a control parameter that regulates the decay speed. Therefore, for an illustrative case where $\alpha = 0.2$, the fitness value can reduce from $f_{\max}$ to f_{v} within five iterations. The value of α needs to be empirically determined according to the speed of the moving source — a lower value of α indicates a slower memory decay.

The behavior of the proposed SWIPF for the alternating source tracking can be described as follows. The particles are first propagated using $\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)})$ where the alternating-source model $\mathcal{G}(\cdot)$ defined in (5.13) and (5.15) is used. In practice, a random sample r is drawn from a uniform distribution and if $r > P_{\mathrm{alt}}$, indicating that alternation does not occur, $\mathbf{x}_{k|k-1}^{(p)}$ is obtained by using (5.13). Otherwise, if $r \leq P_{\mathrm{alt}}$, $\mathbf{x}_{k|k-1}^{(p)}$ is obtained from (5.15). Figure 5.7 illustrates the particle

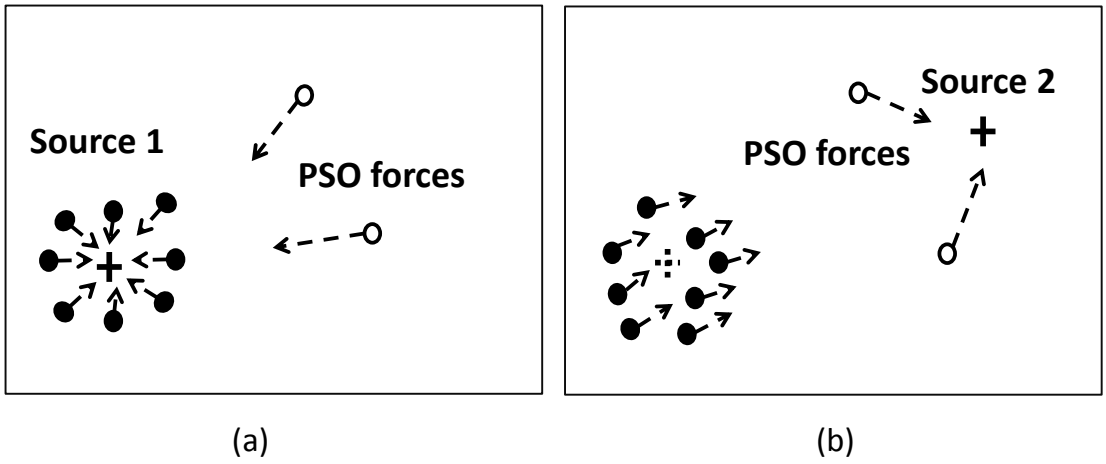


Figure 5.7: The proposed swarm intelligence based particle sampling. (a) When the first talker is active, both drifted particles and newly initiated particles are driven by the PSO force. (b) When the second talker is active, the newly initiated particles become the early members to detect the new active source.

propagation process based on the proposed source-dynamic model. From a particle sampling point of view, P_{alt} regulates the division of particles. Therefore, if $P_{\text{alt}} = 0.2$, 80% of the particles (according to (5.13) and shown by the solid dots) will be perturbed within the neighborhood of its previous estimated source position. The remaining 20% of the particles (according to (5.15) and shown by the hollow circles) will be re-initialized to facilitate the detection of any new active source. As opposed to the use of single-source motion model, this propagation step increases the likelihood of detecting any new active source while preserving the capability of tracking an existing source.

Although the use of alternating-source model $\mathcal{G}(\cdot)$ improves particle prediction, model mismatch may still occur. Firstly, an alternation is assumed with probability P_{alt} for every frame and this assumption may not be valid when the source is continuously active. As shown in Fig. 5.7 (a), some of the particles $\mathbf{x}_{k|k-1}^{(p)}$ (denoted by the hollow circle) are uniformly re-initiated resulting in them being far away from an existing source. In this case, the memory mechanism will be enabled to compensate the model mismatch effect: since the wrongly predicted $\mathbf{x}_{k|k-1}^{(p)}$ is not in line with the location information obtained from $\mathbf{z}_k$, $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\text{pb}}^{(p)}$ and the

historical information $\mathbf{x}_{\text{pb}}^{(p)}$ and $f_{\text{pb}}^{(p)}$ (from (5.28) and (5.29)) will be preserved. The swarm intelligence will then move the wrongly predicted $\mathbf{x}_{k|k-1}^{(p)}$ back to its previous best-fit positions in (5.30). Otherwise, if alternation has occurred, the uniformly re-initiated particles (shown by the hollow circle in Fig. 5.7 (b)) become the early members to detect the new source. For these particles, $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \geq f_{\text{pb}}^{(p)}$ and the historical information will be updated according to (5.28) and (5.29). The interaction mechanism will share the information among all the particles using $\mathbf{x}_{\text{gb}}$ and converge the particles to the global best-fit position given by (5.30).

Finally, for frames with low SIR, SNR and/or SRR, $p(\mathbf{z}_k|\mathbf{x})$ and hence $\mathcal{F}_k(\mathbf{x})$ will not provide accurate source-location information for particles to converge. Knowing that the current erroneous $\mathbf{z}_k$ will cause the newly evaluated fitness $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)})$ to be lower than the previous best-fit fitness derived from the reliable measurement, i.e., $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\text{pb}}^{(p)}$, the memory mechanism will also be exploited and the particles will remain at their previous best-fit state estimates.

5.5.3 Algorithmic Complexity

This section compares the computational complexity of the proposed SWIPF with that of SIRPF [10, 28] and EKPF [12]. Since the dimension of state vector is usually fixed for talker tracking application (e.g., $\mathbf{x}_k = [x_k, y_k]$) while the dimension of measurement vector $\mathbf{z}_k = [\tau_k^{(1,2)}, \ldots, \tau_k^{(N-1,N)}]$ varies depending on the number of microphone pairs $M = \mathcal{C}(\mathbf{z}_k)$ to be used, where $\mathcal{C}(\cdot)$ denotes cardinality, the complexity is therefore estimated with respect to M.

Given N_p particles, the algorithmic complexity per iteration of the proposed SWIPF algorithm is $\mathcal{O}(2N_pM)$. This results from fact that evaluation of the fitness function $\mathcal{F}_k(\cdot)$, i.e., $p(\mathbf{z}_k|\mathbf{x}_k)$, involves complexity $\mathcal{O}(M)$ as in (5.23), and such fitness function needs to be evaluated twice (in (5.28)/(5.29) and (5.32)/(5.33), respectively)

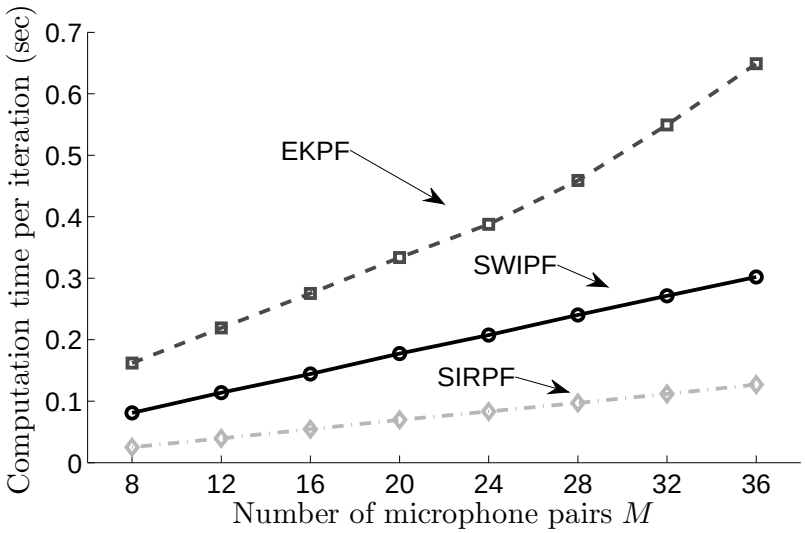


Figure 5.8: Averaged computation time per iteration versus number of microphone pairs M. The number of particles is fixed at $N_p = 100$.

for each particle per iteration. For the SIRPF algorithm, the algorithmic complexity per iteration is $\mathcal{O}(N_p M)$ because computation of $p(\mathbf{z}_k|\mathbf{x}_k)$ involves complexity of $\mathcal{O}(M)$ similar to (5.23) for each particle. The EKPF algorithm requires a complexity of $\mathcal{O}(N_p M^3)$ per iteration. This is because a matrix inverse is required for an $M \times M$ measurement covariance matrix in the Kalman filtering step, which generally requires $\mathcal{O}(M^3)$, and such step has to be repeated for every particle. In view of the above, the SIRPF requires the least computation while the EKPF algorithm requires the highest computational load as M increases. The proposed SWIPF algorithm offers a good tradeoff between computational load and performance.

Figure 5.8 shows the averaged computation time per iteration given a fixed number of particles $N_p = 100$. The computation time is evaluated on Matlab 2013b platform using a desktop PC with a quad-core processor of 3.07 GHz. As expected, the computation time of both SWIPF and SIRPF increases linearly with M. The computation time of the proposed SWIPF is less than that of EKPF. The SWIPF requires 100 ms processing time per iteration for $M = 12$, which is close to a typical frame length of 64 ms (1024 samples when $f_s = 16$ kHz) to 128 ms (2048 samples when $f_s = 16$ kHz), implying that the algorithm is suitable for real-time processing.

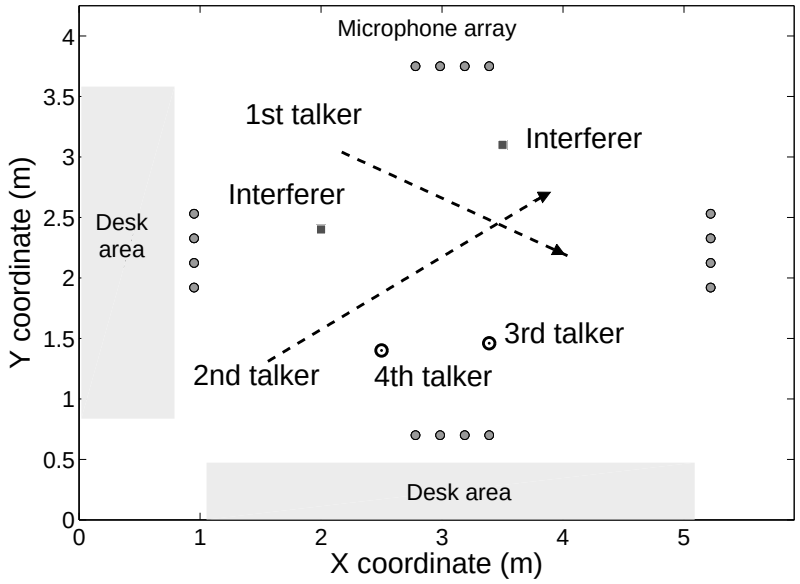


Figure 5.9: Room setup for simulation and experiment. Four microphone arrays with total of 16 microphones are employed. Four alternating talkers including two moving talkers and two stationary talkers are active in turns. A silence period is introduced after the third talker becomes silent and before the fourth talker becomes active. A fan noise interferer and a speech interferer are always present during the entire duration.

5.6 Simulation and Experiment Results

The performance of SWIPF is evaluated both in simulated environment with room impulse responses (RIRs) generated using the method of images [45,74] and an actual room environment. The proposed SWIPF algorithm is compared with SIRPF [10] and EKPF tracking algorithms [12]. For SIRPF, both the Langevin process (single-source) [10] and the proposed alternating-source model have been applied to evaluate the particle sampling impoverishment problem. These algorithms are denoted as SIRPF-single and SIRPF-alternating, respectively. In addition, TDOA measurements are considered for SIRPF as in [28] for consistency with other two algorithms. For EKPF, the IS density is estimated using the extended Kalman filter for tracking alternating talkers. Fig. 5.9 shows a 5.9 m $\times$ 4.25 m $\times$ 2.3 m room environment where the experiments were conducted. Similar to [12], four microphone arrays with total of 16 microphones are deployed and each array consists of three pairs of adjacent microphones with 0.2 m inter-microphone spacing. Four talkers (two

moving and two stationary) are used and they are activated alternatively to generate an alternating-source scenario. A fan noise and an interfering speech signal are present during the entire duration. In addition, a silence period is introduced between the third and fourth talkers' active periods, during which only interferers are present. This is to examine whether the tracking system is able to resume tracking the fourth talker after being interrupted by interferers during the long break. For moving talkers, similar to [10,12,28], the speed of the talkers was set to approximately 0.3 m/s. Since a person is expected to move considerably slower indoors compared to outdoors, the speed used was approximately a quarter that of a pedestrian's [75]. Speech signals were obtained from the TIMIT database [44] and were of duration 26 s for both simulation and experiment. All audio data are sampled at 16 kHz. For simulations, synthetic RIRs were also generated using the same room dimension and microphone-source configuration. To simulate the moving talker, RIRs were generated in a manner similar to [10, 12, 28, 41] for discrete talker positions sampled frame-wise along a pre-defined path trajectory (code available online [45, 74]). The speech signals were then convolved with these RIRs frame-wise to generate signals corresponding to moving talkers.

For the algorithms, TDOA measurements are computed using a frame size of 1024 samples and no overlapping is applied between consecutive frames. The frequency range of interest is set as $\Omega = \{2\pi \times 100 \text{ Hz} \leq \omega \leq 2\pi \times 6 \text{ kHz}\}$ corresponding to frequencies where speech components mainly exist [49]. The PF resampling threshold is set as $N_{\text{thr}} = 37.5$ [10, 28]. In addition, no prior knowledge of initial talker location is assumed for all three algorithms and the initial state $\widehat{\mathbf{x}}_0$ in Table 5.2 is initialized at the center of the room. The parameters of the proposed alternating source-dynamic model was set as $\sigma_{\mathbf{u}}^2 = 0.01$ and $P_{\text{alt}} = 0.2$. For SWIPF,

Additional simulations show that an increase in walking speed will reduce the performance of SIRPF and EKPF while the impact for the proposed SWIPF algorithm is modest due to SWIPF's high rate of convergence

$\sigma_\tau^2 = 5 \times 10^{-5}$ is used for (5.19) and the fitness decay factor was $\alpha = 0.2$. No priority is assumed for the acceleration on individual influence and social influence, i.e., $\varphi_1 = \varphi_2 = 2.1$ are applied.

The tracking performance is evaluated using

$$e_k = ||\widehat{\mathbf{x}}_k - \mathbf{x}_k^{\mathrm{s}}||, \tag{5.40}$$

where $\widehat{\mathbf{x}}_k$ is the estimated position, and $\mathbf{x}_k^{\mathrm{s}}$ is the true talker position. The average tracking error $\bar{e} = \frac{1}{K}\sum_{k=1}^{K} e_k$ quantifies the performance across all audio frames, where K is the total number of frames.

5.6.1 Simulation Results

Tracking results for a single trial with reverberation time $T_{60} = 0.5$ s, SIR = 9 dB, and SNR = 10 dB are presented in Fig. 5.10, where the SNR is computed by involving only the background diffuse noise without the interferers. Figure 5.10 (a) shows the maximum peaks in GCC function (i.e., the measurements) for one pair of microphones and the tracked TDOA trajectory that obtained from the talker position trajectory estimated by SWIPF. The solid bold line denotes the TDOA trajectory corresponding to the actual (active) talker and the two dashed horizontal lines denote the ones with stationary fan and speech interferers. Three alternation instances between the talkers are introduced at the 106th , 256th and 395th frame (after the silence break). It can be observed that the TDOA measurements (shown by the dots) correspond to the desired dominant talker for most of the frames due to his/her highest energy. The occasional outliers (e.g., during frames $250 < k < 350$) are mainly caused by the low SIR. The SWIPF algorithm tracks the desired talker and exhibits robustness to outliers by taking into account both the signal energy and historical activeness information. It can also be observed that during the silence

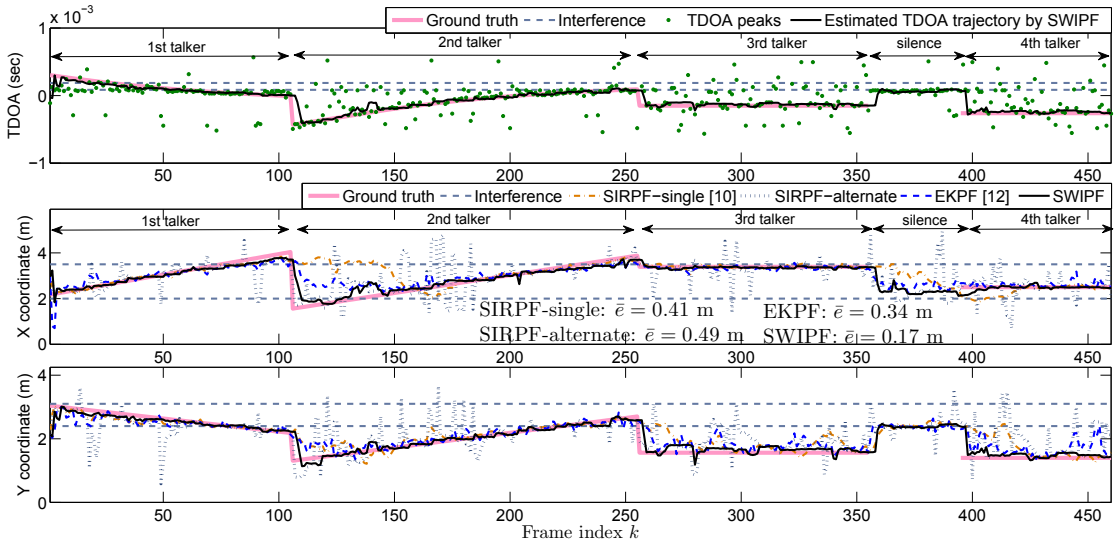


Figure 5.10: A single-trial result using the simulated data when $T_{60} = 0.5$ s, SNR $= 10$ dB and SIR $= 9$ dB for each interferer. A silence period (during frames $355 < k < 395$) is introduced during which the desired talker is absent and only two interferers are present. (a) measured TDOA versus estimated TDOA trajectory by SWIPF for one microphone pair; (b) x coordinate of the estimated trajectory; (c) y coordinate of the estimated trajectory.

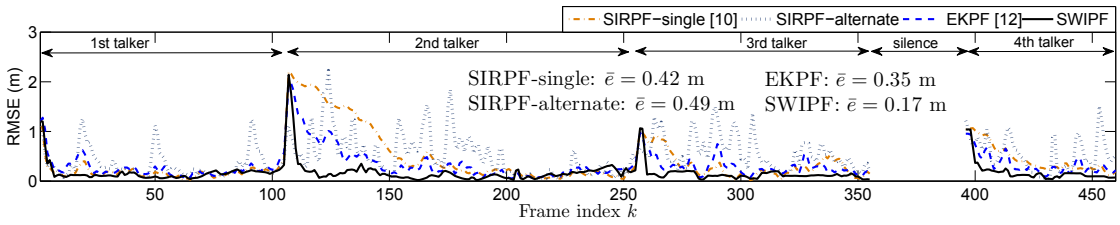


Figure 5.11: RMSE over 50 Monte Carlo trials where particle propagation and swam moves are realized differently in probabilistic approach for each trial. The environment is simulated with $T_{60} = 0.5$ s, SNR $= 10$ dB and SIR $= 9$ dB for each interferer.

period between the 355th and 395 frames, the tracking system inevitably tracks the interferer due to the absence of the desired talker. The SWIPF algorithm resumes tracking the fourth talker immediately after he/she becomes active since the corresponding measurements have been obtained. Note that to distinguish the active and silent phases of the speech source, algorithms such as described in [76, 77] can be considered such that the proposed tracking algorithm will be only enabled when the presence of speech is detected.

Figures. 5.10 (b) and (c) compare tracking results in terms of position trajectory estimates from a single trial. The SIRPF-single algorithm (dash-dotted line)

requires the longest transition time to track the new active talker after alternation is introduced. This is caused by the mismatch between the assumed single-source model and the actual alternation between talkers. When another talker is activated at a new position, the particles still remain in the neighborhood of the previous talker position. Higher number of iterations is therefore required to resample the particles into an area where the active talker is located. For the same reason, long transition time is required to resume tracking the fourth talker after being interrupted by the interference during the silence period. The SIRPF-alternating algorithm (dotted line) shows the effectiveness of the proposed alternating-source model in terms of the transition time. This algorithm, however, exhibits high error during periods where only a single talker is active (i.e., no alternation). This is because the alternating-source model assumes possibility of alternation at each iteration and this assumption is violated during periods when a single talker is continuously moving/stationary. When this occurs, particles are poorly sampled and they converge toward a non-existent "new" talker caused by interferers, noise and/or reverberation. The EKPF algorithm (dashed line) is able to strike a balance between transition time and robustness in continuous tracking period. This is because EKPF approximates a better IS density by taking into account information derived from measurements. However, particles may not be appropriately sampled in EKPF due to the incorporation of erroneous measurements in frames with low SNR, SIR and/or SRR. The resultant trajectory therefore frequently deviates from the ground truth. The proposed SWIPF algorithm (solid line) achieves the shortest transition time as well as the least fluctuation. This is because SWIPF utilizes the proposed alternating source-dynamic model and interaction mechanism to achieve higher rate of convergence to the new active talker. In addition, the memory mechanism is exploited to compensate the mismatch effect brought by the alternating-source model and disturbance from interference, noise and reverberation. The particles remain at their

best-fit positions when no alternation occurs, leading to reduced fluctuation.

Figure 5.11 shows the root-mean-square error (RMSE) over fifty Monte Carlo simulation trials where, for each trial, sampling of particles is realized differently in terms of particle propagation from the probabilistic source-dynamic model and random swarm moves. The room and source configuration setup remain the same as the previous simulation. Three high RMSE peaks indicate the discontinuity in RMSE when the alternation is introduced. The RMSE between the 355th and 395th frames is not plotted due to absence of the speech source. These multiple trials validate the performance of SWIPF compared to the other three algorithms. The SWIPF algorithm requires less than 0.64 s (10 frames with frame length of 1024 samples in 16 kHz sampling rate) on average to successfully switch to the second source. This implies that that SWIPF is expected to perform well provided that the interval between any alternation occurrences is beyond 0.64 s, compared to 2.56 and 4.48 s for the EKPF and SIRPF-single algorithms, respectively. In addition, the tracking error remains low for the period after the 256th frame, indicating that SWIPF is also suitable for stationary sources.

Figure 5.12 (a) shows how the RMSE varies with reverberation time when SNR = 15 dB and SIR = 9 dB. This result shows that the tracking error increases with reverberation time, as expected. Both SIRPF-single and SIRPF-alternating algorithms exhibit high tracking error due to the model mismatch problem. Comparatively, the EKPF algorithm achieves lower tracking error. The proposed SWIPF algorithm achieves the lowest tracking error in all the cases being considered since it utilizes memory information in frames with low SRR. Figure 5.12 (b) shows how the RMSE varies with SNR for $T_{60} = 0.3$ s and SIR = 9 dB. The RMSE increases with reducing SNR for all algorithms as expected. As before, both SIRPF algorithms suffer from the highest tracking error due to particle sampling impoverishment. The SWIPF algorithm achieves the lowest tracking error in a noisy environment.

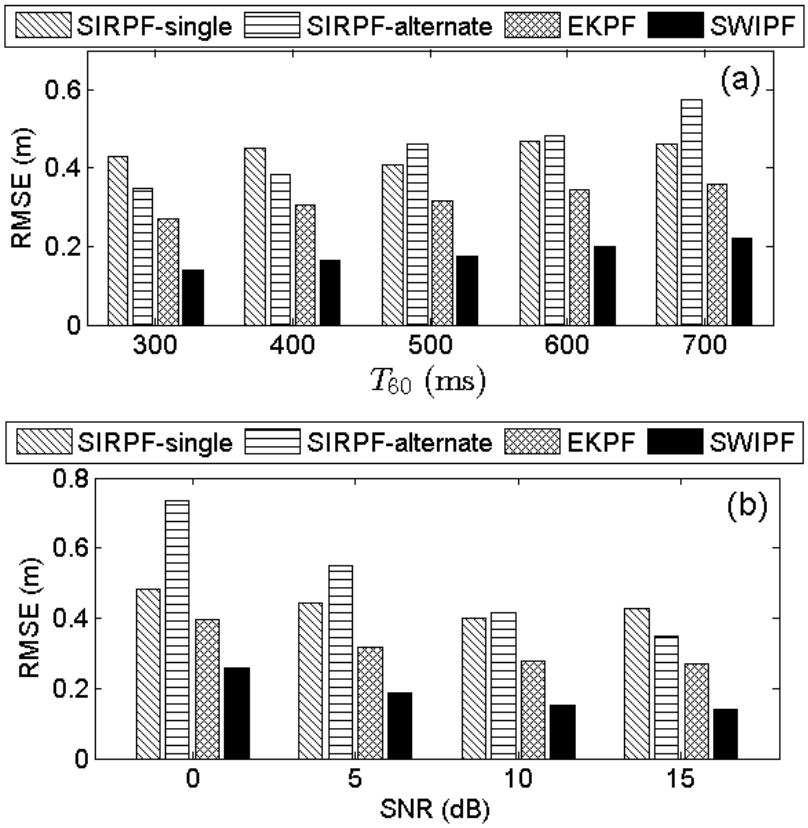


Figure 5.12: Variation of RMSE for (a) different reverberation time at SNR=15 dB and SIR = 9 dB for each interferer, and (b) different SNRs at $T_{60} = 0.3$ s and SIR = 9 dB for each interferer.

5.6.2 Experiment Results

An experiment was conducted in an actual room environment (see Fig. 5.9). Four speech signals were generated, in a manner similar to [12, 31], using loudspeakers to simulate two moving and two stationary talkers. The moving loudspeakers were faced in the direction of motion, while the stationary loudspeaker was faced towards the center of the room. Additionally, both interferers were also played through the loudspeakers towards the center of the room. For this experiment, $T_{60} \approx 0.35$ s and the background noise without any interference was estimated to be at SNR$\approx$ 7 dB. The SIR was estimated to be 12 dB and 15 dB for the interfering speech and fan noise, respectively. Parameters for the algorithms were configured similar to that described in the simulation setup.

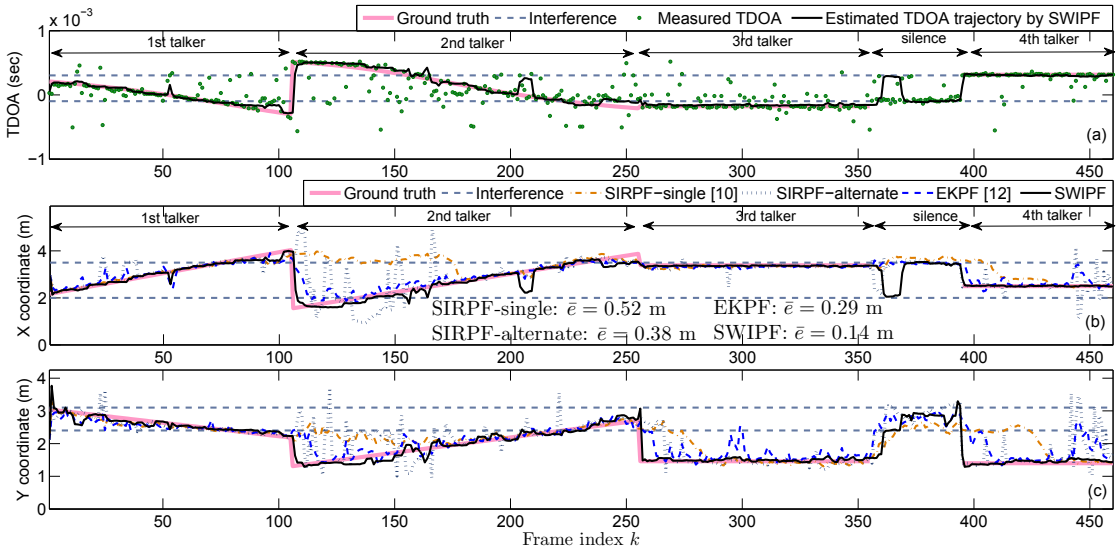


Figure 5.13: A single-trial result using recorded data from an actual environment with an estimated $T_{60} = 0.35$ s and SNR = 7 dB, SIR = 12 dB and 15 dB for the interfering speech signal and fan, respectively. A silence period (during frames $355 < k < 395$) is introduced during which the desired talker is absent and only two interferers are present. (a) measured TDOA versus estimated TDOA trajectory by SWIPF for one microphone pair; (b) x coordinate of the estimated trajectory; (c) y coordinate of the estimated trajectory.

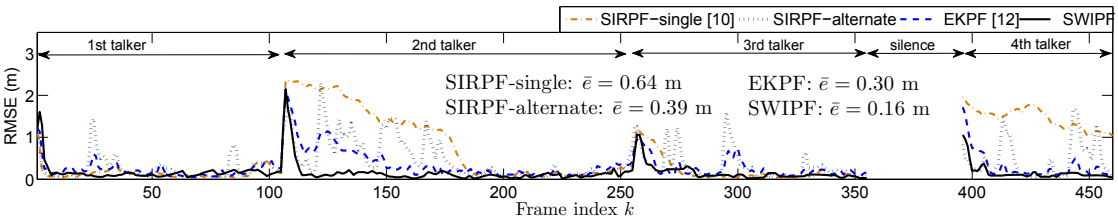


Figure 5.14: RMSE over 50 Monte Carlo trials where particle propagation and swam moves are realized differently in probabilistic approach for each trial. Signals are recorded in an actual environment with an estimated $T_{60} = 0.35$ s and SNR = 7 dB, SIR = 12 dB and 15 dB for the interfering speech signal and fan, respectively.

Figure 5.13 (a) shows the tracking result of SWIPF in terms of TDOA trajectory versus TDOA measurement. The SWIPF algorithm filters out the unreliable measurements and achieves short transition time for alternations and after the silence period. Figure 5.13 (b) and (c) show tracking results among three algorithms in terms of position trajectory estimation. Similar to that of simulations, the SIRPF-single algorithm fails to "lock on" to the new talker after alternation has occurred. While the SIRPF-alternating algorithm achieves shorter transition time, it suffers from high fluctuations during the period when the talker is continuously active.

The EKPF algorithm achieves a shorter transition time than SIRPF-single and less fluctuation than SIRPF-alternating. The proposed SWIPF algorithm achieves the highest tracking accuracy among the considered algorithms in terms of the convergence rate and tracking stability. Similar results can also be found in Monte Carlo trials as in Fig. 5.14.

5.7 Chapter Summary

An SWIPF algorithm is proposed to track alternating talkers. The proposed algorithm exploits PF and swarm intelligence jointly to achieve optimal particle sampling. As opposed to propagating the particles independently in PF, SWIPF incorporates the interaction mechanism in swarm intelligence to improve the particle convergence to the active talker region. In addition, the memory mechanism enables particles to be retained at the previous best-fit positions when signals are corrupted by interference, noise and reverberation. Simulation and experiment results show that SWIPF can locate and track the alternating talkers with short transition period, resulting in the lowest tracking error compared to EKPF and SIRPF in a noisy and reverberant environment.

Part II

DOA Estimation Using Acoustic Vector Sensor

Chapter 6

Literature Review

Part I of this thesis focuses on acoustic source localization and tracking (ASLT) using conventional microphone arrays. Such a microphone array often requires a large aperture to achieve good performance and this limits their application to space-constrained applications. In Part II of this thesis, the focus is on acoustic vector sensor (AVS) [78] which has emerged in recent years. As shown in Fig. 6.1, an AVS consists of one monopole pressure sensor element collocated with three orthogonally oriented dipole elements. Unlike the conventional array which requires spacing between microphones, a single AVS can accomplish ASLT with a compact configuration in which the spacing between sensor elements is negligible. In the remainder of this thesis, the problem of direction-of-arrival (DOA) estimation using a single AVS is considered.

The organization of this chapter is as follows: in Section 6.1, signal models of an AVS are discussed. The short-time Fourier transform representation of the received signal is shown in Section 6.2. In Section 6.3, the state-of-the-art DOA estimation algorithms for AVS will be reviewed. Finally, Section 6.4 concludes the chapter.

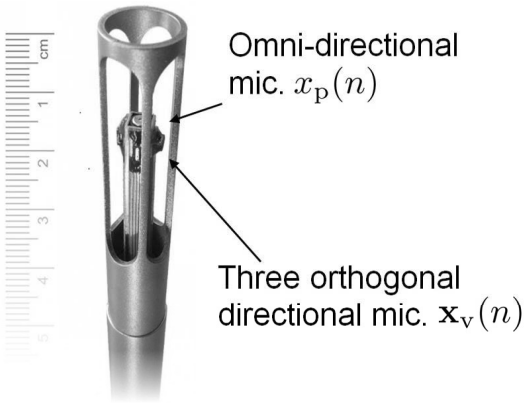


Figure 6.1: An example of acoustic vector-sensor which consists of an omni-directional microphone and three orthogonal directional microphones in x, y and z directions, respectively. (after [79])

6.1 Received signal model of an AVS

6.1.1 Single-source free-space model

As illustrated in Fig. 6.2, consider an AVS consisting of one monopole and three orthogonal dipole elements co-located at the origin in a free space. A source signal $s(t)$ emanates from a particular direction $[\phi^{\mathrm{src}},\ \psi^{\mathrm{src}}]^{\mathsf{T}}$ where ϕ^{src} and ψ^{src} denotes the azimuth and elevation incident angles, respectively. Omitting any attenuation due to signal propagation, the waveform at the sensor position can be modeled as a time-delayed version of the original signal $s(t-\Delta t)$, where Δt is the direct-path propagation time from the source to the sensor. The omni-directional pressure sensor-element is assumed to have a unity gain for the incident signal which is independent of the source direction, while the dipole sensor-elements have incident-angle dependent gains for spatial filtering. The received signal $\mathbf{y}(t)$ can be written

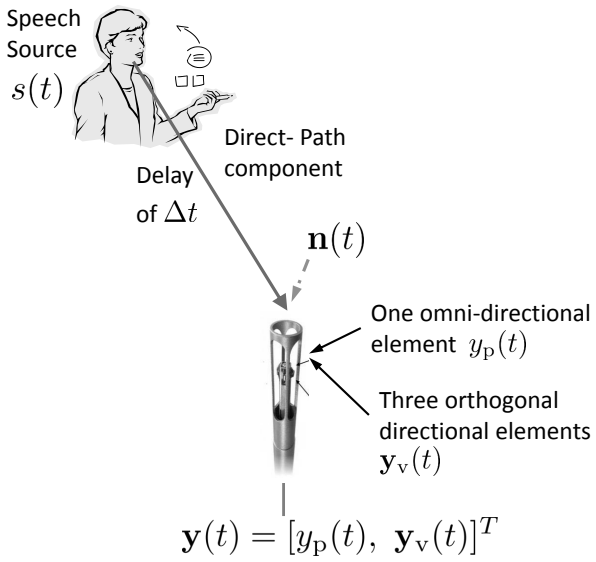


Figure 6.2: Single-source reverberant model for single acoustic vector sensor.

as [78]

$$\begin{aligned}\mathbf{y}(t) &= \begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} \\ &= s(t-\Delta t) \begin{bmatrix} 1 \\ \mathbf{u}^{\mathrm{src}} \end{bmatrix} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \end{aligned} \tag{6.1}$$

where $y_{\mathrm{p}}(t)$ and $\mathbf{y}_{\mathrm{v}}(t)$ are the monopole element and dipole element outputs, respectively, and $\mathbf{u}^{\mathrm{src}} = [\cos\psi^{\mathrm{src}}\cos\phi^{\mathrm{src}},\ \cos\psi^{\mathrm{src}}\sin\phi^{\mathrm{src}},\ \sin\psi^{\mathrm{src}}]^{\mathsf{T}}$ defines the manifold of the dipole elements pointing towards the source. The variables $n_{\mathrm{p}}(t)$ and $\mathbf{n}_{\mathrm{v}}(t)$ are defined, respectively, as the additive noise at the monopole sensor-element and dipole sensor-elements. Equation (6.1) is hereby referred as *single-source free-space model* for the AVS. The aim is to estimate $\mathbf{u}^{\mathrm{src}}$ given received signals $\mathbf{y}(t)$ at the sensor-elements.

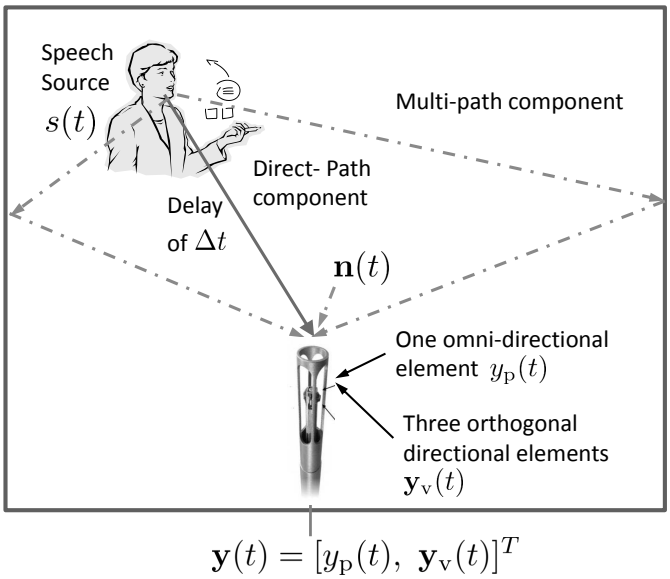


Figure 6.3: Single-source reverberant model for a single acoustic vector sensor.

6.1.2 Single-source reverberant model

Next consider a more realistic scenario where the source and AVS are located in an enclosed environment, as shown in Fig. 6.3. An AVS is located at the origin and a source signal $s(t)$ emanates from a particular direction $[\phi^{\mathrm{src}},\ \psi^{\mathrm{src}}]^{\mathsf{T}}$ with respect to the sensor. In this scenario, the AVS captures not only the direct-path component but also reflected signals from the room boundaries. To formulate both the direct-path component and any reflections in this case, propagation from the source position to the AVS including both direct-path and reflections can be approximated as a linear time-invariant system. Defining $h_{\mathrm{p}}(t)$ as the impulse response from the source to the monopole pressure element, $\mathbf{h}_{\mathrm{v}}(t)$ as a 3×1 impulse response sample vector from the source to the dipole elements, and $*$ as the convolution operator,

the received signal $\mathbf{y}(t) = [y_{\mathrm{p}}(t),\ \mathbf{y}_{\mathrm{v}}^{\mathsf{T}}(t)]^{\mathsf{T}}$ can be written as [80]

$$\begin{aligned}\mathbf{y}(t) &= \begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} \\ &= s(t) * \begin{bmatrix} h_{\mathrm{p}}(t) \\ \mathbf{h}_{\mathrm{v}}(t) \end{bmatrix} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \end{aligned} \tag{6.2}$$

where $n_{\mathrm{p}}(t)$ and $\mathbf{n}_{\mathrm{v}}(t)$ are defined as the noise signals.

Furthermore, the impulse response $[h_{\mathrm{p}}(t),\ \mathbf{h}_{\mathrm{v}}^{\mathsf{T}}(t)]^{\mathsf{T}}$ can be decomposed into the direct-path and reflection components. Equation (6.2) can be rewritten as

$$\begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} = s(t-\Delta t) \begin{bmatrix} 1 \\ \mathbf{u}^{\mathrm{src}} \end{bmatrix} + s(t) * \begin{bmatrix} h'_{\mathrm{p}}(t) \\ \mathbf{h}'_{\mathrm{v}}(t) \end{bmatrix} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \tag{6.3}$$

where the first term in the summation corresponds the direct-path propagation and the second term corresponds to the reflections. The variables $h'_{\mathrm{p}}(t)$ and $\mathbf{h}'_{\mathrm{v}}(t)$ are the remaining impulse responses without any direct-path component. Note that the direct-path propagated signal has an incident angle of $[\phi^{\mathrm{src}},\ \psi^{\mathrm{src}}]^{\mathsf{T}}$ indicated by $\mathbf{u}^{\mathrm{src}}$, while the reflected components, in general, have different incident angles as opposed to the direct-path component. It can be observed that in a reverberant environment, DOA estimation is much more challenging due to the interference caused by reflections.

6.1.3 Multi-source free-space model

For a multiple-source free-space scenario, by defining $s_l(t)$ as the lth source signal and L as the total number of sources, (6.1) can be extended as [78]

$$\begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} = \sum_{l=1}^{L} \left\{ s_l(t-\Delta t_l) \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix} \right\} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \tag{6.4}$$

where $\mathbf{u}_l^{\mathrm{src}} = [\cos\psi_l^{\mathrm{src}} \cos\phi_l^{\mathrm{src}},\ \cos\psi_l^{\mathrm{src}} \sin\phi_l^{\mathrm{src}},\ \sin\psi_l^{\mathrm{src}}]^{\mathsf{T}}$ defines the sensor manifold pointing towards the lth source, Δt_l is the propagation delay from the lth source to the sensor, and $n_{\mathrm{p}}(t)$ and $\mathbf{n}_{\mathrm{v}}(t)$ are defined as the noise signals. The aim is therefore to estimate $\mathbf{u}_l^{\mathrm{src}}$ for L number of sources. It can be observed that overlapping between source signals presents challenges for multi-source DOA estimation.

6.1.4 Multi-source reverberant model

Finally, consider the scenario where the multiple sources are present in an enclosed environment. By defining $s_l(t)$ as the lth source signal, $h_{\mathrm{p},l}(t)$ and $\mathbf{h}_{\mathrm{v},l}(t)$ as the corresponding impulse responses, and L as the total number of sources, the single-source reverberant model in (6.2) can be extended to a multi-source case as

$$\begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} = \sum_{l=1}^{L} \left\{ s_l(t) * \begin{bmatrix} h_{\mathrm{p},l}(t) \\ \mathbf{h}_{\mathrm{v},l}(t) \end{bmatrix} \right\} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix} \tag{6.5}$$

in which the received signals are the summation of the filtered signals plus noise signals. Similarly, by decomposing the impulse response to direct-path and reflection components, (6.2) can be extended by

$$\begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} = \sum_{l=1}^{L} \left\{ s_l(t-\Delta t_l) \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix} + s_l(t) * \begin{bmatrix} h'_{\mathrm{p},l}(t) \\ \mathbf{h}'_{\mathrm{v},l}(t) \end{bmatrix} \right\} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \tag{6.6}$$

where $\mathbf{u}_l^{\text{src}} = [\cos\psi_l^{\text{src}}\cos\phi_l^{\text{src}},\ \cos\psi_l^{\text{src}}\sin\phi_l^{\text{src}},\ \sin\psi_l^{\text{src}}]^{\mathsf{T}}$ defines the sensor manifold pointing towards the lth source, Δt_l is the propagation delay from the lth source to the sensor, and $h'_{\text{p},l}(t)$ and $\mathbf{h}'_{\text{v},l}(t)$ are the remaining impulse responses for the lth source.

6.2 STFT representation of the received signals

Before reviewing the existing DOA estimation algorithms, the short-time Fourier transform (STFT) of the received signals is first discussed for the two simple free-space models in (6.1) and (6.4), which will then be used in the subsequent sections. Defining T as the frame length, the kth frame of the received signal $\mathbf{y}(t)$ can be denoted as

$$\mathcal{Y}(k) = [\mathbf{y}(kT),\ \mathbf{y}(kT+1),\ \ldots,\ \mathbf{y}(kT+T-1)]. \tag{6.7}$$

The STFT transform of the received signal $\mathsf{STFT}(\mathcal{Y}(k))$ for the single-source free-space model in (6.1) can be expressed as

$$\begin{bmatrix} \underline{y}_{\text{p}}(\omega,k) \\ \underline{\mathbf{y}}_{\text{v}}(\omega,k) \end{bmatrix} = \underline{s}(\omega,k)e^{-\jmath\omega\Delta t}\begin{bmatrix} 1 \\ \mathbf{u}^{\text{src}} \end{bmatrix} + \begin{bmatrix} \underline{n}_{\text{p}}(\omega,k) \\ \underline{\mathbf{n}}_{\text{v}}(\omega,k) \end{bmatrix}, \tag{6.8}$$

where $\underline{s}(\omega,k)$, $\underline{n}_{\text{p}}(\omega,k)$ and $\underline{\mathbf{n}}_{\text{v}}(\omega,k)$ are the STFT coefficients of the source signal and noise, ω denotes the angular frequency and k denotes frame index. Similarly, for the multi-source free-space model in (6.4), the STFT transform of the received signal is expressed as

$$\begin{bmatrix} \underline{y}_{\text{p}}(\omega,k) \\ \underline{\mathbf{y}}_{\text{v}}(\omega,k) \end{bmatrix} = \sum_{l=1}^{L}\left\{\underline{s}_l(\omega,k)e^{-\jmath\omega\Delta t_l}\begin{bmatrix} 1 \\ \mathbf{u}_l^{\text{src}} \end{bmatrix}\right\} + \begin{bmatrix} \underline{n}_{\text{p}}(\omega,k) \\ \underline{\mathbf{n}}_{\text{v}}(\omega,k) \end{bmatrix}, \tag{6.9}$$

where $\underline{s}_l(\omega,k)$ is the STFT coefficients of the lth source signal.

6.3 Existing DOA estimation algorithms

For DOA estimation using an AVS, an initial work [78] derives the Cramér-Rao lower bound (CRLB) for a free-space scenario where Gaussian additive noise is considered. Two DOA estimators, namely, the intensity and velocity-covariance based estimators were proposed for the use of a single AVS. In [81], a steered-response-power beamformer (SRP-beamforer) estimator was proposed, which can be regarded as a generalized version of the intensity and velocity-covariance based algorithms with a balance parameter between each other. However, it was unclear that which of the aforementioned algorithms with what parameters can achieve the CRLB. In [82], as a specific realization of the SRP algorithm, a maximum likelihood estimator was proposed in which the optimal parameter to attain the CRLB can be obtained by assuming the noise follows Gaussian distribution with known statistics. In [80], the effect of reverberation was examined for the intensity-based estimator. By deriving a statistical model, this work suggests that in the presence of reverberation, the DOA estimate can be biased. Apart from single-source DOA estimation, multi-source DOA estimation can be achieved by extending the use of the well-known multiple signal classification (MUSIC) algorithm [50] for conventional microphone array to AVS. Furthermore, a single-source point based algorithm has emerged recently which exploits the sparsity between source spectra [83]. In addition to the use of a single AVS, an array of AVSs has also been employed and investigated, in order to improve the DOA estimation performance and these methods are mainly based on beamforming and subspace techniques [84–88]. In [89], quaternion based approach has been proposed, which is found to be able to achieve a more accurate subspace decomposition for DOA estimation. In [90–92], sound source tracking algorithms using Bayesian filters have been developed for AVS.

In the following, details of several benchmarking algorithms for both single-

source and multiple-source scenarios will be discussed. All of the following state-of-the-art algorithms are derived from the free-space models in (6.1) and (6.4). To derive the algorithms, the signal $s(t)$ and the noise $n_{\mathrm{p}}(t)$, $\mathbf{n}_{\mathrm{v}}(t)$ are assumed to be independent identically distributed zero-mean Gaussian process. Furthermore, $s(t)$, $n_{\mathrm{p}}(t)$ and $\mathbf{n}_{\mathrm{v}}(t)$ are assumed to be mutually uncorrelated. The above property can be described as

$$\mathbb{E}\left\{\begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}\right\} = \mathbf{0}_{4\times 1} \tag{6.10}$$

and

$$\mathbb{E}\left\{\begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}\begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}^{\mathsf{T}}\right\} = \begin{bmatrix} \sigma_{\mathrm{p}}^2 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{3\times 1} & \sigma_{\mathrm{v}}^2\mathbf{I}_{3\times 3} \end{bmatrix}, \tag{6.11}$$

where σ_{p}^2 and σ_{v}^2 are the variances of the noise signals at pressure sensor-element and dipole sensor-elements, respectively. The matrices $\mathbf{0}$ and $\mathbf{I}$ are, respectively, the zero-element matrix and identity matrix.

6.3.1 Single-source DOA estimation algorithms

Intensity based DOA estimation [78]

The intensity based DOA estimation algorithm considers the single-source free-space model in (6.1). By considering (6.1) and (6.11), the following relationship

$$\begin{aligned} \mathbf{r}_{\mathrm{pv}} &= \mathbb{E}\{y_{\mathrm{p}}(t)\mathbf{y}_{\mathrm{v}}(t)\} \\ &= \sigma_{\mathrm{s}}^2\mathbf{u}^{\mathrm{src}} \end{aligned} \tag{6.12}$$

can be obtained, where σ_{s}^2 is defined as the variance of the source signal. In practice, $\mathbf{r}_{\mathrm{pv}}$ can be estimated by time averaging the noisy signal snapshots, given by

$$\widehat{\mathbf{r}}_{\mathrm{pv}} = \frac{1}{T_{\mathrm{snap}}} \sum_{t=1}^{T_{\mathrm{snap}}} y_{\mathrm{p}}(t)\mathbf{y}_{\mathrm{v}}(t), \tag{6.13}$$

where T_{snap} is the number of snapshots considered. Hence, the estimate of $\mathbf{u}^{\mathrm{src}}$ can be computed by

$$\widehat{\mathbf{u}}^{\mathrm{src}} = \frac{\widehat{\mathbf{r}}_{\mathrm{pv}}}{\|\widehat{\mathbf{r}}_{\mathrm{pv}}\|}. \tag{6.14}$$

The statistical analysis of the intensity based estimator in (6.14) can be found in [78]. It has been shown that $\widehat{\mathbf{u}}^{\mathrm{src}} \to \mathbf{u}^{\mathrm{src}}$ as $T_{\mathrm{snap}} \to \infty$ when the noise is zero-mean Gaussian distributed. The statistical analysis of (6.14) in the presence of reverberation has also been investigated in [80], and it was found the estimator is biased when reverberation exists.

Velocity-covariance based DOA estimation [78]

The velocity-covariance based DOA estimation algorithm uses only the dipole sensor elements. With reference to (6.1) and (6.11), the covariance of the received signals at dipole sensor elements can be derived as

$$\begin{aligned} \mathbf{R}_{\mathrm{vv}} &= \mathbb{E}\{\mathbf{y}_{\mathrm{v}}(t)\mathbf{y}_{\mathrm{v}}^{\mathsf{T}}(t)\} \\ &= \sigma_{\mathrm{s}}^2 \mathbf{u}^{\mathrm{src}}(\mathbf{u}^{\mathrm{src}})^{\mathsf{T}} + \sigma_{\mathrm{v}}^2 \mathbf{I}_{3\times 3}, \end{aligned} \tag{6.15}$$

where σ_{s}^2 is defined as the variance of the source signal and, in general, $\sigma_{\mathrm{s}}^2 > \sigma_{\mathrm{v}}^2$. In practice, the covariance $\mathbf{R}_{\mathrm{vv}}$ can be computed by time-averaging of sensor snapshots given by

$$\widehat{\mathbf{R}}_{\mathrm{vv}} = \frac{1}{T_{\mathrm{snap}}} \sum_{t=1}^{T_{\mathrm{snap}}} \mathbf{y}_{\mathrm{v}}(t)\mathbf{y}_{\mathrm{v}}^{\mathsf{T}}(t), \tag{6.16}$$

where T_{snap} denotes the number of snapshots. The vector $\mathbf{u}^{\text{src}}$ can be estimated by searching for the unit eigenvector of $\widehat{\mathbf{R}}_{\text{vv}}$ that is associated with its largest eigenvalue with the sign chosen such that the corresponding eigenvalue is positive.

SRP beamformer based DOA estimation [81]

Given the steering vector defined as $\mathbf{u} = [\cos\psi\cos\phi,\ \cos\psi\sin\phi,\ \sin\psi]^{\mathsf{T}}$, the SRP beamformer computes the response signal for a steered direction as a weighted sum of monopole and dipole elements, given by

$$y_{\text{SRP}}(t) = \alpha y_{\text{p}}(t) + (1-\alpha)\mathbf{u}^{\mathsf{T}}\mathbf{y}_{\text{v}}(t), \tag{6.17}$$

where $\alpha \in [0,\ 1]$ is the weight parameter. Similar to conventional beamforming approaches, $\mathbf{u}^{\text{src}}$ can be estimated by searching for a unity steering vector that maximizes the power of the response signal via

$$\widehat{\mathbf{u}}^{\text{src}} = \arg\max_{\mathbf{u}} \left\{ \frac{1}{T_{\text{snap}}} \sum_{t=1}^{T_{\text{snap}}} y_{\text{SRP}}(t)^2 \right\}, \quad \text{s.t.}\ \ \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1. \tag{6.18}$$

Substituting (6.17) into (6.18), the target function in (6.18) can be simplified by removing some fixed energy term as

$$\mathcal{J}_{\text{SRP}}(\mathbf{u}) = \alpha\mathbf{u}^{\mathsf{T}}\widehat{\mathbf{r}}_{\text{pv}} + (1-\alpha)\frac{1}{2}\mathbf{u}^{\mathsf{T}}\widehat{\mathbf{R}}_{\text{vv}}\mathbf{u}, \tag{6.19}$$

where $\widehat{\mathbf{r}}_{\text{pv}}$ and $\widehat{\mathbf{R}}_{\text{vv}}$ have been defined in (6.13) and (6.16), respectively, and the source direction can be estimated by

$$\widehat{\mathbf{u}}^{\text{src}} = \arg\max_{\mathbf{u}} \left\{ \mathcal{J}_{\text{SRP}}(\mathbf{u}) \right\}, \quad \text{s.t.}\ \ \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1. \tag{6.20}$$

It can be observed that in (6.19), when $\alpha \to 1$, the estimate is equivalent to the intensity based estimator as in (6.14). When $\alpha \to 0$, the SRP beamformer estimator is equivalent to the eigenvector search as in the velocity-covariance based estimator described by (6.16). Therefore, the SRP beamformer estimator can be considered as a generalization for the intensity and velocity-covariance based estimators. In addition, an iterative gradient based approach for computing the maximization in (6.20) has been proposed in [81]. However, although the SRP beamformer estimator has been proposed, the value of $\alpha \in [0,\ 1]$ for the SRP-beamformer to achieve the optimal solution is not known.

Maximum likelihood based DOA estimation [82]

The maximum likelihood based DOA estimator relied on the statistical assumption made in (6.10) and (6.11). By assuming that the source signal is zero-mean Gaussian distributed and independent with the the noise signal, it can be shown that the received signal $\mathbf{y}(t) = [y_{\mathrm{p}}(t),\ \mathbf{y}_{\mathrm{v}}^{\mathsf{T}}(t)]^{\mathsf{T}}$ is zero-mean Gaussian distributed with the covariance given by

$$\begin{aligned}\mathbf{R} &= \mathbb{E}\{\mathbf{y}(t)\mathbf{y}^{\mathsf{T}}(t)\} \\ &= \sigma_{\mathrm{s}}^2 \begin{bmatrix} 1 \\ \mathbf{u}^{\mathrm{src}} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{u}^{\mathrm{src}} \end{bmatrix}^{\mathsf{T}} + \begin{bmatrix} \sigma_{\mathrm{p}}^2 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{3\times 1} & \sigma_{\mathrm{v}}^2\mathbf{I}_{3\times 3} \end{bmatrix}. \end{aligned} \tag{6.21}$$

Hence, given the identically independent snapshots of $\mathbf{y}(t)$ for $1 \le t \le T_{\mathrm{snap}}$, the pdf of $\mathcal{Y} = \{\mathbf{y}(t) | 1 \le t \le T_{\mathrm{snap}}\}$ is given by

$$p(\mathcal{Y}|\mathbf{u}^{\mathrm{src}}) = \frac{1}{(2\pi)^{T/2}|\mathbf{R}|^{1/2}} \sum_{t=1}^{T_{\mathrm{snap}}} \exp\left\{-\frac{1}{2}\mathbf{y}^{\mathsf{T}}(t)\mathbf{R}^{-1}\mathbf{y}(t)\right\}. \tag{6.22}$$

The maximum likelihood estimator estimates $\mathbf{u}^{\text{src}}$ by maximizing the above pdf given as

$$\widehat{\mathbf{u}}^{\text{src}} = \arg\max_{\mathbf{u}} \left\{ p(\mathcal{Y}|\mathbf{u}) \right\}, \quad \text{s.t.} \;\; \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1. \tag{6.23}$$

Substituting $\mathbf{u} = [\cos\psi\cos\phi, \;\; \cos\psi\sin\phi, \;\; \sin\psi]^{\mathsf{T}}$ into $\mathbf{R}$ in (6.21) and further in (6.22), the target function in (6.22) can be simplified as

$$\mathcal{J}_{\text{ML}}(\mathbf{u}) = \alpha_0 \mathbf{u}^{\mathsf{T}} \widehat{\mathbf{r}}_{\text{pv}} + (1-\alpha_0)\frac{1}{2}\mathbf{u}^{\mathsf{T}}\widehat{\mathbf{R}}_{\text{vv}}\mathbf{u}, \tag{6.24}$$

where $\widehat{\mathbf{r}}_{\text{pv}}$ and $\widehat{\mathbf{R}}_{\text{vv}}$ have been defined in (6.13) and (6.16), respectively, and the parameter α_0 is derived as

$$\alpha_0 = \frac{\sigma_{\text{v}}^2}{\sigma_{\text{p}}^2 + \sigma_{\text{v}}^2}. \tag{6.25}$$

the variables σ_{p}^2 and σ_{v}^2 have been defined in (6.11) as the variances of noises at pressure sensor-element and dipole sensor-elements, respectively. Note that, as opposed to (6.19) where the weighting α is manually chosen, the weighting α_0 is chosen based on the known statistics of the source signal and noise. Therefore, it has been observed in [82] that the maximum likelihood estimator achieves the optimal solution given the knowledge of noise statistics which, in practise, can be approximately estimated during source silence period. Finally, the source direction can be estimated by

$$\widehat{\mathbf{u}}^{\text{src}} = \arg\max_{\mathbf{u}} \left\{ \mathcal{J}_{\text{ML}}(\mathbf{u}) \right\}, \quad \text{s.t.} \;\; \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1. \tag{6.26}$$

6.3.2 Multi-source DOA estimation algorithms

MUSIC algorithm [50]

Now consider the multi-source free-space model in (6.4) and discuss the application of the conventional sensor-array based MUSIC algorithm [50] to AVS. By assuming the statistics of noise in (6.10) and (6.11) and that the sources are independent

with each other and uncorrelated with noise, the covariance of the received signals is derived as

$$\begin{aligned}\mathbf{R} &= \mathbb{E}\{\mathbf{y}(t)\mathbf{y}^{\mathsf{T}}(t)\} \\ &= \sum_{l=1}^{L} \sigma_{\mathrm{s}_l}^2 \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix}^{\mathsf{T}} + \begin{bmatrix} \sigma_{\mathrm{p}}^2 & \mathbf{0}_{1\times 3} \\ \mathbf{0}_{3\times 1} & \sigma_{\mathrm{v}}^2 \mathbf{I}_{3\times 3} \end{bmatrix},\end{aligned} \tag{6.27}$$

where $\sigma_{\mathrm{s}_l}^2 = \mathbb{E}\{s^2(t)\}$ is the variance of the lth source signal and L is the number of sources. In MUSIC algorithm, the number of sources is assumed to be known and is assumed to be less than the number of sensors. In addition, $L < 3$ has to be assumed in AVS since only 3 sensor-elements are configured to have direction-dependent gains. The MUSIC algorithm first performs an eigenvalue decomposition of the covariance as

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}}, \tag{6.28}$$

where $\mathbf{\Lambda}$ is the diagonal matrix consisting of the eigenvalues in a descending order and $\mathbf{Q}$ is a matrix with columns being the associated eigenvectors. The MUSIC algorithm exploits the fact that the steering vector, in general, is orthogonal to the eigenvectors corresponding to the noise subspace. Therefore, by defining $\mathbf{U}$ as the $4 \times (4 - L)$ matrix consisting of the eigenvectors associated with the $4 - L$ smallest eigenvalues, and $\mathbf{q} = [1,\ \mathbf{u}^{\mathsf{T}}]^{\mathsf{T}}$ as the steering vector, the MUSIC spatial spectrum is defined as

$$\mathcal{J}_{\mathrm{MUSIC}}(\mathbf{u}) = \frac{1}{\|\mathbf{q}^{\mathsf{T}}\mathbf{U}\mathbf{U}^{\mathsf{T}}\mathbf{q}\|}. \tag{6.29}$$

The directions of the sources can therefore estimated by

$$\widehat{\mathbf{u}}_l^{\mathrm{src}} = \arg \overset{\text{local}}{\max_{\mathbf{u}}} \mathcal{J}_{\mathrm{MUSIC}}(\mathbf{u}) \quad \text{s.t.} \quad \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1, \tag{6.30}$$

where $\overset{\text{local}}{\max}\, \mathcal{J}(\cdot)$ denotes the local maximum of the function $\mathcal{J}(\cdot)$.

Single-source point based algorithm [83]

The single-source point (SSP) based algorithm [83] operates in the time-frequency (TF) domain as in (6.9). It assumes that there exists a few TF points (ω, k) in which only one source signal is dominant. This W-disjoint assumption is, in general, valid for the mixture of speech signals in the TF domain. Based on this assumption, considering a TF point (ω, k) in which only the lth source signal is dominant and omitting the effect of noise, (6.9) can be rewritten as

$$\begin{bmatrix} \underline{y}_{\mathrm{p}}(\omega,k) \\ \underline{\mathbf{y}}_{\mathrm{v}}(\omega,k) \end{bmatrix} = \underline{s}_l(\omega,k)e^{-\jmath\omega\Delta t_l} \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix}. \tag{6.31}$$

Furthermore, by defining $\underline{\mathbf{y}}(\omega,k) = [\underline{y}_{\mathrm{p}}(\omega,k),\ \underline{\mathbf{y}}_{\mathrm{v}}^{\mathsf{T}}(\omega,k)]^{\mathsf{T}}$, the SSP algorithm identifies such a single-source point by computing the covariance across adjacent TF points over time frames given by

$$\begin{aligned} \underline{\mathbf{R}} &= \mathbb{E}\big\{\underline{\mathbf{y}}(\omega,k)\underline{\mathbf{y}}(\omega,k)^{\mathsf{H}}\big|k-k_0 \le k \le k+k_0\big\} \\ &\approx \frac{1}{K}\sum_{k'=k-k_0}^{k+k_0} \underline{\mathbf{y}}(\omega,k')\underline{\mathbf{y}}(\omega,k')^{\mathsf{H}} \\ &= \sigma_{s_l}^2(\omega,k) \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{u}_l^{\mathrm{src}} \end{bmatrix}^{\mathsf{T}}, \end{aligned} \tag{6.32}$$

where $\sigma_{s_l}^2(\omega,k) = \mathbb{E}\big\{|\underline{s}_l^2(\omega,k)|\big|k-k_0 \le k \le k+k_0\big\}$ and $\mathbb{E}\big\{\cdot\,\big|k-k_0 \le k \le k+k_0\big\}$ denotes the expectation over time frames and k_0 is the number of adjacent frames being considered. It can be observed that in (6.32), the rank of $\underline{\mathbf{R}}$ approximates to one. It can be further derived that when multiple sources exist in the TF point (ω, k) as in (6.9), the rank of $\underline{\mathbf{R}}$ will be greater than one. Therefore, the SSP algorithm

identifies a TF point (ω, k) as a single-source point using the following condition

$$(\omega, k) \quad \text{is an SSP if} \quad \text{rank}(\underline{\mathbf{R}}) \to 1. \tag{6.33}$$

After identifying the single-source TF points $\{(\omega, k)|\text{rank}(\underline{\mathbf{R}}) \to 1\}$, a clustering algorithm can be applied based on the direction-dependent features extracted from each single-source TF point [83]. After the clustering process, each cluster is supposed to contain only TF points corresponding to one of the sources and hence any single-source DOA estimation algorithm, e.g., intensity-based or velocity-covariance based estimator can be used to estimate the DOA. In addition, the MUSIC algorithm can also be used as a single-source DOA estimator by setting the number of sources to one.

6.4 Chapter summary

In this chapter, the AVS received signal models for single-source and multi-source scenarios are firstly formulated. Several existing algorithms have been discussed for single-source DOA estimation and the optimal estimator can be derived with knowledge of noise statistics. For multi-source DOA estimation, the MUSIC and SSP algorithms have been reviewed and these algorithms serve as benchmark algorithms for the proposed multi-source DOA estimation algorithm in the following chapter.

Chapter 7

Multi-source DOA Estimation by Exploiting Low-reverberant-single-source zones

In Chapter 6, several state-of-the-art algorithms have been reviewed. Although significant progress has been achieved, most of the algorithms assume free-space received signal model. multi-source DOA estimation in an enclosed environment is still challenging due to room reverberation, environmental noise and the overlapping of source spectra. In this chapter, by exploiting a unique structure of the acoustic vector sensor (AVS), i.e., the sensor elements are co-located, a multi-source DOA estimation algorithm is proposed to achieve robustness against reverberation and noise. The algorithm identifies time-frequency (TF) zones of the received signals in

Part of this chapter has been published as K. Wu, V. G. Reju and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in *Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Process. (ICASSP)*, 2015, ©[2015] IEEE.

which only one source is dominant with a high signal-to-reverberation ratio. DOA estimation is then achieved via clustering of the Hermitian angle feature. Simulation will be conducted to show the robustness of the proposed algorithm in a reverberant and noisy environment.

7.1 Introduction

In Section 6.3.2, the existing MUSIC and SSP based algorithms for DOA estimation of multiple simultaneously active sources have been reviewed. However, multi-source DOA estimation in a reverberant environment is still an open problem and attracts research attentions. Conventional multiple signal classification (MUSIC) algorithm requires more number of sensors than the number of sources [50] and with the use of a single AVS, the maximum number of sources that can be addressed is limited to be less than three. The Single-source point (SSP) based algorithm assumes the sparsity of speech source signals in the time-frequency (TF) domain [83]. By identifying and clustering the single-source TF points, DOA estimation can be achieved with higher accuracy than the conventional beamforming and subspace algorithms [93, 94]. Although the aforementioned MUSIC [50] and SSP based [83] algorithms would perform well in a low reverberant environment, the accuracy is expected to reduce with increasing reverberation time due to that the free-space environment model (6.4) is assumed for both algorithms. The co-location structure of the AVS elements and its intrinsic advantages have not been fully exploited for multi-source DOA estimation in a reverberant environment.

In this chapter, a multi-source DOA estimation algorithm using a single AVS is proposed for a moderate and highly reverberant environment. By noting that the effect of reverberation varies with frequency bins, the proposed algorithm assumes some low-reverberant-single-source (LRSS) zones would exist in the TF domain in

which only one source is dominant compared to the other sources with high signal-to-reverberation ratio. Therefore, different from SSP based algorithm [83], the proposed algorithm identifies the TF points where not only the single-source but also the low-reverberant conditions are satisfied. To achieve this, different from the use of free-space model in [83], the proposed algorithm formulate the problem based on the reverberant environment model (6.5). The proposed algorithm then discovers and exploits a fact that the unique structure of the AVS, i.e., the co-location of the sensor elements would be advantageous for identifying these LRSS zones. Compared to the use of conventional microphone arrays, This discovery makes the use of AVS interesting not only because of its compact size but also the performance gain due to its unique physical structure.

After the identification of LRSS zones, the algorithm exploits the Hermitian angle feature [95, 96], which is then used to partition the identified LRSS zones into clusters corresponding to sources. Finally, Multi-source DOA estimation can be achieved by applying a single-source DOA estimator on each cluster.

7.2 Received Signal Formulation

Consider L active sound sources in a reverberant environment and an AVS with one monopole and three orthogonal dipole elements co-located at the origin. The received signals can be formulated as

$$\begin{bmatrix} y_{\mathrm{p}}(t) \\ \mathbf{y}_{\mathrm{v}}(t) \end{bmatrix} = \sum_{l=1}^{L} s_l(t) * \begin{bmatrix} h_{\mathrm{p},l}(t) \\ \mathbf{h}_{\mathrm{v},l}(t) \end{bmatrix} + \begin{bmatrix} n_{\mathrm{p}}(t) \\ \mathbf{n}_{\mathrm{v}}(t) \end{bmatrix}, \tag{7.1}$$

where $y_{\mathrm{p}}(t)$ and $\mathbf{y}_{\mathrm{v}}(t)$ are, respectively, the monopole and the three dipole element outputs, t is the discrete time index, $s_l(t)$ denotes the lth source signal, $h_{\mathrm{p},l}(t)$ denotes the impulse response from the lth source to the monopole pressure element,

$\mathbf{h}_{\mathrm{v},l}(t)$ is a 3×1 impulse response sample vector from the lth source to the dipole elements and $*$ denotes the convolution operator. The variables $n_{\mathrm{p}}(t)$ and $\mathbf{n}_{\mathrm{v}}(t)$ are defined as the noise signals. Using the short-time Fourier transform (STFT), (7.1) can be represented as

$$\underline{\mathbf{y}}(\omega, k) = \sum_{l=1}^{L} \underline{s}_l(\omega, k)\underline{\mathbf{h}}_l(\omega) + \underline{\mathbf{n}}(\omega, k), \tag{7.2}$$

where $\underline{\mathbf{y}}(\omega, k) = [\underline{y}_{\mathrm{p}}(\omega, k),\ \underline{\mathbf{y}}_{\mathrm{v}}^{\mathsf{T}}(\omega, k)]^{\mathsf{T}}$ is the 4×1 STFT coefficient vector of the received signals, $\underline{s}_l(\omega, k)$ is the STFT coefficient of the lth source signal, $\underline{\mathbf{h}}_l(\omega) = [\underline{h}_{\mathrm{p},l}(\omega),\ \underline{\mathbf{h}}_{\mathrm{v},l}^{\mathsf{T}}(\omega)]^{\mathsf{T}}$ is the 4×1 vector formed by the STFT coefficients of the impulse responses, $\underline{\mathbf{n}}(\omega, k) = [\underline{n}_{\mathrm{p}}(\omega, k),\ \underline{\mathbf{n}}_{\mathrm{v}}^{\mathsf{T}}(\omega, k)]^{\mathsf{T}}$ is the vector of noise STFT coefficients, m is the frame index and k is the frequency-bin index.

To analyze the effect of reverberation, $\underline{\mathbf{h}}_l(\omega)$ can be further decomposed into direct-path component $\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ and reflection components $\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)$ such that (7.2) can be rewritten as

$$\underline{\mathbf{y}}(\omega, k) = \sum_{l=1}^{L} \underline{s}_l(\omega, k)\left[\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega) + \underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)\right] + \underline{\mathbf{n}}(\omega, k). \tag{7.3}$$

Since $\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ contains only direct-path component from the direction of the source it can be expressed as

$$\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega) = e^{-\jmath\omega\Delta t_l}\mathbf{q}_l^{\mathrm{src}}, \tag{7.4}$$

where $\mathbf{q}_l^{\mathrm{src}} = [1,\ \mathbf{u}_l^{\mathrm{src}\mathsf{T}}]^{\mathsf{T}}$, $\mathbf{u}_l^{\mathrm{src}} = [\cos\psi_l^{\mathrm{src}}\cos\phi_l^{\mathrm{src}},\ \cos\psi_l^{\mathrm{src}}\sin\phi_l^{\mathrm{src}},\ \sin\psi_l^{\mathrm{src}}]^{\mathsf{T}}$ define the sensor manifold pointing towards the lth source, ϕ_l^{src} and ψ_l^{src} are the azimuth and elevation direct-path incident angles, respectively. The variable Δt_l denotes the direct-path propagation delay from the lth source to the sensor and ω is the angular frequency. On the other hand, since the vector variable $\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)$ contains all reflected

components that are dependent on the environment, it can be expressed as

$$\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega) = \sum_r \alpha_l^r e^{-\jmath\omega\Delta t_l^r} \mathbf{q}_l^r, \tag{7.5}$$

where $\mathbf{q}_l^r = [1,\ \mathbf{u}_l^{r\mathsf{T}}]^{\mathsf{T}}$, $\mathbf{u}_l^r = [\cos\psi_l^r \cos\phi_l^r,\ \cos\psi_l^r \sin\phi_l^r,\ \sin\psi_l^r]^{\mathsf{T}}$ define the manifold pointing towards the rth reflection component, ϕ_l^r and ψ_l^r are the corresponding reflection incident angles, Δt_l^r is the propagation-delay of that reflection and α_l^r is the attenuation due to absorption at the room boundaries. Finally, the objective of this chapter is to estimate $\mathbf{u}_l^{\mathrm{src}}$ which, in turn, provides DOA estimates of the sources.

7.3 The Proposed DOA Estimator

As shown in (7.4) and (7.5), the impulse response $\underline{\mathbf{h}}_l(\omega)$ varies across frequencies. This implies that reverberation effect varies across frequency bins. Due to this fact, the LRSS zones are expected to exist in the TF domain in which only one of the source signals is dominant with high signal-to-reverberant ratio. To identify these LRSS zones, an algorithm is proposed which shows that the unique structure of AVS, i.e., the co-location of the sensor elements can be advantageous for detection of LRSS zones. The detected LRSS zones are then used for DOA estimation. The flow diagram of the proposed algorithm is illustrated in Fig. 7.1.

7.3.1 Identification of Low-reverberant-single-source zone

As illustrated in Fig. 7.2, consider a TF zone $\mathcal{Z}(\omega', k')$ where (ω', k') denotes the two-dimensional zone index across frequencies and time frames. Consider that the TF zone has its centroid located at $\underline{\mathbf{y}}(\omega_{\mathrm{c}}, k_{\mathrm{c}})$, and the size of the zone is $\Omega_{\mathrm{z}} \times K_{\mathrm{z}}$, where Ω_{z} is the zone width across the frequency bins and K_{z} is the zone length

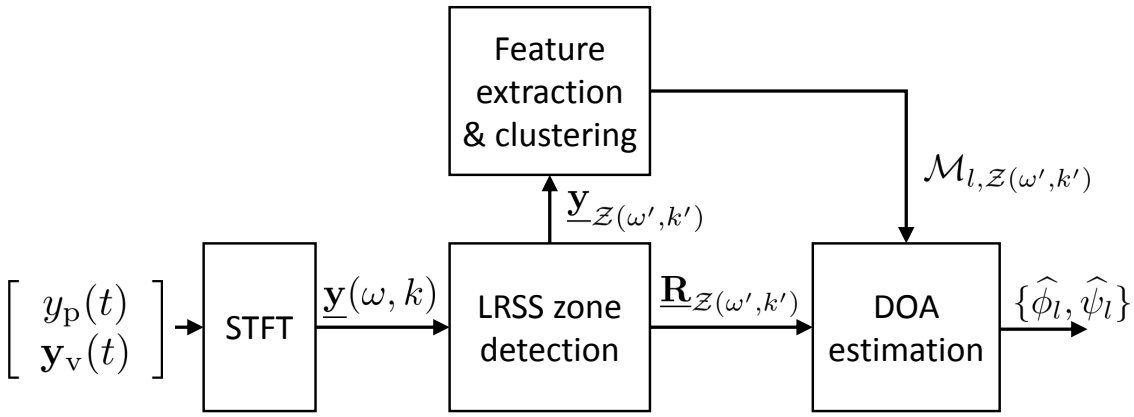


Figure 7.1: Block diagram of the proposed LRSS based DOA estimation algorithm.

across time frames, the definition for the zone is given by

$$\mathcal{Z}(\omega',k') \triangleq \left\{ (\omega,k) \;\middle|\; |\omega-\omega_{\mathrm{c}}| \le \frac{\Omega_{\mathrm{z}}}{2},\ |k-k_{\mathrm{c}}| \le \frac{K_{\mathrm{z}}}{2} \right\}. \tag{7.6}$$

Within such a TF zone, if only the lth source is dominant and that the direct-path component is significantly larger than the reflection components and noise, the received signal in (7.3) can be approximated by

$$\underline{\mathbf{y}}(\omega,k) \approx \underline{s}_l(\omega,k)\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega), \tag{7.7}$$

where $\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ is the direct-path component defined in (7.4). The covariance of $\underline{\mathbf{y}}(\omega,k)$ across all the TF points within the TF zone can then be estimated as

$$\begin{aligned} \underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')} &= \mathbb{E}\left\{ \underline{\mathbf{y}}(\omega,k)\underline{\mathbf{y}}^{\mathsf{H}}(\omega,k) \;\middle|\; \omega,k \in \mathcal{Z}(\omega',k') \right\} \\ &\approx \mathbb{E}\left\{ |\underline{s}_l(\omega,k)|^2 \underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)\underline{\mathbf{h}}_l^{\mathrm{d}\mathsf{H}}(\omega) \;\middle|\; \omega,k \in \mathcal{Z}(\omega',k') \right\} \\ &= \sigma_l^2 \mathbf{q}_l^{\mathrm{src}} \mathbf{q}_l^{\mathrm{src}\mathsf{T}}, \end{aligned} \tag{7.8}$$

where $\sigma_l^2 = \mathbb{E}\{|\underline{s}_l(\omega,k)|^2 \mid \omega,k \in \mathcal{Z}(\omega',k')\}$ is the variance of the lth source signal and $\mathbb{E}\{\cdot \mid \omega,k \in \mathcal{Z}(\omega',k')\}$ denotes the expectation over the TF points within the

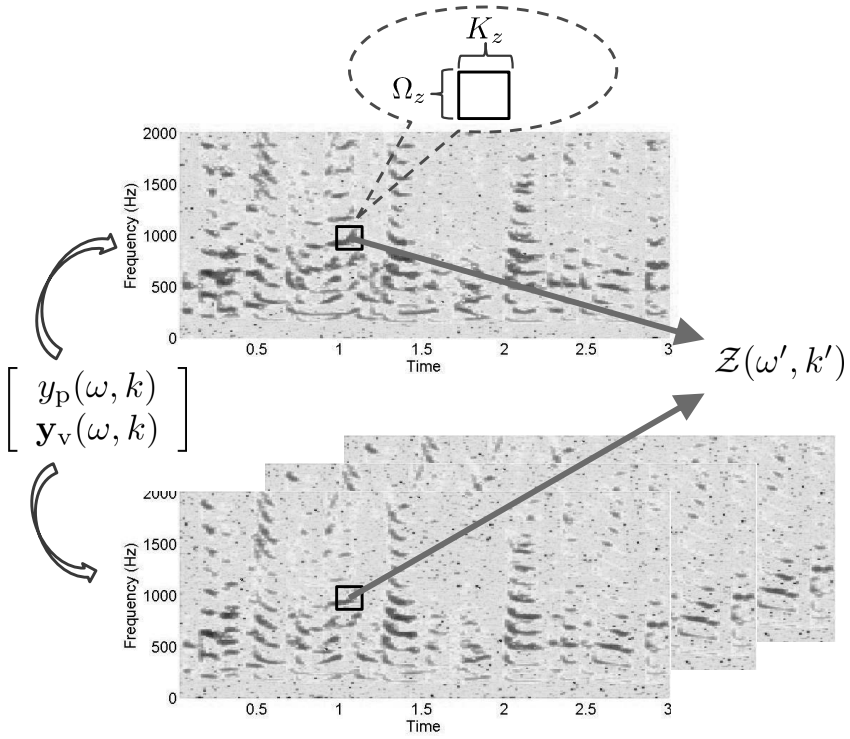


Figure 7.2: Illustration of an LRSS zone in the TF domain of the four-channel received signals. Covariance $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ for zone $\mathcal{Z}(\omega', k')$ is computed by taking expectation of $\underline{\mathbf{y}}(\omega, k)\underline{\mathbf{y}}^{\mathsf{H}}(\omega, k)$ over all the TF points within the zone.

TF zone $\mathcal{Z}(\omega', k')$.

It can be observed from (7.8) that for a LRSS zone, the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')} \to 1$. On the contrary, as the number of sources or α_l^r in (7.5) increases, the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ will increase. To illustrate this point, consider an example case in (7.3) where multiple sources are present in a reverberant-free environment. The covariance of the received signal $\underline{\mathbf{y}}(\omega, k) = \sum_{l=1}^{L} \underline{s}_l(\omega, k)\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ can be expressed by

$$\begin{aligned}
\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')} &= \mathbb{E}\Big\{ \underline{\mathbf{y}}(\omega, k)\underline{\mathbf{y}}^{\mathsf{H}}(\omega, k) \;\Big|\; \omega, k \in \mathcal{Z}(\omega', k') \Big\} \\
&\approx \mathbb{E}\Big\{ \Big(\sum_{l=1}^{L} \underline{s}_l(\omega, k)\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega) \Big) \Big(\sum_{l=1}^{L} \underline{s}_l^*(\omega, k)(\underline{\mathbf{h}}_l^{\mathrm{d}})^{\mathsf{H}}(\omega) \Big) \;\Big|\; \omega, k \in \mathcal{Z}(\omega', k') \Big\} \\
&= \sum_{l=1}^{L} \sigma_l^2 \mathbf{q}_l^{\mathrm{src}} \mathbf{q}_l^{\mathrm{src}\mathsf{T}},
\end{aligned} \tag{7.9}$$

where in the last expression, independence between sources has been assumed. It

can be observed that the covariance will result in a rank that is greater than one. Consider another contradictory case where one source is present with increased reverberation, the covariance of the received signal $\underline{\mathbf{y}}(\omega,k) = \underline{s}_l(\omega,k)\left[\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega) + \underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)\right]$ can be expressed by

$$\begin{aligned}\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')} &= \mathbb{E}\Big\{\underline{\mathbf{y}}(\omega,k)\underline{\mathbf{y}}^{\mathsf{H}}(\omega,k) \;\Big|\; \omega,k \in \mathcal{Z}(\omega',k')\Big\} \\ &\approx \mathbb{E}\Big\{\Big(\underline{s}_l(\omega,k)\left[\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega) + \underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)\right]\Big)\Big(\underline{s}_l^*(\omega,k)\left[(\underline{\mathbf{h}}_l^{\mathrm{d}})^{\mathsf{H}}(\omega) + (\underline{\mathbf{h}}_l^{\mathrm{r}})^{\mathsf{H}}(\omega)\right]\Big) \\ &\qquad \Big| \; \omega,k \in \mathcal{Z}(\omega',k')\Big\} \\ &= \sigma_l^2 \mathbf{q}_l^{\mathrm{src}}\mathbf{q}_l^{\mathrm{src}\mathsf{T}} + \mathbb{E}\Big\{|\underline{s}_l(\omega,k)|^2 \underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)(\underline{\mathbf{h}}_l^{\mathrm{r}})^{\mathsf{H}}(\omega)\Big\}, \end{aligned} \tag{7.10}$$

where in the last expression, independence between $\underline{s}_l(\omega,k)\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ and $\underline{s}_l(\omega,k)\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)$ has been assumed. It can be observed that such covariance will result a rank greater than one. This is due to the fact that $\underline{s}_l(\omega,k)\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)$ is linearly independent across frequency bins as described in (7.5) where $\alpha_l^r e^{-\jmath\omega\Delta t_l^r}$ for each $\mathbf{q}_l^r$ in the summation varies across frequencies. By taking the expectation over frequencies, the last term $\mathbb{E}\{|\underline{s}_l(\omega,k)|^2\underline{\mathbf{h}}_l^{\mathrm{r}}(\omega)(\underline{\mathbf{h}}_l^{\mathrm{r}})^{\mathsf{H}}(\omega)\}$ will have a rank greater than one.

Unlike conventional microphone arrays, it is important to note that the above rank-1 property of $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ for LRSS zone is derived from the unique structure of the AVS, i.e., the co-location of the sensor elements. More specifically, the rank-1 property only exists for AVS because its direct-path component of the impulse response $\underline{\mathbf{h}}_l^{\mathrm{d}}(\omega)$ has the same phase delays across all the four channels. Conversely, the direct-path component of the impulse response of a conventional microphone array is described by

$$\underline{\mathbf{h}}_l^{\mathrm{mic,d}}(\omega) = [e^{-\jmath\omega\Delta t_{i,1}},\; e^{-\jmath\omega\Delta t_{i,2}}, \cdots, e^{-\jmath\omega\Delta t_{i,N}}]^{\mathsf{T}}, \tag{7.11}$$

where $\Delta t_{i,n}$ is the time-delay from the lth source to the nth microphone, and N is

the number of microphones. It can be observed that the vector $\underline{\mathbf{h}}_l^{\mathrm{mic,d}}(\omega)$ is linearly independent across frequency bins and therefore the corresponding $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ does *not* have the rank-1 property; detection of LRSS zones is thus not straightforward.

It is also worth noting that in the proposed algorithm, the definition of the covariance $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ in (7.8) is different from the covariance in (6.32) for the single-source point based algorithm [83]; equation (7.8) is defined as an average across time and frequencies, while the covariance in (6.32) for [83] is averaged across time frames only. It is indeed this manipulation of averaging across frequencies that exploits the common phase-delay property of the four channels in AVS for the detection of the LRSS zones.

To detect the LRSS zones, the TF plane is divided into zones of size $\Omega_{\mathrm{z}} \times K_{\mathrm{z}}$ with 50% overlap between the zones across time frames and frequency bins. Each of these zones will be verified if they are LRSS zones by evaluating the rank of the corresponding covariance matrix. To determine whether the rank of the 4×4 covariance matrix $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ approaches to one, the coherence test [83]

$$\mathcal{C}_{\mathcal{Z}(\omega',k')} = \frac{1}{6} \sum_{a \neq b} \frac{|\underline{R}^{(a,b)}_{\mathcal{Z}(\omega',k')}|^2}{\underline{R}^{(a,a)}_{\mathcal{Z}(\omega',k')} \underline{R}^{(b,b)}_{\mathcal{Z}(\omega',k')}} \tag{7.12}$$

can be used, where $\mathcal{C}_{\mathcal{Z}(\omega',k')}$ is the coherence value for the zone $\mathcal{Z}(\omega', k')$, $\underline{R}^{(a,b)}_{\mathcal{Z}(\omega',k')}$ is the $(a,\ b)$ element of the matrix $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$. In (7.12), $0 \leq \mathcal{C}_{\mathcal{Z}(\omega',k')} \leq 1$ and a higher value of $\mathcal{C}_{\mathcal{Z}(\omega',k')}$ implies that the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}$ is closer to 1. Hence, in order to detect the LRSS zones, A threshold $\mathcal{C}_{\mathrm{thd}}$ is defined such that zones with $\mathcal{C}_{\mathcal{Z}(\omega',k')} > \mathcal{C}_{\mathrm{thd}}$ will be designated as LRSS zones. These zones will then be used for multi-source DOA estimation.

To show the effectiveness of the proposed algorithm on identifying the LRSS zones, a simulation was conducted in a room environment with $T_60 = 300$ ms and SNR = 15 dB and two speech sources are simultaneously active at $\phi_1 = 110°$, $\psi_1 =$

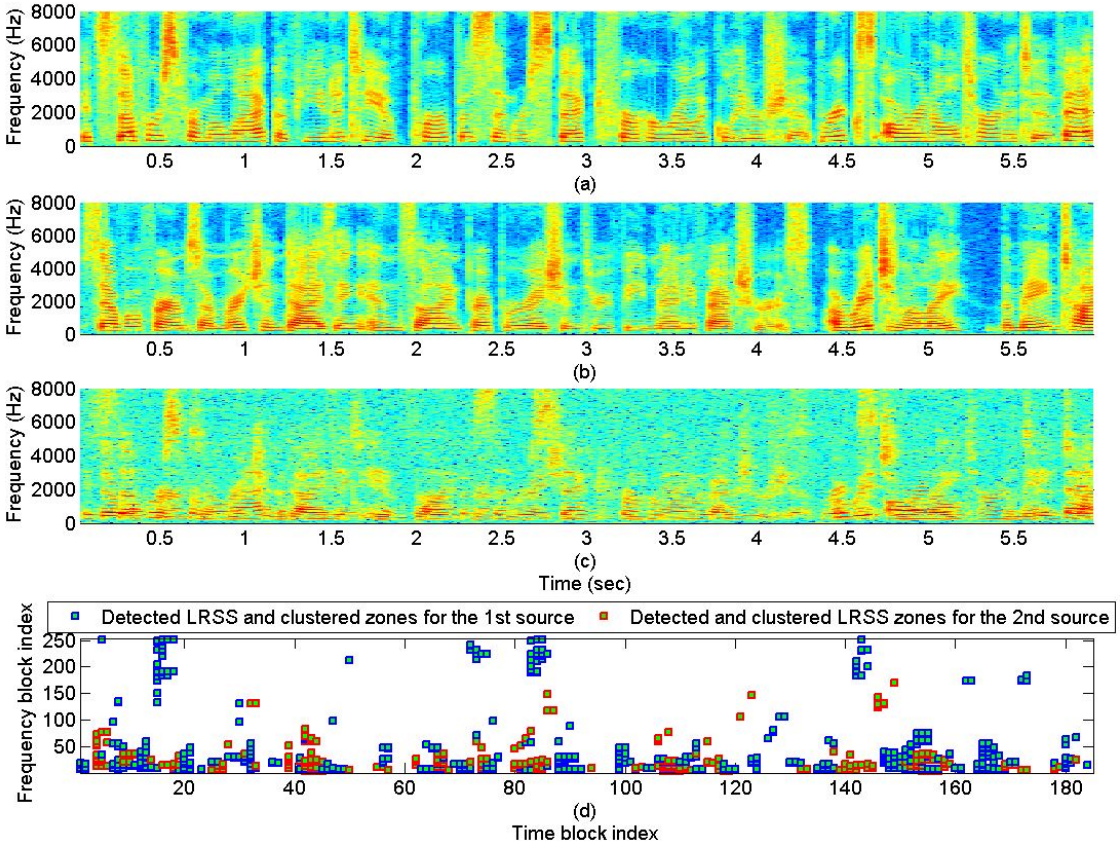


Figure 7.3: Simulated result for the identification and clustering of the LRSS zones. (a) original signal of the 1st source; (b) original signal of the 2nd source; (c) AVS received signal (omni-directional channel) in a room environment with $T_{60} = 300$ ms and SNR = 15 dB; (d) the identified and clustered LRSS zones.

$-10°$ and $\phi_2 = 165°$, $\psi_2 = 15°$ with respect to the AVS. The other configurations and parameters are kept the same as will be elaborated in Sec. 7.4. Fig. 7.3 (a) and (b) show the spectrograms of the original source signals with duration of 6 seconds. Fig. 7.3 (c) shows the AVS received signal (omni-directional channel). Fig 7.3 (d) shows the identified LRSS zones using (7.12). By comparing (d) with (c), it can be observed that the identified zones, shown in squares, correspond to the TF regions with high signal-to-reverberant ratio and signal-to-noise ratio (the dark-color regions) in Fig. 7.3 (c). This result demonstrates that the proposed algorithm is effective in identifying the TF zones which are less affected by reverberation and noise. By comparing (d) with (a) and (b), it can be further observed that the zones

are corresponding to the original sources in (a) and (b), respectively. This result shows that the identified zones are the single-source zones where one of the sources is dominant than others.

7.3.2 Feature extraction

Given a set of LRSS zones Γ that contains all of the identified LRSS zones $\{\mathcal{Z}(\omega', k')|\mathcal{Z}(\omega', k') \in \Gamma\}$ in a time block, these zones are clustered such that each cluster corresponds to a different source and single-source DOA estimation algorithm can be used for each cluster. To cluster the LRSS zones, feature of each zone has to be extracted and in this work, the Hermitian angle will be exploited.

Theorem: The Hermitian angle between two arbitrary complex vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined as [95]

$$\theta = \cos^{-1}(|\cos(\theta_C)|), \tag{7.13}$$

where the cosine of complex-valued angle θ_C is given by $\cos(\theta_C) = \mathbf{r}_1^{\mathsf{H}}\mathbf{r}_2/\|\mathbf{r}_1\|\|\mathbf{r}_2\|$ and $\|\cdot\|$ denotes Euclidian norm. In addition, the Hermitian angle between $\mathbf{r}_1$ and $\mathbf{r}_2$ will remain the same even if the vectors are multiplied by any complex scalars [96].

To extract the feature corresponding to the source DOAs using Hermitian angle, the TF point with the highest energy from each LRSS zone is firstly taken as

$$\underline{\mathbf{y}}_{\mathcal{Z}(\omega',k')} = \arg \max_{\underline{\mathbf{y}}(\omega,k)\in\mathcal{Z}(\omega',k')} ||\underline{\mathbf{y}}(\omega, k)||. \tag{7.14}$$

The elements of $\underline{\mathbf{y}}_{\mathcal{Z}(\omega',k')}$ are then used to form six two-element sub-vectors given by $\breve{\underline{\mathbf{y}}}^{<a,b>}_{\mathcal{Z}(\omega',k')} = [\underline{x}^{(a)}_{\mathcal{Z}(\omega',k')},\ \underline{x}^{(b)}_{\mathcal{Z}(\omega',k')}]^{\mathsf{T}}$, $\{a,\ b\} \subset \{\mathrm{p},\ \mathrm{v_x},\ \mathrm{v_y},\ \mathrm{v_z}\}$. Using (7.4), (7.7) can be rewritten as

$$\breve{\underline{\mathbf{y}}}^{<a,b>}_{\mathcal{Z}(\omega',k')} \approx \underline{s}_l(\omega, k)e^{-\jmath\omega\Delta t_l}\breve{\mathbf{q}}_l^{<a,b>}, \tag{7.15}$$

where $\breve{\mathbf{q}}_l^{<a,b>}$ is the vector consisting of the corresponding two elements of $\mathbf{q}_l^{\text{src}}$. In (7.15), $\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}$ is expressed as a product of a source DOA dependent vector $\breve{\mathbf{q}}_l^{<a,b>}$ with a complex scalar $\underline{s}_l(\omega,k)e^{-\jmath\omega\Delta t_l}$. It is therefore expected that the Hermitian angle between $\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}$ and any reference vector $\mathbf{r}$ will be equal to the Hermitian angle between $\breve{\mathbf{q}}_l^{<a,b>}$ and $\mathbf{r}$. In other words, if the reference vector $\mathbf{r}$ is arbitrarily fixed, the Hermitian angles $\breve{\theta}_{\mathcal{Z}(\omega',k')}^{<a,b>}$ computed from $\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}$ will be uniquely determined by their dominant source DOAs. This uniquely determined Hermitian angle can be treated as a source DOA dependent feature.

Mathematically, the Hermitian angle between $\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}$ and an arbitrarily selected $\mathbf{r}$ can be computed, using (7.13), as

$$\breve{\theta}_{\mathcal{Z}(\omega',k')}^{<a,b>} = \cos^{-1}\left(\left|\frac{\mathbf{r}^{\mathsf{H}}\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}}{\|\mathbf{r}\|\|\underline{\breve{\mathbf{y}}}_{\mathcal{Z}(\omega',k')}^{<a,b>}\|}\right|\right). \tag{7.16}$$

Furthermore, to improve clustering resolution, the above feature can be computed for every pair of sensor elements. A 6×1 vector of Hermitian angles can be hence constructed, i.e.,

$$\boldsymbol{\Theta}_{\mathcal{Z}(\omega',k')} = \left[\breve{\theta}_{\mathcal{Z}(\omega',k')}^{<\text{p,v}_\text{x}>},\ \breve{\theta}_{\mathcal{Z}(\omega',k')}^{<\text{p,v}_\text{y}>},...,\ \breve{\theta}_{\mathcal{Z}(\omega',k')}^{<\text{v}_\text{x},\text{v}_\text{y}>}\right]^{\mathsf{T}}. \tag{7.17}$$

As discussed, $\boldsymbol{\Theta}_{\mathcal{Z}(\omega',k')}$ is a six-dimension feature that depends only on the DOAs of the dominant sources and it will be used to cluster the LRSS zones in the next section.

7.3.3 Clustering and mask estimation

Given $\boldsymbol{\Theta}_{\mathcal{Z}(\omega',k')}$, the clustering of $\boldsymbol{\Theta}_{\mathcal{Z}(\omega',k')}$ and hence the corresponding LRSS zones is performed in a multi-dimensional space. Any one of the well-established data clustering algorithms [97, 98], such as k-means [99] or fuzzy c-means (FCM) [100]

may be used for this purpose. In this work, in order to avoid a binary assignment of partitioning, the FCM algorithm is employed which partitions the data into clusters with membership function that is inversely related to the distance of $\mathbf{\Theta}_{\mathcal{Z}(\omega',k')}$ to the centroid of each cluster. Therefore, defining l as the cluster index, the obtained membership function $\mathcal{M}_{l,\mathcal{Z}(\omega',k')}$ would be a smooth function.

In the use of FCM, the number of clusters/sources is generally assumed to be known a priori. In a practical scenario where the number of sources is unknown, cluster validation techniques can be applied [96, 101, 102]. By assuming a maximum number of possible sources $L_{\max}$, these approaches perform clustering of $\mathbf{\Theta}_{\mathcal{Z}(\omega',k')}$ for each candidate $L = 2, \cdots, L_{\max}$, where L denotes the number of clusters. After all the clustering has been performed, the cluster validity index will be computed for each L, The candidate L which achieves the optimal cluster validation index will be taken as the estimated number of sources. In this chapter, such clustering validity techniques are not examined due to that the focus of this chapter is DOA estimation. The number of sources is assumed to be known in the simulation.

The clustering result for the example of Fig. 7.3 can be seen in Fig. 7.3 (d). By comparing (d) with (a) and (b), it can be observed that based on the Hermitian angle features, the proposed algorithm has properly partitioned the LRSS zones in blue and red colors, respectively, corresponding to the original sources.

7.3.4 DOA estimation

After clustering, the multi-source DOA estimation problem can be solved by applying single-source DOA estimator on each cluster. The membership function $\mathcal{M}_{l,\mathcal{Z}(\omega',k')}$ obtained from the FCM algorithm is used as a mask for each source.

The covariance for the lth source is hence estimated by

$$\underline{\mathbf{R}}_l = \sum_{\mathcal{Z}(\omega',k')\in\Gamma} \mathcal{M}_{l,\mathcal{Z}(\omega',k')} \underline{\mathbf{R}}_{\mathcal{Z}(\omega',k')}. \tag{7.18}$$

In (7.18), since only the LRSS zones are taken into account, the obtained $\underline{\mathbf{R}}_l$ is expected to contain only the signal of the lth source with low signal-to-reverberation ratio. Because the TF zones have been separated, any single-source DOA estimation algorithms, such as velocity-covariance based [78], maximum SRP estimator [81] or maximum likelihood estimator [82], can be applied. In this work, the MUSIC algorithm is used due to its high spatial resolution [50]. For a single source, the MUSIC spatial spectrum is defined as

$$\mathcal{J}_l(\mathbf{u}) = \frac{1}{\|\mathbf{q}^{\mathsf{H}} \underline{\mathbf{U}}_l \underline{\mathbf{U}}_l^{\mathsf{H}} \mathbf{q}\|}, \tag{7.19}$$

where $\mathbf{q} = [1,\ \mathbf{u}^{\mathsf{T}}]^{\mathsf{T}}$ with $\mathbf{u} = [\cos\psi\cos\phi,\ \cos\psi\sin\phi,\ \sin\psi]^{\mathsf{T}}$ being the steering vector, and $\underline{\mathbf{U}}_l$ is the matrix consisting of three eigenvectors corresponding to the smallest eigenvalues of $\underline{\mathbf{R}}_l$. The direction of the lth source is then estimated by

$$\widehat{\mathbf{u}}_l^{\text{src}} = \arg\max_{\mathbf{u}} \mathcal{J}_l(\mathbf{u}), \quad \text{s.t.} \quad \mathbf{u}^{\mathsf{T}}\mathbf{u} = 1. \tag{7.20}$$

For different active sources, DOA estimation is performed using (7.18) to (7.20) for each of the identified clusters.

7.4 Simulation Results

Simulations were conducted for a room environment with room dimension of 6 m × 6 m × 4 m. An AVS was located at [3 m, 3 m, 1.3 m]. Similar to [80], room impulse responses were generated using the method of image [103]. The pressure

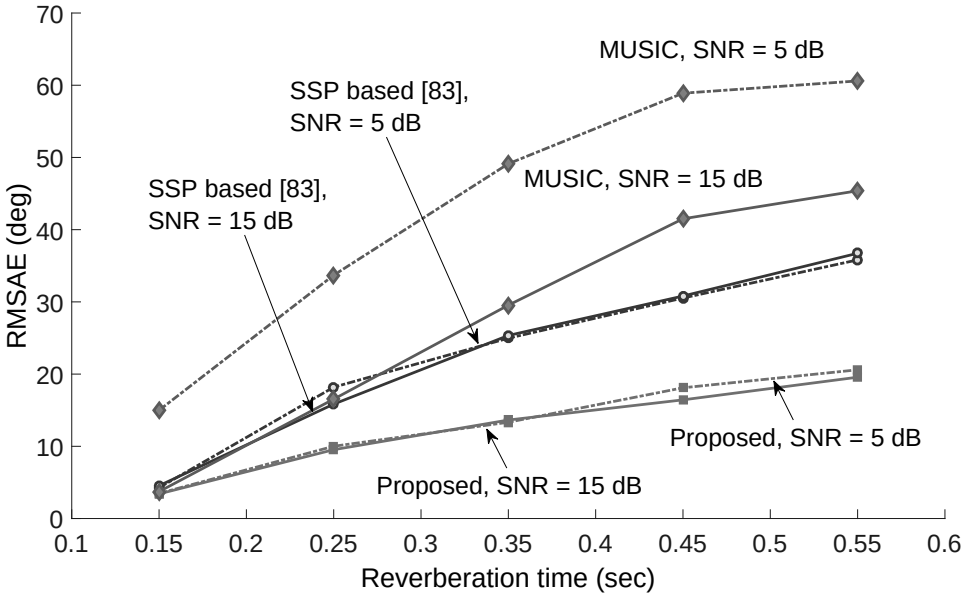


Figure 7.4: RMSAE for different reverberation time and SNR, when two sources are present at $\phi_1 = 110°$, $\psi_1 = -10°$ and $\phi_2 = 165°$, $\psi_2 = 15°$.

element was set as omni-directional and each of the vector-sensor elements was set as bi-directional with orthogonal orientation. Both male and female speech signals sampled at 16 kHz from the TIMIT database [44] were used as source signals. The sources were placed 1.7 m away from the sensor. White Gaussian noise at different signal-to-noise ratios (SNRs) were added to each of the four channels. The DOAs of the sources were estimated using 3 s block data during which the LRSS zones are identified. The frame length of STFT was 1024 samples. The TF-zone size was set to 62.5 Hz $\times$ 256 ms and this corresponds to four frequency bins and four time frames with 50% overlap across frequency bins and time frames. The coherence test threshold was set to $\mathcal{C}_{\text{thd}} = 0.75$ and the arbitrarily selected reference vector for Hermitian angle computation was $\mathbf{r} = [1 + \jmath,\ 1 + \jmath]^{\mathsf{T}}$.

The accuracy of DOA estimation is evaluated using angular error defined as the angle by which $\widehat{\mathbf{u}}$ deviates from $\mathbf{u}$ [78,82], i.e., $2\sin^{-1}(||\widehat{\mathbf{u}} - \mathbf{u}_l||/2)$. For the case

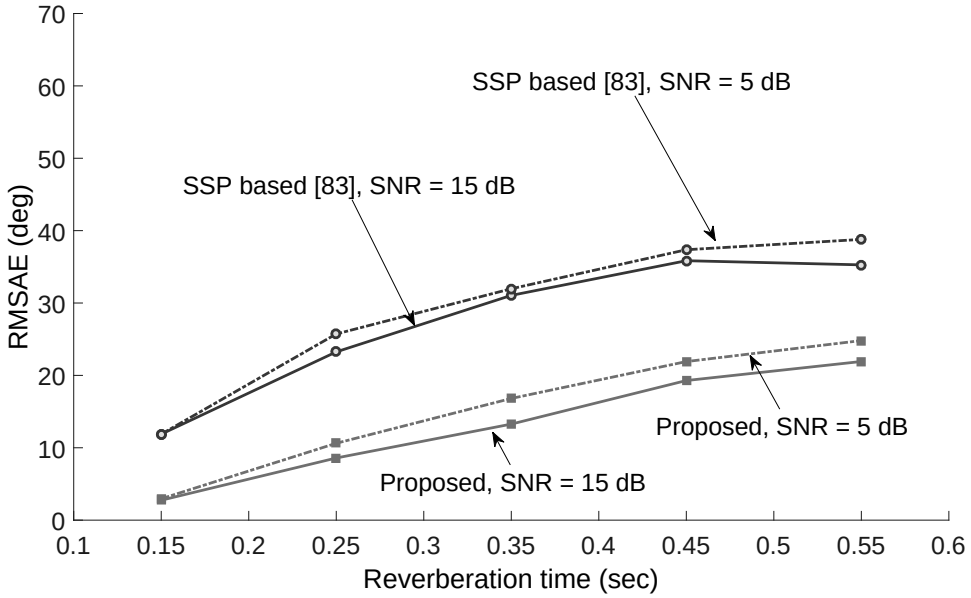


Figure 7.5: RMSAE for different reverberation time and SNR, when three sources are present at $\phi_1 = 110°$, $\psi_1 = -10°$, $\phi_2 = 165°$, $\psi_2 = 15°$ and $\phi_3 = 220°$, $\psi_3 = 20°$.

of multiple sources, it can be modified by averaging over all the sources as

$$e = \frac{1}{L}\sum_{l=1}^{L} 2\sin^{-1}\left(||\widehat{\mathbf{u}}_l^{\text{src}} - \mathbf{u}_l^{\text{src}}||/2\right). \tag{7.21}$$

The performance is then quantified across all the data blocks using the root-mean-square angular error (RMSAE) defined as RMSAE $= \sqrt{\mathbb{E}\{e^2\}}$.

In this chapter, the proposed algorithm is compared with two existing multi-source DOA estimation algorithms. The conventional MUSIC algorithm [50] is used as baseline comparison in which the covariance matrix is computed without LRSS zone detection and clustering. The single-source point (SSP) based algorithm (discussed in Sec. 6.3.2) was also implemented by detecting and clustering the single-source points in TF plane [83]. It worth noting that this SSP based algorithm is unable to detect the low-reverberant TF points/zones since the covariance matrix is obtained by averaging only across time frames (see Sec. 7.3.1 for explanation).

Figure 7.4 shows the variation of RMSAE versus reverberation time for var-

ious noise levels when two sources are simultaneously active. The two sources are located at $\phi_1 = 110°$, $\psi_1 = -10°$ and $\phi_2 = 165°$, $\psi_2 = 15°$. These results show that the performance of the three algorithms degrade with increasing reverberation, as expected. The MUSIC algorithm achieves an low error of less than 5° when $T_{60} = 150$ ms. However, its performance significantly deteriorates with increasing reverberation due to the free-space assumption in its signal model. It is also observed that the increase of noise also affects the performance of MUSIC algorithm. The SSP based algorithm achieves a lower error than the MUSIC algorithm since only single-source points are exploited for DOA estimation. The proposed algorithm achieves the lowest error compared to the other two algorithms and is observed to be less sensitive to reverberation. This is due to the fact that only LRSS zones are identified and utilized for the proposed algorithm, which are less affected by reverberation.

Figure 7.5 shows the performance of the algorithms for three active sources, where the third source is placed at $\phi_3 = 220°$, $\psi_3 = 20°$. In this figure, the results of MUSIC are not included since the MUSIC algorithm requires that the number of sources should be less than the number of dipole elements in an AVS. Similar to previous simulation, the proposed algorithm achieves a lower error than the SSP algorithm. However, the performance reduces with increasing reverberation since the number of LRSS zones is reduced.

Figure 7.6 shows the accuracy of the DOA estimation algorithms for various angular distance between two active sources. It can be observed that the error increases with reducing angular distance for the MUSIC algorithm. Although MUSIC is well-known to achieve high resolution, its performance is reduced when only one AVS is used in a reverberant and noisy environment. On the other hand, the SSP based algorithm and the proposed algorithm are generally less sensitive to the source positions since they cluster the single-source TF points/zones before DOA

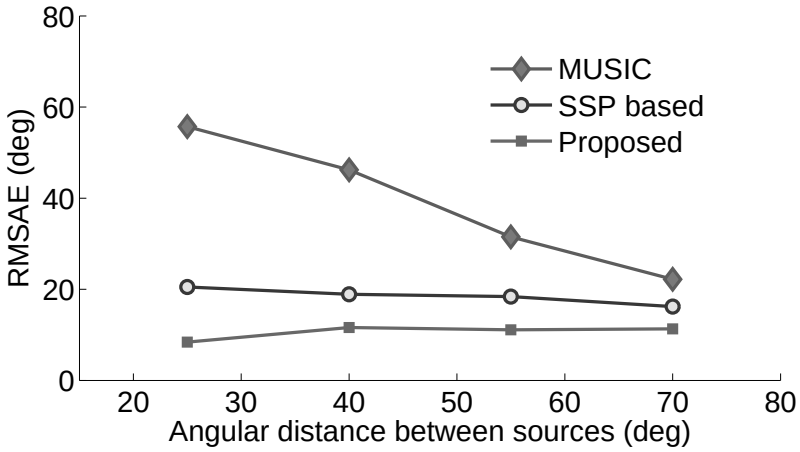


Figure 7.6: Variation of RMSAE against angular distance between two active sources for $T_{60} = 300$ ms and SNR = 15 dB.

estimation. The proposed algorithm achieves a lower error compared to the SSP algorithm due to exploitation of LRSS zones.

7.5 Chapter summary

A multi-source DOA estimation algorithm using a single AVS is proposed in this chapter. By exploiting the co-location of sensor elements, the proposed algorithm identifies the LRSS zones available in the TF plane of the sensor outputs. The LRSS zones are then separated into clusters according to the sources for which the Hermitian angle feature is utilized. DOA estimation is then applied on each of these clusters. Simulation results show that the proposed algorithm achieves lower DOA estimation error than the MUSIC and SSP-based algorithms in a noisy and reverberant environment.

Chapter 8

Conclusions and Future Research Directions

8.1 Conclusions

In this thesis, two applications of acoustic source localization and tracking have been considered, namely, the use of conventional omni-directional microphone array and the use of acoustic vector sensor (AVS).

Part I of this thesis focused on the problem of acoustic source tracking using conventional microphone arrays. Firstly, the single-source tracking scenario has been considered and the sequential-importance-resampling particle filter (SIRPF) has been used as a basic tracking framework. To achieve single-source tracking in a reverberant and noisy environment, an algorithm is proposed which employs a regional steered-response-power beamformer (RSRP beamformer) for approximating the measurement likelihood. Compared to the conventional steered-response-power beamformer (SRP beamfomer), the RSRP beamformer achieves fewer spurious peaks in a reverberant and noisy condition, which, as a consequence, makes it more suit-

able for approximation of the measurement likelihood. The proposed SIRPF-RSRP algorithm is compared with the existing SIRPF-SRP algorithm in a simulated environment with various reverberant and noisy conditions.

Secondly, for tracking of single speech source in the presence of sound interference, a SIRPF based tracking algorithm which exploits the unique speech harmonicity feature has been proposed. The proposed algorithm uses the harmonicity based SRP beamformer (HSRP) for the approximation of the measurement likelihood. Due to the use of harmonicity, speech-sensitive tracking can be achieved and other sound interference can be ignored by the system. Simulation results show that the proposed SIRPF-HSRP algorithms outperforms the existing SIRPF-SRP algorithm.

The thesis next considers the alternating-source scenario where the speakers are active in turns. To achieve a rapid system convergence to the active source when alternation occurs and reduce the tracking error in a reverberant and noisy environment, a swarm intelligence based particle filter (SWIPF) framework has been proposed. The SWIPF framework jointly exploits the advantages of particle filter and particle swarm intelligence. The particle filter is used as a sequential state estimation framework that is suitable for the tracking application. The limitation of a particle filter, which lies in the particle sampling problem, is addressed by the interaction and memory mechanism in the particle swarm intelligence. Simulation and experiment show that the proposed SWIPF framework outperforms the existing SIRPF and extended Kalman particle filter tracking frameworks in the alternating-source scenario.

Part II of this thesis focused on the problem of multi-source direction-of-arrival (DOA) estimation using AVS. Compared to the conventional microphone array, it has been shown that the co-location of the sensor-elements in AVS can be exploited to achieve reverberation-robust DOA estimation. An algorithm is pro-

posed which identifies the low-reverberant-single-source (LRSS) zones of the received signal is then proposed. By using only these LRSS zones of the received signal, the proposed algorithm achieves multi-source DOA estimation which is robust to reverberation. Simulation verifies that the proposed algorithm outperforms the existing multiple signal classification and single-source point based multi-source DOA estimation algorithms.

8.2 Future research directions

The followings are some of the possible suggestions for future research:

1. **Tracking time-varying number of sources.** Part I of this thesis focused on issues pertaining to single-source tracking and alternating-source tracking. In recent years, tracking time-varying number of sources has also gained much interest in the research community [17, 37, 104]. In a typical environment, multiple speakers may speak simultaneously and the speakers may be active or inactive at any time instant. This practical situation requires an advanced probabilistic model [17,105] to be incorporated in the existing PF framework to achieve time-varying number of speaker tracking. In addition, such algorithm should be able to detect and initialize a new-born targets and remove inactive targets [37].

2. **Tracking of DOAs of acoustic sources using AVS.** In Part II of this thesis, the problem of multi-source DOA estimation in a reverberant environment has been described. The number of the sources, however, is assumed to be known and unchanged. Estimating the number of sources can further be investigated for a more practical scenario. Furthermore, fusion of DOA estimates across time frames can be considered using the tracking framework

as discussed in Part I of this thesis [92].

References

[1] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Magazine*, vol. 28, pp. 20–32, Jan. 2011.

[2] E. A. P. Habets and J. Benesty, "A perspective on frequency-domain beamformers in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 947–960, March.

[3] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, May, pp. 2592–2595.

[4] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[5] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *Proc. 7th ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2012*. IEEE, 2012, pp. 431–438.

[6] G. Valenzise, L. Gerosa, M. Tagliasacchi, E. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS' 07)*, Sept., pp. 21–26.

[7] A. W. H. Khong and M. Brookes, "The effect of calibration errors on source localization with microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, April, vol. 1, pp. 137–140.

[8] W. Zeng and X. Li, "High-resolution multiple wideband and nonstationary source localization with unknown number of sources," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3125–3136, June.

[9] J. Dmochowski, J. Benesty, and S S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327–1339, 2007.

[10] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP J. on Adv. Signal Process.*, vol. 2007, 2007.

[11] M. F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1228–1242, 2010.

[12] X. Zhong and J. R. Hopgood, "Particle filtering for TDOA based acoustic source tracking: nonconcurrent multiple talkers," *Signal Processing*, vol. 96, Part B, pp. 382 – 394, 2014.

[13] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1490–1502, Nov. 2008.

[14] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[15] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[16] J. H. DiBiase, *A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays*, Ph.D. thesis, Brown Univ., 2000.

[17] W. K. Ma, B. N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, 2006.

[18] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, 2006.

[19] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[20] G. Clifford Carter, Albert H. Nuttall, and P. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.

[21] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, 1987.

[22] K. Levenberg, "A method for the solution of certain problems in least squares," *Quat. Applied Math.*, vol. 2, pp. 164–168, 1944.

[23] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. of Soc. for Industrial Applied Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[24] J. DiBiase, H. Silverman, and M Brandstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications.*, pp. 157–180, 2001.

[25] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[26] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.

[27] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, April, vol. 1, pp. I–121–I–124.

[28] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 826–836, 2003.

[29] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[30] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'01)*, 2001, pp. 3021–3024.

[31] A. Levy, S. Gannot, and E. A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1540–1555, Aug. 2011.

[32] R. V. D. Merwe, A. Doucet, N. D. Freitas, and E. Wan, "The unscented particle filter," in *NIPS*, 2000, pp. 584–590.

[33] F. Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1806–1817, Sep. 2010.

[34] E. Lehmann, A. M. Johansson, and S. Nordholm, "Modeling of motion dynamics and its influence on the performance of a particle filter for acoustic speaker tracking," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoust.* IEEE, 2007, pp. 98–101.

[35] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *Proc. IEE -F, Radar and Signal Process.* IET, 1993, vol. 140, pp. 107–113.

[36] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP J. on Applied Signal Process. (special issue on microphone arrays)*, vol. 2006, pp. 1–17, 2006.

[37] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[38] T. Li and W. Ser, "Three dimensional acoustic source localization and tracking using statistically weighted hybrid particle filtering algorithm," *Signal Processing*, vol. 90, no. 5, pp. 1700–1719, 2010.

[39] A. Brutti and F. Nesta, "Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs," *Computer Speech & Language*, vol. 27, no. 3, pp. 660–682, 2013.

[40] X. Zhong and J.R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, 2009, pp. 253–256.

[41] A. Masnadi-Shirazi and B.D. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 828–841, 2013.

[42] X. Zhong and J.R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 293–296.

[43] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[44] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, 1993.

[45] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, July 2008.

[46] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–11, 2012.

[47] M. Kepesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. and Microphone Arrays, 2008*, May, pp. 85–88.

[48] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[49] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time Processing of Speech Signals*, Wiley-IEEE Press, 2000.

[50] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon & Schuster, 1992.

[51] S. Timofeev, A. R. S. Bahai, and P. Varaiya, "Adaptive acoustic beamformer with source tracking capabilities," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2812–2820, 2008.

[52] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 854–867, 2012.

[53] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[54] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 994–1006, 2012.

[55] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, no. 8, pp. 1223–1235, Aug. 1988.

[56] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2914–2919, 1999.

[57] X. Zhong and J.R. Hopgood, "A time–frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2356–2370, 2015.

[58] A. Quinlan and F. Asano, "Tracking a varying number of speakers using particle filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 297–300.

[59] Y. Oualil and D. Klakow, "Multiple concurrent speaker short-term tracking using a Kalman filter bank," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'14)*. IEEE, 2014, pp. 1444–1448.

[60] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Proc. Classification of Events, Activities and Relationships (CLEAR)*. Springer, 2006, pp. 137–150.

[61] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[62] A. P. Engelbrecht, *Fundamentals of Computational Swarm Intelligence*, vol. 1, Wiley Chichester, 2005.

[63] X. Zhang, W. Hu, W. Qu, and S. Maybank, "Multiple object tracking via species-based particle swarm optimization," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1590–1602, 2010.

[64] M. Thida, H. Eng, D. N. Monekosso, and P. Remagnino, "A particle swarm optimisation algorithm with interactive swarms for tracking multiple targets," *Applied Soft Computing*, vol. 13, no. 6, pp. 3106–3117, 2013.

[65] R. Parisi, P. Croene, and A. Uncini, "Particle swarm localization of acoustic sources in the presence of reverberation," in *Proc. IEEE Int. Symposium on Circuits and Systems. (ISCAS 2006)*. IEEE, 2006, pp. 4739–4742.

[66] E. Antonacci, D. Riva, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking of two acoustic sources in reverberant environments using a particle swarm optimizer," in *Proc. IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS' 07)*. IEEE, 2007, pp. 567–572.

[67] M. Hirakawa and K. Suyama, "Multiple sound source tracking by two microphones using pso," in *Proc. Intelligent Sig. Process. and Comm. Systems (ISPACS 2013)*, 2013, pp. 467–470.

[68] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu, "Sequential particle swarm optimization for visual tracking," in *Proc. Computer Vision and Pattern Recognition (CVPR 2008)*. IEEE, 2008, pp. 1–8.

[69] X. Zhang, W. Hu, and S. Maybank, "A smarter particle filter," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 236–246.

[70] B. Ristic, S. Arulampalm, and N. J. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House Publishers, 2004.

[71] M. P. Wand and M. C. Jones, *Kernel Smoothing*, CRC Press, 1994.

[72] C. Musso, N. Oudjane, and F. LeGland, "Improving regularised particle filters," in *Sequential Monte Carlo Methods in Practice*, pp. 247–271. Springer, 2001.

[73] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 26, CRC press, 1986.

[74] E. A. Lehmann, "Room impulse response generator," `www.eric-lehmann.com`, (Accessed: 23/01/2015).

[75] K. Aspelin, "Establishing pedestrian walking speeds," *Portland State University*, pp. 5–25, 2005.

[76] K. Wu, S. T. Goh, and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, 2013.

[77] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[78] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, Sep. 1994.

[79] Microflown AVISA, "A comercialized avs product," `http://http://microflown-avisa.com/products/acoustic-multi-mission-sensor/`, (Accessed: 15/09/2015).

[80] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 1800–1811, 2010.

[81] D. Levin, S. Gannot, and E. A. P. Habets, "Direction-of-arrival estimation using acoustic vector sensors in the presence of noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, 2011, pp. 105–108.

[82] D. Levin, E. A. P. Habets, and S. Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. 1240–1248, 2012.

[83] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[84] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," *IEEE Trans. Signal Process.*, vol. 46, no. 9, pp. 2291–2304, 1998.

[85] M. Hawkes and A. Nehorai, "Wideband source localization using a distributed acoustic vector-sensor array," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1479–1491, 2003.

[86] D. Rahamim, J. Tabrikian, and R. Shavit, "Source localization using vector sensor array in a multipath environment," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3096–3103, 2004.

[87] H. Chen and J. Zhao, "Coherent signal-subspace processing of acoustic vector sensor array for DOA estimation of wideband sources," *Signal Processing*, vol. 85, no. 4, pp. 837–847, 2005.

[88] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3D sound localization system with miniature microphone array for virtual reality," in *Proc. 7th IEEE Int. Conf. Industrial Electronics and Applications (ICIEA)*, 2012, pp. 1853–1857.

[89] S. Miron, N. L. Bihan, and J. I. Mars, "Quaternion-MUSIC for vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1218–1229, 2006.

[90] X. Zhong and A. B. Premkumar, "Particle filtering approaches for multiple acoustic source detection and 2-D direction of arrival estimation using a single acoustic vector sensor," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4719–4733, 2012.

[91] M. K. Awad and K. T. Wong, "Recursive least-squares source tracking using one acoustic vector sensor," *IEEE Trans. Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3073–3083, 2012.

[92] X. Zhong, A. B. Premkumar, and H. Wang, "Multiple wideband acoustic source tracking in 3-D space using a distributed acoustic vector sensor array," *IEEE Sensors Journal*, vol. 14, no. 8, pp. 2502–2513, Aug. 2014.

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

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

[95] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Math.*, vol. 69, no. 1, pp. 95–103, 2001.

[96] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101–116, 2010.

[97] A. K. Jain K, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[98] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE. Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[99] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symp. on Math. Statist. and Prob.*, 1967, vol. 1, pp. 281–297.

[100] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Kluwer Academic Publishers, 1981.

[101] H. Sun, S. Wang, and Q. Jiang, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2027–2037, 2004.

[102] Y. Zhang, W. Wang, X. Zhang, and Y. Li, "A cluster validity index for fuzzy clustering," *Information Sciences*, vol. 178, no. 4, pp. 1205–1218, 2008.

[103] E. A. P. Habets, "Room impulse response (RIR) generator," `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`, (Accessed: 22/07/2016).

[104] M. R. Morelande, C. M. Kreucher, and K. Kastella, "A Bayesian approach to multiple target detection and tracking," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 1589–1604, 2007.

[105] B. T. Vo, B. N. Vo, and C. Antonio, "Bayesian filtering with random finite set observations," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1313–1326, 2008.

Author's Publications

Book chapters

[1] K. Wu and A. W. H. Khong, "Sound source localization and tracking for social robots," *Context Aware Human-Robot and Human-Agent Interaction*, pp 55-78, Springer, 2016

Journals

[2] K. Wu, V. G. Reju, A. W. H. Khong and S. T. Goh, "Swarm Intelligence Based Particle Filter for Alternating Talker Localization and Tracking Using Microphone Arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 25, no. 6, pp. 1384-1397, 2017.

Conference Proceedings

[3] K. Wu, S. T. Goh and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Process. (ICASSP)*, 2013.

[4] K. Wu and A. W. H. Khong, "Acoustic source tracking in reverberant environment using regional steered response power measurement," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, 2013, invited paper.

[5] K. Wu, V. G. Reju and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor," in

Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Process. (ICASSP), 2015.

[6] K. Wu, V. G. Reju and A. W. H. Khong, "Single-channel speech enhancement in a transient noise environment by exploiting speech harmonicity," in *Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Process. (ICASSP)*, 2015.