A Bayesian framework for multiple acoustic source tracking

Xionghu Zhong

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
October 2010

Abstract

Acoustic source (speaker) tracking in the room environment plays an important role in many speech and audio applications such as multimedia, hearing aids and hands-free speech communication and teleconferencing systems; the position information can be fed into a higher processing stage for high-quality speech acquisition, enhancement of a specific speech signal in the presence of other competing talkers, or keeping a camera focused on the speaker in a video-conferencing scenario. Most of existing systems focus on the single source tracking problem, which assumes one and only one source is active all the time, and the state to be estimated is simply the source position. However, in practical scenarios, multiple speakers may be simultaneously active, and the tracking algorithm should be able to localise each individual source and estimate the number of sources. This thesis contains three contributions towards solutions to multiple acoustic source tracking in a moderate noisy and reverberant environment.

The first contribution of this thesis is proposing a time-delay of arrival (TDOA) estimation approach for multiple sources. Although the phase transform (PHAT) weighted generalised cross-correlation (GCC) method has been employed to extract the TDOAs of multiple sources, it is primarily used for a single source scenario and its performance for multiple TDOA estimation has not been comprehensively studied. The proposed approach combines the degenerate unmixing estimation technique (DUET) and GCC method. Since the speech mixtures are assumed window-disjoint orthogonal (WDO) in the time-frequency domain, the spectrograms can be separated by employing DUET, and the GCC method can then be applied to the spectrogram of each individual source. The probabilities of detection and false alarm are also proposed to evaluate the TDOA estimation performance under a series of experimental parameters.

Next, considering multiple acoustic sources may appear nonconcurrently, an extended Kalman particle filtering (EKPF) is developed for a special multiple acoustic source tracking problem, namely "nonconcurrent multiple acoustic tracking (NMAT)". The extended Kalman filter (EKF) is used to approximate the optimum weights, and the subsequent particle filtering (PF) naturally takes the previous position estimates as well as the current TDOA measurements into account. The proposed approach is thus able to lock on the sharp change of the source position quickly, and avoid the tracking-lag in the general sequential importance resampling (SIR) PF.

Finally, these investigations are extended into an approach to track the multiple unknown and time-varying number of acoustic sources. The DUET-GCC method is used to obtain the TDOA measurements for multiple sources and a random finite set (RFS) based Rao-blackwellised PF is employed and modified to track the sources. Each particle has a RFS form encapsulating the states of all sources and is capable of addressing source dynamics: source survival, new source appearance and source deactivation. A data association variable is defined to depict the source dynamic and its relation to the measurements. The Rao-blackwellisation step is used to decompose the state: the source positions are marginalised by using an EKF, and only the data association variable needs to be handled by a PF. The performances of all the proposed approaches are extensively studied under different noisy and reverberant environments, and are favorably comparable with the existing tracking techniques.

Declaration of originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the School of Engineering at The University of Edinburgh.

Xionghu Zhong
Edinburgh
October 2010

Acknowledgements

I could probably have written a full chapter which is of comparable size to the contribution parts of this thesis to acknowledge those, who, in different ways, have helped me in my four-year PhD life at the University of Edinburgh.

First, my sincere gratitude to Dr. James R. Hopgood for his supervision and support throughout the last four years. I have benefited tremendously from his suggestions on my research as well as his patience in word correction of all my manuscripts.

I am also grateful to my second supervisor, Professor Bernard Mulgrew, for his critical review of my technical report and papers. Thanks due to Dr. Wenwu Wang (The University of Surrey) and Dr. Zhuo Zhang (Symantec company) who helped me in applying the Wing-Yip bursary for my last year study.

My thanks go to all members in Espresso club: David, Gabriel, George, Graham, Ioannis, Mehrdad, Mohammad, Renato and Xiaoyan, for many stimulating discussions over the coffee time. Also my thanks to all other colleges and staffs in IDCOM.

I have had many fun times with my friends at Edinburgh, particularly those badminton friends, yansong, yufeng, chen-zhan, xu-zhao and Laoliu, and have benefited from the time I have spent with them.

Finally, my deepest gratitude to my wife, Le Xu, who has been a constant source of happiness and motivation, and takes care of our new born baby at Manchester during the final stage of thesis writing up. Also deepest gratitude to my parents who have helped me throughout my whole life.

This dissertation is dedicated to my baby: Chenlin Zhong.

Xionghu Zhong
Edinburgh, October 2010

Contents

List of figures

List of tables

Acronyms and abbreviations

AED	adaptive eigenvalue decomposition
CC	cross-correlation
cdf	cumulative density function
CL	curvilinear
CLT	central limit theorem
CRLB	Cramer-Rao lower bound
CU	coordinate uncoupled
DTFT	discrete time Fourier transform
DUET	degenerate unmixing estimation technique
EKF	extended Kalman filter
EKPF	extended Kalman particle filtering
FIR	finite impulse response
GCC	generalised cross-correlation
GR	gain-ratio
IIR	infinite impulse response
IKF	iterative Kalman filter
ILD	interaural level difference
IMM	interacting multiple model
JPDA	joint probabilistic data association
KF	Kalman filter
LMS	Least mean square
LS	least square
MAST	multiple acoustic source tracking
MCMC	Markov chain Monte Carlo
MHT	multiple hypothesis tracking
ML	maximum likelihood
PDA	probability data association
pdf	probability density function
PF	particle filtering

PHAT	phase transform
PHD	probability hypothesis density
PSR	preserved signal ratio
RBPF	Rao-Blackwellised particle filtering
RFS	random finite set
RIR	room impulse response
RMS	root-mean-square
ROC	receiver operating characteristic
SBF	steered beamforming
SIR	sequential importance resampling
SIS	sequential importance sampling
SMC	sequential Monte Carlo
SNR	signal-to-noise ration
SPD	speech pause detector
SRR	signal-to-reverberation ratio
STFT	short time Fourier transform
TDE	time-delay estimation
TDOA	time-delay of arrival
TF	time-frequency
	transfer function
UKF	unscented Kalman filtering
UT	unscented transform
VAD	voice activity detector
WDO	window-disjoint-orthogonality
WGN	white Gaussian noise
2-D	two dimensional

Nomenclature

$\star$	convolution operation
$\cup$	set union
$\cap$	intersection
$\setminus$	set minus
$\|\| \cdot \|\|$	Euclidean norm
$\| \cdot \|$	absolute value/module operation/cardinality
$\angle \cdot$	phase operation
$\mathcal{F}\{\cdot\}$	Fourier transform
$\mathcal{F}^{-1}\{\cdot\}$	inverse Fourier transform
$\mathbb{E}(\cdot)$	expectation operation
$*$	conjugate
$\Longrightarrow$	convergence in distribution
$\overset{a.s.}{\rightarrow}$	converges almost surely
$\delta_{x'}(x)$	dirac function
t	continuous time index
ω	discrete frequency index
k	discrete time step (frame) index
m	source number index
ℓ	microphone pair index
i	microphone/particle/wall index
M	number of source
L	number of microphone pair
N	number of particles
a	amplitude attenuation
r	distance between microphone and source
d	microphone separation
c	sound velocity
T	length of frame
	superscript denoting transpose

E	signal energy
f_s	sampling frequency
V	room volumn
$\mathcal{A}$	wall reflection area
ρ	wall reflection coefficient
α	wall absorption coefficient
T_{60}	reverberation time
h_{d}	direct impulse response
h_{r}	reverberation impulse response
z_{d}	received direct path signal
z_{r}	received reverberant signal
$\boldsymbol{p}$	microphone position
τ	time delay
x	x-coordinate position
$\dot{x}$	x-coordinate velocity
y	y-coordinate position
$\dot{y}$	y-coordinate velocity
$\mathbf{x}$	source state
$\mathcal{X}$	state set
$s(t)$	source signal
$\mathbf{s}_m(k)$	kth frame signal for mth source
$S_m(k,\omega)$	Fourier transform of kth frame signal for mth source
$z(t)$	received signal
$z_{\ell,i}(t)$	received signal at ith microphone of ℓth microphone pair
$\mathbf{z}_{\ell,i}(k)$	kth frame of received signal
$Z_{\ell,i}(k,\omega)$	Fourier transform of kth frame of received signal
z_k	singleton measurement set
$\mathbf{z}_k^{\ell}$	measurement vector at ℓth microphone pair
$\mathcal{Z}_k$	compete measurement set
$\phi_\ell(k,\omega)$	phase term
$\phi^{\text{high}}_{(k,\omega)}$	phase term for the higher frequency band
$\phi^{\text{low}}_{(k,\omega)}$	phase term for the low frequency band
$\phi_{(\text{pred},\omega)}$	predicted phase term

$v_{\ell,i}(t)$	noise term
$v_{\ell,i}(t)$	Fourier transform of noise term
$R_{\ell,i}(k)$	autocorrelation function
R_{ss}	autocorrelation function of source signal
R_{vv}	autocorrelation function of noise term
$R_{\ell}(k)$	cross-correlation function
$G_{\ell,i}(k)$	power spectral density
$G_{\ell}(k)$	cross power spectral density
$\Phi_{\ell}(k,\tau)$	PHAT-GCC weighting term
R_{TH}	threshold of the GCC function
P_{D}	probability of detection
P_{F}	probability of false alarm
$\Lambda_i(k,\omega)$	binary mask indicator for source i at TF bin (k,ω)
$\Lambda^{\ell}_{(A,D)}(k,\omega)$	GR and TD information indicator at microphone pair ℓ
$\mathbb{I}^{\ell}_{n}(k,\omega)$	TF bin indicator for nth cluster at ℓth microphone pair
(A,D)	GR and TD resolution parameter
(ζ,η)	GR and TD index pair
ϵ	anomaly error
T_c	correlation time
$\mathbf{F}_k$	state transition matrix
$\mathbf{G}_k$	measurement process matrix
$\mathbf{Q}_k$	process noise variance matrix
$\mathbf{R}_k$	measurement noise matrix
$\mathbf{K}_k$	Kalman information gain
$\mathbf{P}_k$	state variance matrix
$\mathbf{O}_{\mathbf{x}}(\mathbf{x}_{k-1})$	higher order error in state linearisation
$\mathcal{H}_{i,k}$	hypothesis for ith TDOA measurement
$\mathbb{I}_{\tau}(\hat{\tau}^{(\ell)}_{i,k})$	time-delay indicator, $\hat{\tau}^{(\ell)}_{i,k} \in \tau = [-\tau_{\max}\ \tau_{\max}]$
$\mathcal{U}_{\tau}(\cdot)$	uniform distribution over the time-delay range τ
$\mathcal{N}(\cdot)$	normal distribution
$\mathcal{G}(\cdot)$	gamma distribution
$p(\mathbf{x})$	probability density of $\mathbf{x}$
$q(\mathbf{x})$	importance distribution of $\mathbf{x}$

$f_k(\cdot)$	dynamic model
$g_k(\cdot)$	measurement model
$\mathbf{x}_{1:k}$	state vector series from time step 1 to k
$\mathbf{z}_{1:k}$	measurement vector series from time step 1 to k
$\mathbf{x}^{(i)}$	particles
$w^{(i)}$	particle weights
(α, β)	Gamma pdf parameter
$\hat{I}_N(f(\mathbf{x}_{0:k}))$	expected estimation of function $f(\mathbf{x}_{0:k})$
P_b	birth prior
P_d	death prior
$\mathcal{B}_k$	new born states
$\mathcal{D}_k$	death states
$\boldsymbol{\theta}_k$	source dynamic and association vector
$\boldsymbol{b}_k$	birth variable
$\boldsymbol{d}_k$	death variable
$\boldsymbol{\gamma}_k$	data association variable
P_k	probability of the correct number estimation
ϵ_k	cardinality error of the source number estimation
ξ_k	mean deviation of the position estimation

Chapter 1
Introduction

Estimating the position of acoustic source in the room environments is a fundamental problem, and lies at the heart of many speech and audio processing applications. Nowadays, more and more advanced signal processing techniques are developed to meet the requirements of tracking acoustic sources in adverse environments and particularly, handling multiple simultaneously active sources. In this thesis, a number of novel approaches will be investigated for multiple acoustic source tracking (MAST) in the room environments. This introductory chapter first presents the motivation of this thesis in Section 1.1. Acoustic source localisation (ASL) problem and elements for MAST are then outlined in Section 1.2. Section 1.3 gives a brief overview of the state of the art towards solutions to MAST problem subsequently. Finally, Section 1.4 specifies our contributions and provides an overview of the work in this thesis.

1.1 Motivation

As a development of science and technology, people becomes increasingly reliant on computers and human-computer systems. Recent advances have made some new applications of speech and audio processing such as multimedia, hearing aids, and hands-free speech communication and teleconferencing systems to appear feasible. Figure 1.1 depicts a typical modern hands-free speech environment. Unlike the traditional speech recording and communication systems which require the speaker to hold a microphone or a telephone, the acoustic sources (or talkers) in the hands-free speech environment are allowed to move around in the room freely.

The acoustic arrays in the room environment usually consist of multiple microphones, which receive ambient acoustic waves emitted by the speakers as well as the other objects. In general, it is well-known that the speech signals received at microphone receivers are not only disrupted by the background noise, but also distorted severely by the room reverberation. This background noise, together with the room reverberation, cause significant difficulties in designing a speech signal processing system and thus impedes its real applications. One solution to this

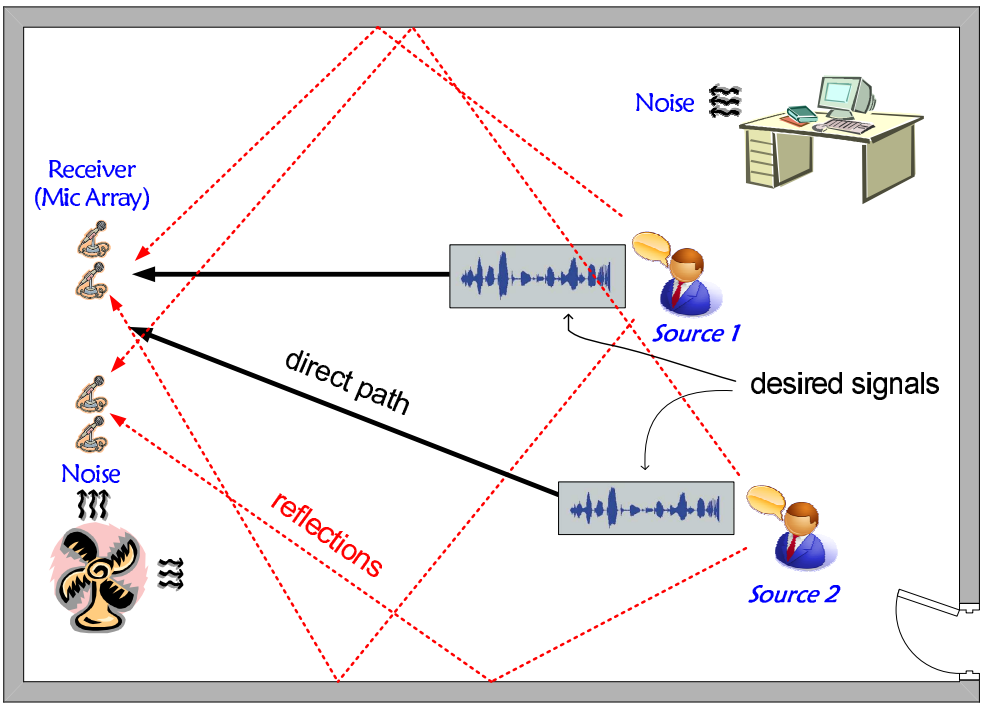


Figure 1.1: *A typical hands-free speech environment.*

problem is localising the position of acoustic source by using the received signal from the microphones first. Once the position of the speaker is known, the valuable information can then be fed into a higher processing stage for:

- high-quality speech acquisition;
- keeping a camera focused on the speaker in a video conferencing scenario;
- enhancement of a specific speech signal in the presence of other competing talkers.

For example, for a speech enhancement system, it is desired that the system is able to reduce the effects caused by the noise and the reflections, and restore the received signal to its original as much as possible. It is clear that if the information of the source position is available, beamforming can be used to extract the desired signal and suppress the noise and inferences. Further, in some cases such as a teleconferencing system, it requires that the camera always focuses on the active speaker. The source position information derived from the microphone arrays can be straightforwardly fed into such systems to orient the cameras. Obviously, the exact location information is also helpful for selective speech acquisition and source signal separation, etc.

Using the microphone receivers in ASL is attractive since the microphone receivers can be operated passively, and allow a random exploration in the field. Other advantages of microphone arrays are that they are easily implemented and cheap as compared other sensor modalities, for example, cameras. In some applications, the source moves around in the room, and the position of the source may keep changing. This further complicates the ASL problem since at each processing step, the available signal from a given position is very limited. In such cases, acoustic source tracking (AST) is preferably employed to localise the source instantaneously. This thesis is dedicated to develop a number of novel approaches to track the positions of acoustic sources in the room environment, especially when the number of the sources is unknown and time-varying.

1.2 Acoustic source localisation and tracking

1.2.1 Acoustic source position estimation

The room environment considered in our work will be a regular office room, with a medium dimension (say any enclosure smaller than or around a size of $10 \times 10 \times 3$ m^3). The temperature and moisture in the room is assumed to be stationary and their effect can be ignored when formulating the wave propagation. It is further supposed that all the microphones are omnidirectional and their positions are known and fixed, as is the sound speed (which is about 343 m/s [1] propagating in the air). The detailed information about microphone array configurations and source trajectories can be found in each experiment individually. The main issues in developing an acoustic source position estimation system are as follows.

Localisation vs. tracking. The source *localisation* is usually done in an 'open loop' way, in which the position is estimated based on the current measurements, without employing the prior information from the previous position estimates. For a speaker moving in the room, the positions are highly correlated in adjacent time steps. It is thus possible to exploit the information from both the previous position estimates and the current measurements to find the locations. In contrast, *tracking* is an approach that acts like a 'close loop' scheme, in which the position estimated from the previous step is fed back to the tracking system to initialise the position state for the next estimation. The advantage of localising the source position via tracking is that it can be used for a dynamic source and a short frame length can be employed.

[1] The variation of propagation speed in the air caused by temperature is ignored.

Large arrays vs. small arrays. As an application of the AST system in a more constrained environment such as a conference room or an office room is expected, all the experiments will be carried out using a reasonably small number of microphones. Experiments organised by researchers in [2–4] use large microphone arrays, which contain several hundred microphones. The requirement of dedicated computer architectures seriously obstacles the real application of these systems. In some special speech processing scenarios, such as the hearing aids system, it may require the localising capability of just utilising one pair of sensors. Since our solution is intended to be implemented using a real-time online system, it is desirable to minimise the number of microphones and thus reduce the computation power.

Far-field vs. near-field. The acoustic sources can either be located at the far end relative to a microphone or the close end to it. The former case is refereed as a far-field situation, in which the acoustic wave is assumed to propagate in a plane-wave form, as shown, in Fig. 1.1(a). The latter case is a near-field situation, in which the acoustic wave is assumed to propagate in a spherical-wave form. Fig. 1.1(b) illustrates the propagation of a spherical wave. The far-field assumption can seriously simplify the algorithm and system design, especially for large microphone arrays. This is because the source can be assumed at the same direction of arrival (DOA) and with a same distance relative to all the microphones.

The distance d that one can safely use the far-field assumption in the room acoustic is determined by the array separation D (also called aperture). The relation can be expressed as [5]

$$d \geq \frac{D^2 f_s}{c} \tag{1.1}$$

where f_s and c represent the sampling frequency and sound speed respectively, and d is calculated as the radial distance between the source and the center of the microphone array. For example, if the microphone separation is $D = 0.5$m, and sampling frequency is $f_s = 8000$Hz, the minimum source-microphone distance for a valid far-field assumption will be $d = 5.8$ m. For a regular office room which is smaller than $10 \times 10 \times 3\text{m}^3$, the far-field assumption is violated at almost half of the area in the room. In an enclosure with such a dimension, the far-field assumption is thus not satisfied.

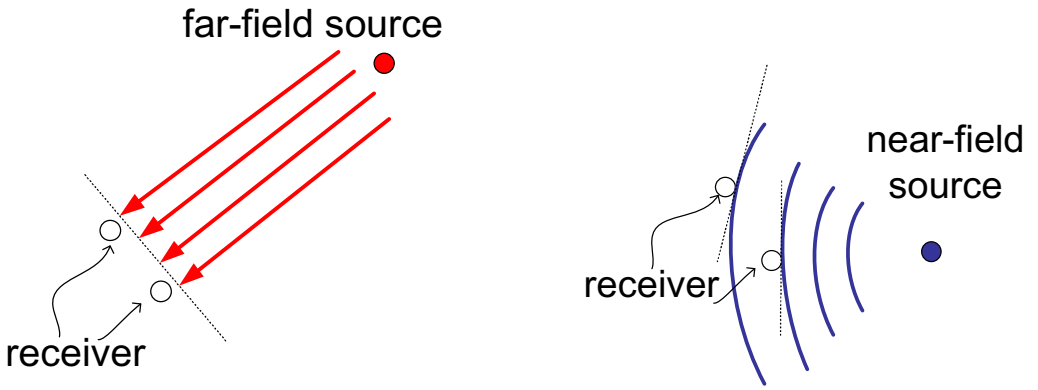


Figure 1.2: *(a) far-field source and plane wave propagation; (b) near-field source and spherical wave propagation.*

1.2.2 Elements of multiple acoustic source tracking

In real room recording systems, not only is the direct path of the speech signal picked up by the microphone receivers, but also the multi-path reflections from the walls and other objects, and the ambient noise. Due to the moving of the source, a dynamic geometry between the sources and the microphone receivers may be explored. Usually, the speech signal at far-end are significantly corrupted by the reflected components and the background noise. This causes problems for source localisation and tracking since the probability of detection is degraded and heavy false alarms may appear. Further more, in the scenario of multiple sources which are simultaneously active, the dominant background noise will be other speech signals and thus the noise level may be extremely high. In addition, tracking a moving acoustic speaker requires a short processing frame so that the position estimates can be updated without a delay. The data available at each processing step is thus limited. Since speech by its nature is temporally discontinued (the active voice and inactive voice are presented alternatively), such short frames give rise to a drastically changing between the appearance and disappearance of the source.

Due to all these facts, acoustic source tracking in a room environment is challenging, especially when considering a multiple source scenario. Generally, following three main tasks are concerned in designing a tracking system, summarised as

- *measurement extraction*, extracting the measurements from the received speech signals;
- *source dynamic modelling*, modelling the motion trajectory of the source;

- *tracking algorithm*, filtering the state of the source given the available measurements and the dynamic model.

The detailed reviews along all these three perspectives in designing an ASL system will be presented in Section 2.3, 2.4 and 2.5 respectively.

The tracking approaches can be mainly divided into two broad categories: 1) direct approaches, which use the position information as the measurements, typically extracted by employing a beamformer. Since the measurements are the position of the source already, the following tracking algorithm is only necessary to detect the sources and smooth these position measurements to obtain the trajectories; and 2) indirect approaches, which estimate the TDOAs from the received signals first, usually by using a generalised cross-correlation (GCC) method, and the source positions are then obtained by the tracking algorithms using these TDOA measurements. For indirect approaches, the source position and the TDOA measurements are with a nonlinear expression (thus the measurements indicate the source state indirectly), the tracking algorithm should be able to solve this nonlinear equation as well as estimate the source trajectories. The direct approaches has its advantage that the measurements extracted are positions, and thus the state-space equations are linear [6]. However, extracting the position measurements is usually computationally expensive. In contrast, the TDOA measurements are simple and easily available in many applications, and are widely used for either localising or tracking the acoustic sources [6–9].

1.3 State of the art

Among all the signal processing approaches introduced into the AST problem in recent years, particle filtering is a most advanced and widely used one [10, 11]. It is firstly introduced in developing an indirect tracking approach for acoustic source by Vermaak et al [12](2001), and later for direct tracking approach by Ward et al [13](2002). The particle filtering based AST approximates the true posterior distribution of the state with a set of samples, and thus requires no triangulation and linearisation of the nonlinear state space model, which are used in the conventional localisation and tracking techniques depicted in [14] and [8] respectively. The particle filtering is novel and found very promising in acoustic source tracking in that:

- The false measurements due to the reverberation and noise can be coped with a multi-

hypothesis likelihood model.

- Particle filtering is able to provide accurate estimation for nonlinear and non-Gaussian state space model.
- Robust against the effects of model mismatch.

A general framework of applying the particle filtering for acoustic source tracking problem is summarised by Ward et al [6], in which both the particle filtering for indirect and direct acoustic source tracking are discussed.

Due to its advantages of handling the nonlinear state space model and robust against the room reverberation, particle filtering is also found effective for multiple acoustic source tracking together with a random finite set (RFS) formulation. Ma et al [1](2006) developed a RFS PF approach to track an unknown time-varying number of speakers using TDOA measurements. In essence, a RFS is a finite collection of elements in which both the elements and the number (cardinality) of elements are random. It is elegant for formulating multiple source tracking problem in that a single RFS is able to capsulate the states (or the measurements) of multiple sources as well as the number of the sources (or the measurements). Similar as the indirect tracking approach in [6], a set of TDOA measurements which includes detections of multiple sources as well as false alarms is collected from each microphone pair. The source state for multiple sources is a finite set which integrates all the state vectors from the potential sources. A RFS based Bayesian filtering is then derived and the PF is employed to estimate the source positions. As the number of sources increases, the computation will be more expensive, and the first-order moment approximation approaches may be more appropriate [15, 16].

Fallon et al [17, 18] also developed a direct approach for time-varying number of multiple acoustic source tracking based on particle filtering. The tracker is based closely on the algorithm in [19], in which the surveillance region is divided into several cells, and a Bayesian filtering framework is designed to evaluate the belief in the existence of source in each of the cells. The measurements are extracted by using steered beamforming (SBF), as that used in [6]. The particle filter is with variable dimension to model different source behaviours: newly active source, source survival and source inactive.

Measurement models, source motion models, and multiple state filtering problems all add to the complexity of designing an acoustic target tracking system. Each of these perspectives is

needed to assume away many issues so that the resulting tracking system is implementable under some criterion. The main novelty of all the work in [1, 15–18] is introducing the updated signal processing techniques, particularly multi-target estimation techniques into the AST problem. However, a good measurement extraction (more source detections and less false alarms) will significantly reduce the complexity in designing the later tracker and enhance the tracking accuracy. In the next section, our development and improvements to the measurement extraction and tracking approaches will be briefly presented.

1.4 Scope of this work

This thesis addresses a number of novel AST algorithms and their applications for estimating the position of one or more dynamic speakers in the room environment. In this section, the contributions to the solutions of AST problem are firstly highlighted. An overview of the work in this thesis is then presented.

1.4.1 Contributions

As mentioned in Section 1.3, the "tracking triple", namely, measurement extraction, tracking techniques and source dynamical models will add to the complexity of designing an AST system. A successful AST system should take all these elements into account, and should be flexible for different environments and practical uses. Traditional AST systems either focus on the measurement extraction or tracking algorithm, and lack a consideration of AST problem in a whole. Since position estimation via tracking is considered here, sequential Bayesian filtering which estimates the state recursively over time steps using incoming measurements and a source dynamic model is naturally an optimum option. Also because the TDOA measurements hold a nonlinear relationship with the source states, the particle filtering which has been verified very efficient for nonlinear problems will be used in this thesis. Based on the framework of Bayesian filtering and its PF implementation, this thesis develops a number of novel approaches for multiple acoustic source tracking in room environments and contains three contributions towards solutions to multiple AST problem.

The first contribution of this thesis is proposing a TDOA estimation approach for multiple sources. Although the phase transform (PHAT) weighted generalised cross-correlation method has been employed to extract the TDOAs of multiple sources, it is primarily used for a single

source scenario and its performance for multiple TDOA estimation has not been comprehensively studied. The proposed approach combines the degenerate unmixing estimation technique (DUET) and GCC method. Since the speech mixtures are assumed window-disjoint orthogonal (WDO) in the time-frequency domain, the spectrograms can be separated by employing DUET, and the GCC method can then be applied to the spectrogram of each individual source to obtain the TDOAs for multiple sources. The probabilities of detection and false alarm are also proposed to evaluate the TDOA estimation performance under a series of experimental parameters.

Next, considering multiple acoustic sources may appear nonconcurrently, an extended Kalman particle filtering (EKPF) is developed for a special multiple acoustic source tracking problem, namely "nonconcurrent multiple acoustic tracking (NMAT)". The extended Kalman filter (EKF) is used to approximate the optimum importance weights, and the subsequent particle filtering naturally takes the previous position estimates as well as the current TDOA measurements into account. The proposed approach is thus able to lock on the sharp change of the source position quickly, and avoid the tracking-lag in the general sequential importance resampling (SIR) PF.

Finally, these investigations are extended into an approach to track the multiple unknown and time-varying number of acoustic sources. The DUET-GCC method is used to obtain the TDOA measurements for multiple sources and a random finite set based Rao-Blackwellised PF is employed and modified to track the sources. Each particle has a RFS form encapsulating the states of all sources and is capable of addressing source dynamics: source survival, new source appearance and source deactivation. A data association variable is defined to depict the source dynamic and its relation to the measurements. The Rao-Blackwellisation step is used to decompose the state: the source positions are marginalised by using an EKF, and only the data association variable needs to be handled by a PF. The performances of all the proposed approaches are extensively studied under different noisy and reverberant environments, and are favorably comparable with the existing tracking techniques.

1.4.2 Overview of this thesis

The work in this thesis is to study the difficulties of AST and develop a series of novel approaches to track the positions of multiple sources in the room environment. It can mainly be divided into two parts: an introduction of the background knowledge (Chapter 2 and 3) and our

contribution along with this problem (Chapter 4 to 6). The detailed work of the main chapters are sumarised as follows:

- **Chapter 2** firstly introduces the acoustic propagation in the room environment and the received signal model. The tracking approaches are then fully reviewed and discussed separately in terms of the "tracking triple". The simulated room environment as well as the real room environment which will be employed to implement all the experiments are then discussed.

- **Chapter 3** presents the basic concepts and implementations of Bayesian filtering, which will form a foundation of the tracking algorithms developed in this thesis. A series of Bayesian filtering approaches, from Kalman filtering to particle filtering, and also a variant of the PF, Rao-Blackwellised particle filtering, is introduced. In particular, some basic materials about the multiple target filtering in the perspective of Bayesian inference are formulated. The PF is used throughout our tracking approaches, but functioning differently; for example, in Chapter 5, it is straightly employed as a position estimator to solve the nonlinear TDOA measurement model; while in Chapter 6, since the Rao-Blackwellisation step is used to decompose the states to be estimated, the PF is employed to extract the appropriate hypothesis and refine the position estimation provided by an extended Kalman filter.

- **Chapter 4** provides the detailed experimental analysis of the performance of TDOA measurements. The first contribution in this chapter is that by defining the probabilities of detection and false alarm, the performance of PHAT-GCC method under different noisy and reverberant environments can be fully examined. Further, based on the WDO assumption of the speech mixtures, we propose a degenerate unmixing estimation technique weighted GCC (DUET-GCC) approach, which is more appropriate for TDOA estimation of multiple simultaneously active sources. The received speech mixtures are separated by employing DUET, and the TDOAs for each source can thus be estimated from the corresponding source signal individually. The TDOA performance of these approaches are extensively studied in different adverse environments.

- **Chapter 5** mainly considers a special case of multiple source tracking problem, non-concurrent multiple source tracking. In a number of scenarios, multiple speakers may appear alteratively within a room environment; one speaker is active for a period, and

then another follows. This special case requires the algorithm to follow sharp changes in a position and lock on the new speaker. The general sequential importance resampling particle filtering approach fails to do so since the importance function only takes the position information estimated at the previous step into account and without considering the innovations from current measurements. The particles are thus not drawn effectively. Although the EKF is able to catch up the sharp change, it can not incorporate a reverberant measurement model as the general SIR-PF does. Our contribution in this chapter is combining the extended Kalman filter and particle filter, namely the extended Kalamn particle filtering, to solve this problem. The core idea is that by employing an EKF, one can obtain the optimum importance function, and then sample the particles based on this importance function. The particles are thus drawn in a more relevant area than simply using a prior density function, which is typically a sampling scheme used in general SIR-PF. The tracking performance is demonstrated in the different simulated room environments as well as the real room environment.

- **Chapter 6** focuses on the highest hierarchy of the acoustic source tracking problem: a time-varying number of acoustic sources. In such case, the number of the sources as well as their positions are unknown *a priori*. The tracker developed is based on the RBPF data association technique. The original source position states together with an association variable, are used to represent the source position and the source dynamics: birth, survival or death. In practice, it is quite often that both the measurement extraction approaches fail to report the TDOA measurements across some sensors. A death process that allows a measurement missing at some microphone pairs are particularly designed. To reduce the estimation variance and sampling efficiently, a Rao-Blackwellisation technique is employed, by which the position states are marginalized by using an EKF, and only the association variable is needed to be handled by a PF. Rather than the traditional data association approaches which use a heuristic technique to prune or determine the hypothesis, the proposed particle filtering data association approach theoretically admits a random hypothesis-pruning. Unlike the existing multiple source tracking approaches which are only tested in the simulated room environment, both the different simulated room environments and the real room experiments are organised to fully examine the tracking performance of our approaches.

Finally, all the results are summarised and some conclusions are drawn in **Chapter 7**. Suggestions for future research directions are also presented.

Chapter 2
Background knowledge on room acoustic source tracking

This chapter presents a review of the various basic materials relevant to the room acoustic source tracking problem. The illustration of the room acoustic propagation model and received signal model is described first in Section 2.1. A broad range of tracking systems in terms of different measurement extraction approaches are then discussed. Some periphery techniques which are able to further enhance the tracking ability are also introduced. At the end of the chapter, a description of the simulated room environment as well as the real recording environment are presented.

2.1 Room acoustics

To gain a basic knowledge how the acoustic waves propagate in the room, the direct path and multi-path propagation model is illustrated first. The room impulse response (RIR) and its length, reverberation time are then introduced. Some assumptions are listed here to simplify the propagation model:

- *The propagation medium is homogeneous, nondispersive and lossless*. The assumptions of homogeneous and nondispersive dictate that the propagation speed of sound c is constant everywhere in the closure, and it does not vary with frequency. The lossless assumption further indicates that the attenuation of the wave energy in the room is equal everywhere.

- *The Doppler effect is negligible*. The source can be moving around in the room, but its speed is far less than the speed of sound. Hence, it is not necessary to take the Doppler shift in frequency into account.

- *The microphones are identical and omnidirectional*. The microphones are assumed to be identical and omnidirectional. This assumption will simply the tracking problem since

it is unnecessary to consider the microphone scaling problem. The actual frequency response of microphones is shown in Fig. A.1, in Appendix A.1, on page 192.

2.1.1 Acoustic wave propagation

In free space, sound wave propagates without any interference from the objects. Suppose a single source emits an acoustic signal, $s(t)$, at the position $\mathbf{x}$, and the position of the microphone receiver is $\boldsymbol{p}$. The direct path impulse response is [20, 21]

$$h_{\mathrm{d}}(\mathbf{x}, \boldsymbol{p}, t) = \frac{a}{r}\delta(t - \tau), \tag{2.1}$$

where $r = \|\mathbf{x} - \boldsymbol{p}\|$ ($\|\cdot\|$ stands for Euclidean distance) is the distance between the source and the microphone receiver, and $\tau = r/c$ is the time delay. $\delta(t-\tau)$ is a delta-Dirac function which equals one when $t = \tau$ and zero otherwise. The attenuation factor is inversely proportional to the distance from the source, and the constant a is used to model the effect of the medium and the system gain. Based on the assumptions proposed at the beginning of this section, the acoustic propagating in a closure can be modelled as a linear system. The output signal will be the original source signal convolved with the impulse response. The direct path wave which arrives at the receiver can thus be expressed as

$$\begin{aligned} z_{\mathrm{d}}(\mathbf{x}, \boldsymbol{p}, t) &= s(t) \star h_{\mathrm{d}}(\mathbf{x}, \boldsymbol{p}, t) \\ &= \frac{a}{r}s(t - \tau), \end{aligned} \tag{2.2}$$

where $s(t-\tau)$ is a delayed version of the original source signal. All the localisation and tracking approaches rely on this direct path component since it parameterises the distance between the source and the microphone receivers.

In a room environment depicted in Fig. 1.1, the sound waves are reflected by the objects in the room and the walls. Usually the walls of most rooms are reflective enough to generate significant reverberation, and such reverberation is much easier to be simulated than the reflections from other objects. Hence, in our simulated reverberant environments, only the wall reflections will be considered and all other reflections will be ignored. Suppose the wall reflection impulse response is h_{r}. The signal generated by the wall reflections is

$$z_{\mathrm{r}}(\mathbf{x}, \boldsymbol{p}, t) = s(t) \star h_{\mathrm{r}}(\mathbf{x}, \boldsymbol{p}, t), \tag{2.3}$$

The complete wave propagates to the microphone receiver is

$$\begin{aligned} z(\mathbf{x}, \boldsymbol{p}, t) &= s(t) \star (h_{\mathrm{d}}(\mathbf{x}, \boldsymbol{p}, t) + h_{\mathrm{r}}(\mathbf{x}, \boldsymbol{p}, t)) \\ &= \frac{a}{r} s(t - \tau) + s(t) \star h_{\mathrm{r}}(\mathbf{x}, \boldsymbol{p}, t). \end{aligned} \tag{2.4}$$

which is a summation of the direct path propagation and the reverberations.

2.1.2 Room impulse response

When the geometry between the source and microphone is determined, the wave propagation is globally characterised by the room impulse response (RIR). As depicted in the previous section, the complete RIR $h = h_{\mathrm{d}} + h_{\mathrm{r}}$ is the combination of the direct path response and the wall reflection response. Fig. 2.1 gives an illustration of the RIR, which typically consists of the following three parts:

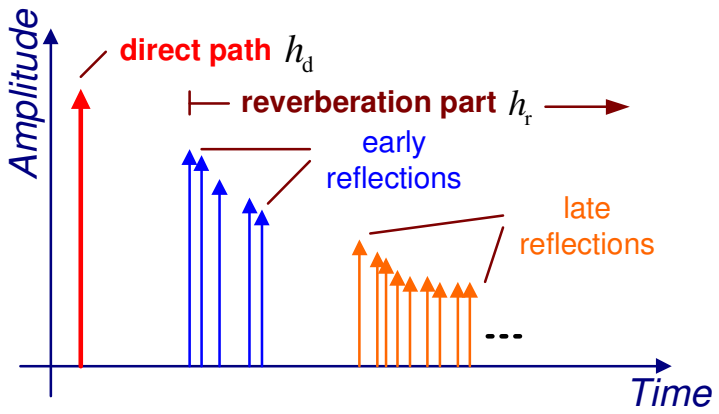


Figure 2.1: *Illustration of the RIR.*

- *Propagation time for direct path*, the time needed for acoustic propagation from the source to the microphone along the shortest path.
- *Early reflections*, the first several individual low-order reflections of the direct signal off the surfaces in the room.
- *Late reflections*, the high-order reflections that decay exponentially in time and characterise the room's reverberation.

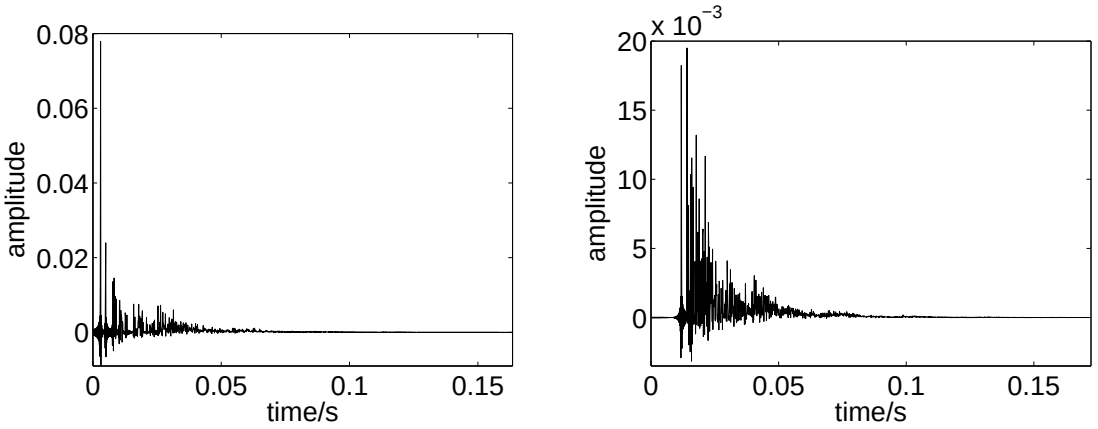


Figure 2.2: *RIRs in a room with the room dimension* $5 \times 4 \times 3$ m^3 *and microphone position at* $[0.50 \;\; 1.75 \;\; 1.70]$ *m. The average value of the wall reflection coefficients is* $\rho = 0.6$. *The source is located at (a) close-end* $[0.90 \;\; 0.85 \;\; 1.70]$ *m; (b) far-end* $[4.25 \;\; 3.25 \;\; 1.70]$ *m.*

In speech signal enhancement, the early reflections and late reflections are dealt differently: the former is actually able to enhance the speech signal, and only the later is needed to be reduced. However, for the localisation and tracking problem, both these two components will significantly deteriorate the location estimation performance since only the direct path impulse response has a contribution to the problem.

For the tracking problem, since the geometry between the source and the microphone is changing all the time, the RIR varies even in the same room environment. Fig. 2.2 presents two different RIR in the same room environment. The room dimension and the position of the receiver are fixed, but the source is active at different positions: one is located at close-end, and the other far-end. The RIR varies significantly due to the change in the source position: the direct path response is relatively strong at the close-end, but may be very weak at the far-end locations. This causes difficulties to the tracking problem since even in the same room environment, the speech signals are deteriorated at different level; high direct signal to reverberation ratio at the close-end, and vice versa. If the walls are with a strong reflectivity, only the the received signals from the sources which are close to the microphone receiver are reliable.

Given a certain room environment, the RIR can easily be simulated by using the image method [22]. Rather than tracing all the reflections from the walls, the image method creates an image for the source with respect to the each room boundary. The received signal is thus simply the

summation of a large number of direct path components propagating from the image sources to the microphone. The order of the image is determined by the number of reflections that may occur. The RIR is usually modelled using a finite impulse response (FIR) filter, which may have hundreds or even thousands filter taps. The infinite impulse response (IIR) filter can also be used to reduce the filter order. However, such order reduction is very limited, and the filter still needs several hundreds taps [23]. Further, inverting a RIR is not readily possible since it has been shown that the RIR is generally a non-minimum phase system [24].

2.1.3 Reverberation time

The length of the RIR is determined by the reverberation time T_{60}, which is defined as the time taken for the sound pressure level to decay by 60 dB of its original level after the source signal has been switched off. Different formulas have been developed to estimate the reverberation time. Eyring's expression for T_{60} is given by [20]

$$T_{60} = \frac{-0.163V}{\mathcal{A}\ln(1-\bar{\alpha}) + 4mV}, \tag{2.5}$$

with

$$\bar{\alpha} = \frac{1}{\mathcal{A}}\sum_i \mathcal{A}_i\alpha_i, \tag{2.6}$$

where V in m^3 is the room volume, and $\mathcal{A}_i$ in m^2 is the reflection area of the ith wall, and $\mathcal{A}$ is the whole reflection area, i.e., $\mathcal{A} = \sum_i \mathcal{A}_i$. m is the intensity attenuation constant, and the term $4mV$ is related to the air absorption which can usually be neglected for small rooms. The corresponding absorption coefficient α_i holds a relationship with the wall reflection coefficients ρ_i as

$$\alpha_i = 1 - \rho_i^2. \tag{2.7}$$

By expanding the logarithm term in (2.5) into a series, and ignoring all the higher terms except the first order, Sabine [20] further simplifies the expression for T_{60} as following,[1]

$$T_{60} = \frac{0.163V}{\sum_i \mathcal{A}_i\alpha_i}. \tag{2.8}$$

Equations (2.5) and (2.8) are widely used to estimate the reverberation time in the room envi-

[1]The higher terms can be neglected since, in practice, it is always safe to assume that the average absorption coefficient $\bar{\alpha}$ is small compared with unity [20].

ronment. Suppose an office room consists of a carpet floor (heavy on concrete), and concrete block walls and ceiling (painted), where the absorption coefficients are 0.6 and 0.09 respectively [21].[2] For a small office room with a dimension of $5 \times 4 \times 3$ m^3, the Eyring's and Sabine's reverberation time T_{60} will be 0.470s and 0.524s respectively. For a larger enclosure with a dimension of $10 \times 10 \times 3$ m^3, the Eyring's and Sabine's reverberation time T_{60} will be 0.533s and 0.613s respectively.

Other reverberation formulae are inclusively discussed in [20]. Although the math expressions are different from each other, the main difference can be summarised into one: the manner how the absorption coefficients of the various areas of wall are averaged. The same conclusion can be drawn from all these reverberation time expressions; the larger and less absorbent the wall surfaces, the longer the decay time. In this thesis, the length of RIR of the simulated room environment will be calculated by Sabine's reverberation time T_{60}.

2.2 Signal model

In this section, the microphone received signal model will be formulated. Since our work is concentrated on the multiple source tracking problem, the multiple source signal model is introduced here directly. The single source signal model can be obtained by simply setting the number of sources to one. Taking all multi-path propagation components into account will make the estimation problem extremely complicated. The reverberant signal model is thus simplified into a free-field one, from which the direct path time delays can be pulled out explicitly.

2.2.1 Recording model

The recording signals in a room environment can be described as follows. Let $\boldsymbol{p}_{\ell,i}$ and $\mathbf{x}_{m,t}$ denote the position of ith microphone in the ℓth microphone pair and the position of mth source at time t, respectively. The signals generated by simultaneously active sources can be modelled as a summation of the multiple individual source signals in the room environment. Suppose

[2]The frequency band of acoustic signal considered here is 2000Hz. For same materials, the absorption coefficients vary with different frequency band.

there are totally M_t sources, the discrete time signal at time instance t can be written as

$$z_{\ell,i}(t) = \sum_{m=1}^{M_t} s_m(t) \star h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t}, t) + \bar{v}_{\ell,i}(t), \tag{2.9}$$

where $s_m(t)$ and $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t}, t)$ are the mth source signal and the corresponding room impulse response (RIR) respectively. The RIR term $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t}, t)$ has been fully illustrated in the Section 2.1.2, on page 15. The noise term $\bar{v}_{\ell,i}(t)$ consists of the contributions from the following two parts:

- Unknown noise sources $\tilde{s}_j(t)$. This noise term may be generated by other speakers or objects such as ventilator, fans and footsteps. Suppose that the corresponding position of the noise source is $\tilde{\mathbf{x}}_j(t)$. This part of contribution can be written as

$$\bar{v}^1_{\ell,i}(t) = \sum_j \tilde{s}_j(t) \star h(\boldsymbol{p}_{\ell,i}, \tilde{\mathbf{x}}_{j,t}, t), \tag{2.10}$$

  where the summation is taken over the number of all noise sources.

- Independent channel noise $\bar{v}^2_{\ell,i}(t)$. This part of noise is usually generated by the microphone receiver or other equipment of the recording.

Both these two noise terms are assumed to be independent and uncorrelated with the source signals and across different microphone receivers.

2.2.2 Free-field model

As shown in the propagation model depicted in Section 2.1.1, on page 14, only the direct path component contains the necessary delay information which contributes to the localisation and tracking solutions. Following equation (2.4), the impulse response can be decomposed into the direct path and multi-path components as

$$\begin{aligned} z_{\ell,i}(t) &= \sum_{m=1}^{M_t} \frac{1}{4\pi r^m_{\ell,i}(t)} s_m(t - \tau^m_{\ell,i}(t)) + \sum_{m=1}^{M_t} s_m(t) \star h_{\mathrm{r}}(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t}, t) + \bar{v}_{\ell,i}(t) \\ &= \sum_{m=1}^{M_t} \frac{1}{4\pi r^m_{\ell,i}(t)} s_m(t - \tau^m_{\ell,i}(t)) + v_{\ell,i}(t), \end{aligned} \tag{2.11}$$

where $s_m(t-\tau_{\ell,i}^m(t))$ is the pure delayed source signal; $h_{\mathrm{r}}(\boldsymbol{p}_{\ell,i},\mathbf{x}_{m,t},t)$ is the reverberation part of the room impulse response (RIR); $r_{\ell,i}^m(t)=\|\mathbf{x}_{m,t}-\boldsymbol{p}_{\ell,i}\|$ is the distance between the source and microphone; $\tau_{\ell,i}^m(t)=r_{\ell,i}^m(t)/c$ is the direct path time-delay with c representing the speed of sound. The main difference between the free-field model and the recording model is the noise term $v_{\ell,i}(t)$. Other than the two parts of noise included in $\bar{v}_{\ell,i}(t)$, the new noise term $v_{\ell,i}(t)$ includes the contribution from the reverberation:

- known source noise generated by the multiple reflections of the original source signals, which can be stated as

$$\bar{v}_{\ell,i}^3(t)=\sum_{m=1}^{M_t} s_m(t)\star h_{\mathrm{r}}(\boldsymbol{p}_{\ell,i},\mathbf{x}_{m,t},t), \tag{2.12}$$

The noise term $\bar{v}_{\ell,i}^3(t)$ here is constructed by many replications of the original source signals, and is therefore correlated with the source signals. This will violate the independence and Gaussian assumptions about the noise term in many signal processing cases. The free-field model simplifies the expression of the time delay parameter. However, it is only appropriate in the low and moderate reverberation scenarios. Since, in such scenarios, the replications of the source signal are relatively weak, the assumptions of independency and Gaussian about the noise term are still satisfied.

2.2.3 Speech processing basics

The speech signal itself is non-stationary and its statistics change over time. In practice, the speech signals are usually split into small consecutive intervals and one can thus assume that the signals in these intervals are stationary. Hence the signal received at each microphone are processed in frames. Let T and k denote the length of the frame and the time index of the frame respectively. The source signal and the signal collected at the ith microphone of ℓth pair can then be written as

$$\begin{aligned}\mathbf{s}_m(k)&=[s_m(kT),s_m(kT+1),\ldots,s_m(kT+T-1)]\,,\\ \mathbf{z}_{\ell,i}(k)&=[z_{\ell,i}(kT),z_{\ell,i}(kT+1),\ldots,z_{\ell,i}(kT+T-1)]\,.\end{aligned} \tag{2.13}$$

Further we assume that in such frames, the position of the source is stationary as well. All the parameters about the source are thus fixed at the kth frame, e.g., the number of sources M_k,

source position $\mathbf{x}_{m,k}$, and the corresponding room impulse responses.

The frequency domain representation of each frame can be obtained by applying the discrete-time Fourier transform (DTFT), stated as

$$\begin{aligned} Z_{\ell,i}(k,\omega) &= \mathcal{F}\{\mathbf{z}_{\ell,i}(k)\} \\ &= \sum_{m=1}^{M_k} \frac{1}{4\pi r_{\ell,i}^{m}(k)} e^{-j\omega\tau_{\ell,i}^{m}(k)} S_m(k,\omega) + V_{\ell,i}(k,\omega), \end{aligned} \tag{2.14}$$

where ω is the discrete frequency index; and Z, S and V are the DTFT of the received signal $\mathbf{z}$, source signal $\mathbf{s}$, and a frame of noise term $\mathbf{v}$ respectively. Cascading all these DTFTs leads to the discrete short time Fourier transform (STFT) of the whole received speech signal.

The power spectral density for the received signal and the cross power spectral density across the ℓth microphone pair can be computed as

$$G_{\ell,i}(k,\omega) = \mathbb{E}\{|Z_{\ell,i}(k,\omega)|^2\}, \tag{2.15}$$

$$G_{\ell}(k,\omega) = \mathbb{E}\{Z_{\ell,1}(k,\omega)Z_{\ell,2}^{*}(k,\omega)\}, \tag{2.16}$$

where the superscript $*$ denotes the conjugate, and $\mathbb{E}$ is the expectation operator. From the Wiener-Khintchine theorem [25], it is known that the autocorrelation function and cross-correlation function are the inverse Fourier transform of the power spectral density and cross power spectral density respectively, given as

$$R_{\ell,i}(k) = \mathcal{F}^{-1}\{G_{\ell,i}(k,\omega)\}, \tag{2.17}$$

$$R_{\ell}(k) = \mathcal{F}^{-1}\{G_{\ell}(k,\omega)\}. \tag{2.18}$$

The transfer function in the frequency domain across the ℓth microphone pair is defined as the ratio of their DTFTs,

$$H_{\ell}(k,\omega) = \frac{Z_{\ell,1}(k,\omega)}{Z_{\ell,2}(k,\omega)}. \tag{2.19}$$

2.3 Measurement extraction

In terms of microphone array based localisation and tracking, the measurements extracted from the received signal can be mainly divided into two categories: location measurement and TDOA

measurement. The former is regarded as direct information since it directly shows the source position. This category of measurement is typically extracted from beamforming methods, such as steered beamforming [13,26]. To localise and track the source, since the position of the source is coarsely represented by the measurements, the task of later locator/tracker is simply to smooth and refine the position estimates. The TDOA measurement contains the position information in an indirect way; it usually requires the following locator/tracker to be capable of solving a nonlinear relationship between the TDOA measurement and the source positions.

In this section, the steered beamforming based direct measurement extraction approach is firstly presented. The TDOA measurement extraction approaches which have been used in the tracking problem are then introduced. These approaches include the generalised cross-correlation, phase unwrapping, and adaptive eigenvalue decomposition approaches. Other measurements such as interaural level difference and joint audio and video information are also briefly reviewed.

2.3.1 Direct measurements: position extraction

Steered beamforming is a direct method applied to the acoustic localisation and tracking problem. It can be regarded as a steered response obtained from the output of a delay-and-sum beamformer. It was shown in the free-field model (2.11) that each received signal is actually a delayed and noise corrupted version of the original speech signal. The delay-and-sum beamformer compensates the direct path delay of the received signal and gathers them together to preserve the original signal from a spatial location. It is defined as [26,27]:

$$y(\mathbf{x}, k) = \sum_{\ell,i} \mathbf{z}_{\ell,i}(k + \tau_{\ell,i}), \tag{2.20}$$

where $\mathbf{z}_{\ell,i}(k)$ is the received signal, and the summation is taken over all the received signals. $\tau_{\ell,i}$ is the steering-delay which is normally defined as a delay relative to a reference point, given as

$$\tau_{\ell,i} = c^{-1}(\|\mathbf{x} - \boldsymbol{p}_{\ell,i}\| - d_{\text{ref}}), \tag{2.21}$$

with $\mathbf{x}$ and $\boldsymbol{p}_{\ell,i}$ denoting the steered position and sensor position respectively. The reference distance d_{ref} is the distance between the steered position and some reference position, which is typically chosen as the center of the sensor array. The steering-delay $\tau_{\ell,i}$ is thus the time deference between a steered position and the reference position.

Given the DTFT version of the received signal $Z_{\ell,i}(k)$, the delay-and-sum beamformer can be written in the frequency domain as

$$Y(\mathbf{x}, k) = \int_{\Omega} \sum_{\ell,i} \left| W_{\ell,i}(k,\omega) Z_{\ell,i}(k,\omega) e^{j\omega\tau_{\ell,i}} \right|^2 d\omega, \tag{2.22}$$

where $W_{\ell,i}(k,\omega)$ is a frequency weighting term, which is used to compensate the noise effects and is dependent on the individual signals. The phase transform (PHAT) weight $W_{\ell,i}(k,\omega) = (|Z_{\ell,i}(k,\omega)|)^{-1}$ is chosen to normalise the contribution of each frequency component, and has been proved advantageous for practical situations where the ideal filters are unavailable [18]. The frequency range Ω over which the integration is implemented is extensively discussed in [18]. It is shown that only little benefit can be gained for frequencies above 2kHz, and thus there is no significant localisation information that can be obtained in such frequency bands. This is true since the speech information is rich in the lower formant area, and very sparse at the higher frequency band.

The location estimate is found by maximizing the output power of the beamformer in a potential location space, given as

$$\hat{\mathbf{x}}_k = \underset{\mathbf{x}}{\operatorname{argmax}}\, Y(\mathbf{x}, k). \tag{2.23}$$

The source position can be estimated by implementing a multidimensional search over the position vector space. For the case that the noise term in the free-filed model is additive, uncorrelated and with a uniform variance, SBF is able to indicate the source position in the position space by a sharp peak. Fig. 2.3 shows the SBF response from a speech frame. The source position $[2.0, 2.5]$m is indicated by a peak obviously. However, in the reverberant environment, these assumptions are always violated by the convolutional channel effects and the correlated noise, and the performance of SBF is thus degraded.

Strobel et al [26] fully present a SBF algorithm for acoustic source localisation problem. The authors also show that the performance can be further enhanced by incorporating a speech pause detector (SPD). A particle filtering framework with SBF measurements is developed in [6], in which SBF presents better tracking performance in dealing with reverberation than TDOA measurements. Based on this PF framework, Lehmann et al [28] also implement SBF particle filtering approach with a voice activity detector (VAD) and further present several different importance sampling schemes [29]. Although localisation and tracking based on SBF is simple and robust in the moderate noisy and reverberant environment, the potential computation can be

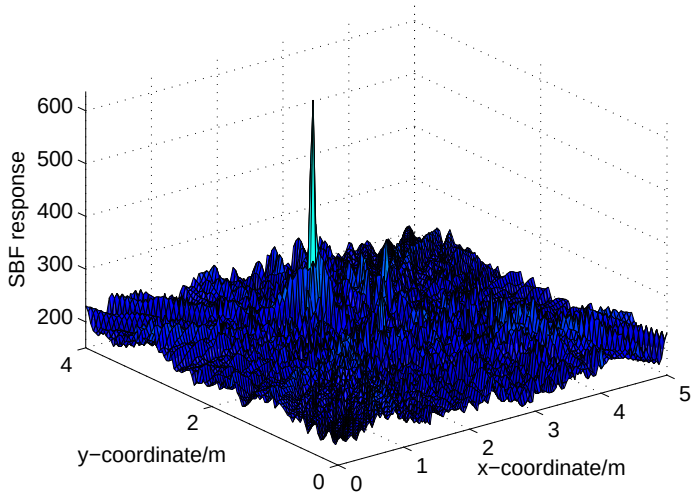


Figure 2.3: *SBF response from a frame of speech signal. The integration frequency range is 300 to 3500Hz. The true source position is at [2.0, 2.5]m. The grid density is set to 0.04cm, and the integration in equation (2.22) will be calculated for* $125 \times 100 = 12500$ *times.*

very demanding to achieve a high location accuracy, and thus impedes its practical applications.

Valin et al [30] introduces a beamforming based approach to track simultaneously moving sound sources. With a data association consideration and particle filtering implementation, the algorithm is able to track a time-varying number of multiple sources. However, the algorithm does not really track the position of the sources since it can only tell the direction of the sources. More sophisticated tracking based on beamforming can be found in Fallon's work [18], in which the number of sources can also be time-varying. But the sources are supposed to move extremely slow in the room environment (in the experiments, the sources are moving at a velocity less than 0.15m/s, which is very slow compared to the moving velocity around 0.6m/s in [1]).

2.3.2 Indirect measurements: TDOA extraction

TDOA measurement has attracted a considerable amount of attention in that it can be obtained easily. A large number of TDOA measurement estimation techniques have been developed to estimate TDOAs in anechoic environments as well as in the reverberant and noisy environment. Here we only fully present several approaches such as generalised cross-correlation, phase

unwrapping and adaptive eigenvalue decomposition due to their popularity in the localisation and tracking problem.

2.3.2.1 GCC method

Given the speech frames $\mathbf{z}_{\ell,1}$ and $\mathbf{z}_{\ell,2}$ collected at ℓth microphone pair at time step k, the cross-correlation can be approximated as

$$R_\ell(k,\tau) \approx \int_\Omega Z_{\ell,1}(k,\omega) Z^*_{\ell,2}(k,\omega) e^{j\omega\tau} d\omega, \tag{2.24}$$

where Ω is the frequency range over which the integration is carried out. Similar as the integration frequency range in equation (2.22), only the frequency range which is rich in speech signal will contribute significantly for the cross-correlation function. If the received signal satisfies the free-field model (2.11), it can be regarded as a noise corrupted time-delay version of the original source signal. The Fourier transform can be written as

$$Z_{\ell,i}(k,\omega) = e^{-j\omega\tau_{\ell,i}} S(k,\omega) + V_{\ell,i} \qquad i = 1,2, \tag{2.25}$$

and the cross-correlation can thus be expressed as

$$R_\ell(k,\tau) \approx R_{ss}(k,\tau-\tau_\ell) + R_{v_1v_2}(k,\tau), \tag{2.26}$$

where R_{ss} is the autocorrelation of the source signal and $R_{v_1v_2}$ is the cross-correlation of the noise components, $\tau_\ell = \tau_{\ell,2} - \tau_{\ell,1}$ is the time-delay between two microphones. Since the noise terms are assumed to be independent, we have $R_{v_1v_2} = 0$ and the maximum of the cross-correlation function will appear at $\tau = \tau_\ell$. The TDOA measurement at the ℓth microphone pair can thus be estimated by exploring the time-delay value τ that maximizes the cross-correlation function

$$\hat{\tau}^\ell_k = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_\ell(k,\tau), \tag{2.27}$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay which can only happen when the microphone pair and the source lie exactly on an extended line.

An approximation of the cross-correlation, the generalized cross-correlation (GCC) function is stated as [31]

$$R_\ell(k,\tau) = \int_\Omega \Phi_\ell(k,\omega) G_\ell(k,\omega) e^{j\omega\tau} d\omega, \tag{2.28}$$

where $G_\ell(k,\omega)$ is the cross power spectrum density defined in equation (2.16), and $\Phi_\ell(k,\omega)$ is a weighting term. Different choices of the weighting term $\Phi(k,\omega)$ have been inclusively studied in [31]. The phase transform (PHAT) weighting term which is extensively used in the localisation problem and found robust in the noisy and reverberant environments is employed throughout our work. The details of applying this weighting term will be discussed in Chapter 4.

The GCC method is proposed by Knapp and Carter [31] in 1976 and found to be the most popular method in extracting the TDOAs in the last decades. It introduces a pre-filtering processor to enhance the peak from the true delay, and the effects of the noise and reverberation can be suppressed. Azaria et al [32] further studied the problem of estimating time delay by cross correlation methods for the whole class of stationary signals. Due to its easy implementation and effectiveness in TDOA extracting in noisy and reverberant enviroments, GCC method is also extensively used in the source localisation systems [33–35]. By considering the property that the strength of excitation in voiced speech is large around the glottal closure instant, Yegnanarayana et al [36] implement the GCC method based on the Hilbert envelope of the linear prediction (LP) residual, rather than based on the speech frames directly. The TDOA measurement extracted from this approach is further used to localise the source in [37], in which the localisation error is proved to be consistently equal or less than the GCC method.

The performance of the GCC method in different reverberant environments (versus different reverberation time T_{60}) is extensively studied in [38]. It is shown that GCC method is reliable to report the TDOAs in the moderate reverberant environment. However, as the reverberation increases to a certain level (as shown in [38], $T_{60} \geq 0.5$s), it will collapse abruptly. It is worth mentioning that all these studies are based on the static case, i.e, the source is stationary at a position in the room, and the performance of GCC method in extracting the TDOA measurements of multiple sources is still unknown. In Chapter 4, the GCC based TDOA performance for dynamic source, and particularly, for multiple simultaneously active sources will be studied. Further, multiple nonconcurrent and concurrent source tracking in the room environment based on GCC TDOAs will be fully investigated in Chapter 5 and Chapter 6 respectively.

2.3.2.2 Phase unwrapping

TDOA can also be extracted in the frequency domain by unwrapping the phase term of the cross-spectrum across a microphone pair. The phase of the cross-spectrum can be expressed as

$$\begin{aligned}\phi_\ell(k,\omega) &= \arg(Z_{\ell,1}(k,\omega)Z^*_{\ell,2}(k,\omega)) \\ &= \omega\tau_\ell + \epsilon(k,\omega),\end{aligned} \tag{2.29}$$

where $\epsilon(k,\omega)$ is a phase error term summarizing the contributions of the noise to the phase and the inaccuracy from the analysis window. Equation (2.29) shows that the TDOA τ_k^ℓ is actually the slope of the line fitting all the phase terms. The least square (LS) solution of τ_k^ℓ can be written as [25]

$$\begin{aligned}\tau_k^\ell &= \arg\min \|\omega\tau_\ell - \phi_\ell(k,\omega)\|^2 \\ &= \frac{\sum_\omega \omega\Phi_\ell(k,\omega)\phi_\ell(k,\omega)}{\sum_\omega \omega^2\Phi_\ell(k,\omega)},\end{aligned} \tag{2.30}$$

where $\Phi_\ell(k,\omega)$ is a weighting term calculated from the variance of the phase error $\text{var}(\epsilon(k,\omega))$, given as

$$\begin{aligned}\Phi_\ell(k,\omega) &= \frac{1}{\text{var}(\epsilon(k,\omega))} \\ &= \frac{\lambda_{\ell,1}(\omega)}{|Z_{\ell,1}(k,\omega)|^2} + \frac{\lambda_{\ell,2}(\omega)}{|Z_{\ell,2}(k,\omega)|^2},\end{aligned} \tag{2.31}$$

where $\lambda_{\ell,i}(\omega)$, $i = 1, 2$ is the average noise power at the frequency point ω usually calculated using a short frame of background noise.

It must be pointed out that the phase discontinuity may happen when applying this TDOA estimator. The phase estimates $\phi_\ell(k,\omega)$ evaluated in (2.29) is modulo 2π, whereas the linear estimator (2.30) requires a phase angle that varies in a continuous linear fashion along the frequency bins. Hence, the phase unwrapping step must be applied. Several approaches have been proposed for this purpose, and a typical one can be found in Tribolet's work [39]. Normally the phase estimates are very unlikely to be unwrapped at the low frequency band. The initial TDOAs are thus estimated from these low frequency phase components $\phi_{k,\omega}^{\text{low}}$. The prediction of the phase for the higher frequencies can be estimated using these initial estimates and the linear prediction. The phase extracted at the higher frequency band should be $\phi_{k,\omega}^{\text{high}} = \omega\tau_\ell + 2\kappa\pi$.

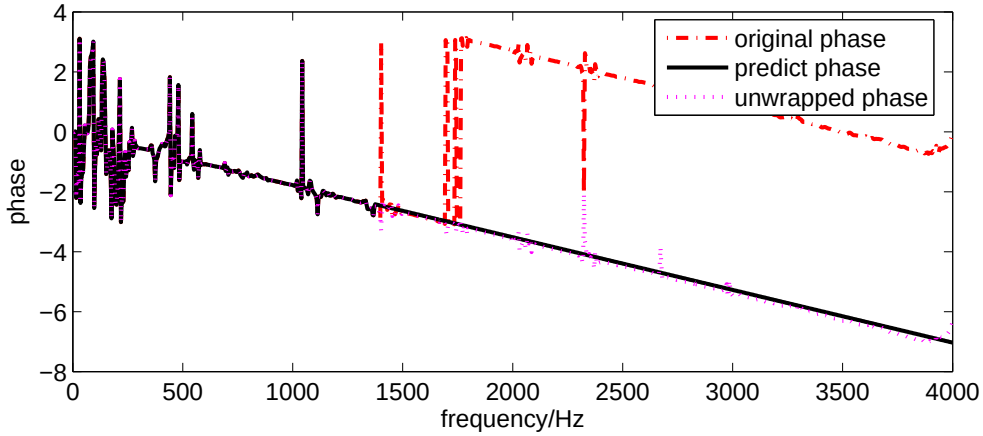


Figure 2.4: *Phase unwrapping. Phases at the higher frequency band ($f > 1715Hz$) are wrapped. The unwrapped phases at lower frequency band are used to predict the phase term at the higher frequency band, and the wrapped phase are adjusted accordingly.*

Supposing that the predicted phase is $\phi_{\text{pred},\omega}$, the integer κ should satisfy

$$\kappa = \arg\min_{\kappa} |\phi_{\text{pred},\omega} - \phi_{k,\omega}^{\text{high}} + 2\kappa\pi|. \tag{2.32}$$

The final phase estimates is the combination of the estimates from the low frequency band $\phi_{k,\omega}^{\text{low}}$ and high frequency band $\phi_{k,\omega}^{\text{high}}$. The TDOA is thus obtained by re-estimate the slope by taking all the phase estimates into account.

Figure 2.4 gives an example of the wrapped phase term and the phase after the unwrapping process. A frame of speech signal and two microphones with a separation of 0.1m are used to generate the received signals. The maximum unwrapped frequency is

$$f_{\text{max}} = \frac{c}{2d} = \frac{343}{2 \times 0.1} = 1715\text{Hz}. \tag{2.33}$$

Detailed derivation of the relationship between the maximum unwrapped frequency and the microphone separation is given in Section 4.4.2, on page 100. For the frequencies $f > 1715$Hz, the phase terms are wrapped and the above unwrapping technique is employed to obtain meaningful phases. Finally, all the phase terms are following a linear fashion and the TDOA can be estimated from the slope of the line.

The phase unwrapping approach is found popular in the localisation problem [7, 40, 41] due

to its features of low computational requirements and high updating rate. In contrast to GCC methods which explores a search method to extract the TDOA estimates, there is no time-resolution limitation problem. The sub-sampling accuracy of TDOAs can be achieved by the analytical model (2.30). However, in practice, excessive separation between the microphone pair may lead a significant degradation of the TDOA estimates. This is mainly because the unequal signal attenuation at the near-field and phase ambiguity may happen at a high frequency band. Since the work in this thesis is developing the approaches for multiple source tracking, the phase unwrapping cannot be applied straightforwardly. In Chapter 4, it will be employed to develop the DUET-GCC method to extract the TDOA measurements for multiple sources when the phase ambiguity happens.

2.3.2.3 Adaptive eigenvalue decomposition

Another well-known TDOA estimation algorithm is the adaptive eigenvalue decomposition (AED) [42], in which the TDOAs are extracted directly from the impulse responses between the source and microphones.

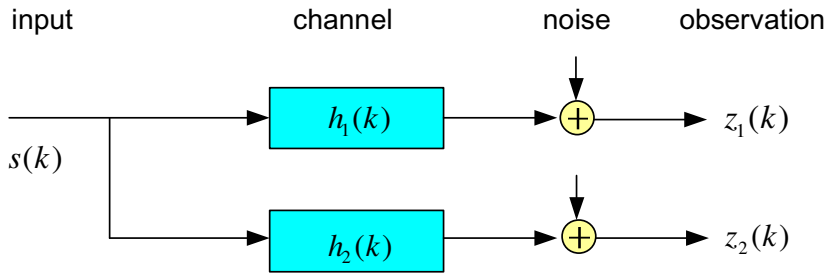


Figure 2.5: *Illustration of the channel model.*

Let $s = s(t)$ denote the source signal and h_i, for $i = 1, 2$ represents the channel response to the ith microphone. The received signal z_i follows the relationship

$$z_1 \star h_2 = s \star h_1 \star h_2 = z_2 \star h_1. \tag{2.34}$$

The detailed channel model for this expression is shown in Fig. 2.5.

For the microphone pair signal model in (2.11), ignoring the background noise, and simply replacing the time domain signal in (2.34) with signal frames, we have

$$\mathbf{z}_{\ell,1}^T(k) \star \mathbf{h}_{\ell,2}(k) = \mathbf{s}(k) \star \mathbf{h}_{\ell,1}(k) \star \mathbf{h}_{\ell,2}(k) = \mathbf{z}_{\ell,2}^T(k) \star \mathbf{h}_{\ell,1}(k), \tag{2.35}$$

where superscript T denotes the transpose of a vector, and $\mathbf{h}$ is the impulse response at the same length with the signal, defined as

$$\mathbf{h}_{\ell,i}(k) = [h_{\ell,i}(kT), h_{\ell,i}(kT+1), \ldots, h_{\ell,i}(kT+T-1)]\,, \qquad i = 1, 2. \tag{2.36}$$

The covariance matrix of the ℓth microphone pair is

$$\mathbf{R}_\ell = \begin{bmatrix} \mathbf{R}_{1,1} & \mathbf{R}_{1,2} \\ \mathbf{R}_{1,2} & \mathbf{R}_{2,2} \end{bmatrix}, \tag{2.37}$$

with $\mathbf{R}_{i,j} = \mathbb{E}\{\mathbf{z}_{\ell,i}(k)\mathbf{z}_{\ell,j}(k)\}$, $i, j = 1, 2$. Define a $2T \times 1$ vector

$$\mathbf{u}_\ell = \begin{bmatrix} \mathbf{h}_{\ell,2} \\ -\mathbf{h}_{\ell,1} \end{bmatrix}. \tag{2.38}$$

The following equation can be achieved

$$\mathbf{R}_\ell \mathbf{u}_\ell = \mathbf{0}. \tag{2.39}$$

This means that the covariance matrix $\mathbf{R}$ has a single eigenvector $\mathbf{u}$ (which contains two impulse responses) corresponding to the eigenvalue 0 conditional on no common zeros in $\mathbf{u}$ and full rank of the matrix $\mathbf{R}$. If the free-field model is fulfilled and the noise term can be ignored, the final TDOA is simply the time difference of the two indices corresponding to the two peaks from the eigenvector $\mathbf{u}$.

In practice, however, estimating the eigenvector $\mathbf{u}$ is not a trivial task since the performance is affected by the content of the speech, the length of the impulse response, and the reberberation and background noise, etc. Following the constrained least mean square (LMS) algorithm [43], Huang and Benesty et al [42, 44] proposed an adaptive algorithm to estimate the vector iteratively. Different from the GCC method which only considers the direct path, AED deals with time-delay estimation problem based on the whole impulse response and includes the multipath process into the model. It is thus expected that AED approach will provide more accurate and robust TDOA estimates than the GCC method [42]. Doclo et al [45] further present two generalised adaptive eigenvalue decomposition algorithms for the time-delay estimation in the large amount noisy and reverberant environment.

While the AED approach is found a great interest in the acoustic localisation problem, its application in the source tracking is very limited. This is mainly because it needs a burn-in period to converge to a good TDOA estimate. On the other hand, the authors in [44] mentioned that the the AED is capable of catching up with the true TDOAs in less than 250 ms, which is tolerable for most localisation applications. Its application in acoustic source tracking is still a problem since it cannot follow a dynamical TDOAs quickly and one has to control the convergence of the measurements and the tracker both.

2.3.3 Interaural level difference

The signals received by the microphone receivers not only differ in their time-delay difference, but also in their attenuation level. This attenuation level difference information, in some applications referred as interaural level difference (ILD), forms the basis of directional sense of human hearing [46], but has received much less attention in the localisation and tracking field. Suppose the source energy received at two microphones of ℓth microphone pair are $E_{\ell,1}$ and $E_{\ell,2}$ respectively. From the signal model in equation (2.11), a simple relationship between the source energies and distances can be expressed as [47]

$$E_{\ell,1} r_{\ell,1}^2 = E_{\ell,2} r_{\ell,2}^2 + \epsilon_\ell, \tag{2.40}$$

where r represent the corresponding distance between the source and the microphone receiver, and ϵ_ℓ is the measurement noise term.

The acoustic source localisation using the ILD measurements is similar as a lot of acoustic energy measurement based localisation scheme in wireless sensor networks [47, 48]. Equation (2.40) shows that the ratio of signal strength between two microphone receiver is inversely proportional to the ratio of the distance to the microphone receivers. This signal strength ratio actually generates a location circle, and by intersection of these circles, the source position can be estimated. Fig. 2.6 illustrates the different location scheme based on TDOA and ILD in a two dimensional space: while the TDOA indicates the potential locations along a line, the ILD measurement uses a circle to represent the potential source locations, unless the two energies are equal (by which a location line will apply).

The interaural level difference has been shown an important cue for the room acoustic source localisation system. Birchfield et al [49] first introduce the ILD cue for a computer based local-

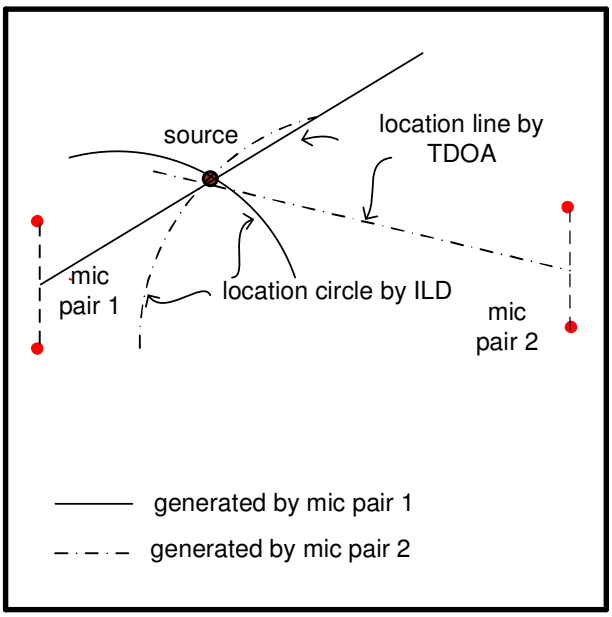


Figure 2.6: *Localisation scheme of TDOA and ILD. TDOA represents the potential source position on a line, and ILD exploits a circle to represent the potential source position.*

isation system, in which the source position is estimated by combining the likelihood functions from several microphone pairs. By employing both the TDOA and ILD information, a closed-form solution for the localisation problem is developed in [50] only using one microphone pair. However, it requires prior information of evaluation regions to obtain the final rational solution. The authors in [51] have also developed an energy based source localisation system for ad hoc microphone arrays.

Another interesting perspective in the binaural localisation and tracking systems is using the joint TDOA and ILD information as the measurements [52]. A Bayes rule based localisation system is proposed in [53], in which the TDOA is extracted by exploiting the PHAT-GCC method, and ILD is calculated at different frequency band. Based on the Window-disjoint-orthogonality (WDO) assumption [54], the binaural cues of each source can be calculated from the ratio of two-channel time-frequency (TF) representation of the received signals. Roman et al [55] develop a hidden Markov model (HMM) tracking algorithm using the binaural cues calculated from the TF representation. The likelihood function is evaluated by integrating the probabilities across reliable frequency channels, and the HMM is employed to detect the number of sources and track the azimuths. Localising source based on a more elaborate model of the relation between azimuth angle and binaural cues can be found in [56]. However, under the reverberant environment, the TF spectrogram is smeared due to the multiple reflections

of the original signal. These measurements extracted from the TF domain are thus degraded significantly.

2.3.4 Joint audio and video measurements

In an audio-video conferencing, there are usually two measurement modalities available: sound signals and video images. A number of algorithms have been developed to fuse these two modalities to track the source. Since our work is focused on the tracking problem with the acoustic signals, these approaches are not going to be used in this work but only the current researches along this perspective are briefly reviewed .

As mentioned in Section 1.1, on page 1, the microphone receiver can be operated passively, and a localisation and tracking system using it is able to allow a random exploration in the room. It is thus good for initialisation of the source position, where vision is relatively expensive. The video information has its advantage for localising the source in that it is free of the room reverberation and it can avoid dynamic occlusions between the acoustic sources. Tracking applying these two modalities jointly is thus able to complement the drawback of each other and enhance the localisation performance. Vermaak et al [57] build a sequential Monte Carlo fusion framework for speaker tracking using joint audio and image modalities. The utilising of the sound signal is based on the TDOA measurements extracted from the GCC method, and a standard approach for visual tracking is applied to the vision modality. The TDOA likelihood and image likelihood are constructed for the corresponding modality, and a particle filter is applied to fuse the TDOA and image measurements. It is also shown in [57] that the sound information is able to provide the position initialisation, and helps considerably with the algorithm recovering from the tracking loss.

By exploiting a small microphone array and multiple uncalibrated cameras, a similar tracking system which fuses 2-D object shape and audio information via importance particle filters is proposed in [58]. A real-time speaker tracker applying particle filter sensor fusion is proposed in [59], in which a novel sensor fusion framework combining both the bottom-up and top down approaches is developed to probabilistically fuse the multiple modalities. Asoh et al [60] use the joint audio and video information to track multiple sound sources. Instead of the general importance particle filtering as in [57], the authors in [61] present a Markov chain Monte Carlo (MCMC) particle filtering approach to jointly track the location and speaking activity of multiple speakers in a meeting room. Due to employing a multi-person dynamical model and a high

sampling efficiency by the MCMC particle filtering approach, it can deal with cases of visual clutter and occlusion and significantly outperforms the traditional sampling-based approaches in [59].

2.4 TDOA based tracking approaches

Amongst such a big number of measurement categories introduced above, the TDOA measurement is the favorite one for either acoustic localisation or tracking due to its own advantages of simplicity. Further, since the TDOA measurement can be estimated by simply exploiting a pair of microphones, it is easily available in many applications. In recent years, a lot of different tracking approaches have been developed to deal with the TDOA based tracking problem, in the single source scenario as well as the multiple source scenario. In this section, a full review of the tracking approaches in terms of these two tracking scenarios will be presented.

2.4.1 Single acoustic source tracking

The single source tracking problem based on TDOA measurements has been extensively investigated in recent years [6–9]. It is usually operated in an indirect way: the TDOAs from microphones are extracted by, for example, firstly employing the generalised cross-correlation (GCC) function [31], and then using a tracker to triangulate the source position based on these

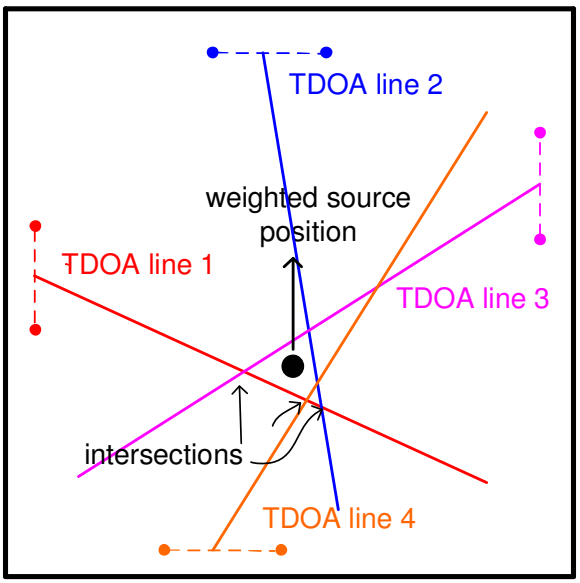


Figure 2.7: *Illustration of the linear intersection approach.*

TDOA measurements. Brandstein et al [27] addressed a maximum likelihood criterion for such triangulating. Since each TDOA can indicate the orientation of the source relative to the microphone pair, the linear intersection (LI) method is employed to estimate the position of the source [14]. Fig. 2.7 shows the intersection of the orientation lines formulated by four TDOAs. Normally the intersection of two TDOA lines is able to report the source position, but multiple TDOAs can be incorporated to enhance the position estimates by weighting the different intersections according to the likelihood of the TDOA measurement.

Since the adjoint positions of a moving source in the room are highly correlated, temporal information available from these adjoint positions is also important and useful to obtain a robust localisation estimates. LI estimator is actually a localisation (not a tracking) approach since it only takes spatial information into account at each time step, without considering the temporal information. Bayesian ASL estimation techniques [6, 8, 9, 12] exploit both the temporal and spatial information to estimate the source position. The temporal information are obtained from the previous source position estimates and the source dynamical model. Klee et al [8] introduce a series of Kalman filters into the speaker tracking problem. The EKF as well as the iterative Kalman filter (IKF) is formulated to handle the nonlinearity of the TDOA measurement function. It is found that these tracking algorithms are able to provide more accurate location results compared with the linear intersection techniques. Gannot et al [9] further employ the unscented transform (UT) to track the source and have proved the advantages of the proposed tracking approaches.

If there is merely a time delay between the received signals, and the TDOA estimates have a Gaussian distribution, the Kalman filtering approaches [8, 9] are able to present satisfactory tracking results. However, the TDOA measurements are often violated by reverberation and different kinds of noise in real-life. The tracking algorithms are thus deteriorated since the inaccuracy from the first-stage measurement extraction is not accounted for by the following location estimator. Recently, particle filtering is introduced into the acoustic source tracking problem to reduce the errors brought by the false TDOA estimates mainly caused by the multi-path reverberant components [6, 12]. The TDOA measurement in the reverberant environment is modelled by a bi-modal distribution (Gaussian for the TDOAs generated by the real source and uniform distribution for those generated by clutter). This reverberant measurement model will be fully introduced in Section 5.2.3, and will be used as the measurement model to develop the EKPF tracking approach in Chapter 5. The PF is completely a nonlinear filter in that it esti-

mates the posterior distribution of the state directly, which is achieved by using a large number of samples representing the distribution of the states and going through a Bayesian estimator. Generally, the error of approximating the TDOA measurement function in EKF can be avoided. A full description of this method can be found in [12] and [6], and a real-time implementation of this approach is demonstrated in [62].

2.4.2 Multiple acoustic source tracking

In the single source tracking problem, it is assumed that there is one and only one source existing throughout the entire tracking period, and all the measurements are either generated by this source or by clutter. Although the approaches [28, 62] have taken silent intervals into account, the silent intervals are assumed short and the number of sources during the tracking period is assumed known and fixed to one. In a wide range of applications such as surveillance, multimedia conferencing, and selective speech enhancement, however, the system is required to localise multiple sources simultaneously. Unlike the single source tracking approaches, in which the Bayesian theory based approaches such as Kalman filter and sequential Monte Carlo implementation are overwhelmingly employed, a benchmark framework appropriate for multiple source tracking is still missing.

The multisensor multi-target filterings have been extensively applied in the sonar and radar target tracking systems [63–66], but less widespread in the field of room acoustics due to the complexity of the tracking approaches as well as the room environment. The same as a general multitarget tracking, the main tasks for a multiple acoustic source tracker are: 1) assigning the measurements to the corresponding sources appropriately; 2) filtering the position of the states individually from the noise-corrupted measurements. The former task is usually performed by a data association technique such as probability data association (PDA) or multiple hypothesis tracking (MHT) [63, 64]. The later is achieved by employing a Bayesian filtering approach such as Kalman filter [67] (or its variants extended Kalman filter [67] and unscented Kalman filter [68]), or particle filtering method [11].

Sturim et al [69] introduce several Kalman filters running parallel to obtain the potential position states, and an interacting multiple model (IMM) approach is used to fuse these states and match them with the individual talkers. A more elaborate IMM based multiple moving speaker tracking system is proposed in [41]. By jointly employing a probabilistic data association (PDA) technique and IMM estimator, the authors develop a general statistical framework that

allows the system to fuse the state estimates obtained by the IMM estimator. This multisensor multitarget technique is later introduced into the speech separation of multiple moving speakers in [70]. By using the probabilistic-data-association technique in conjunction with the IMM estimator, the DOA from each source can be determined. The receptive beams are then formed to lock on each moving speaker. Therefore, the voice separation can be achieved. Joint probabilistic data association (JPDA) filter is also employed to track multiple simultaneous speakers in [71], in which the JPDA is able to make use of several measurements and discriminate them from real sources or clutters. However, all of these approaches suffer from either one or several of following drawbacks:

- a time-varying number of sources cannot be applied;
- the results are presented for a single trial or even just for a single step due to a lack of consideration of new measures which are appropriate for multiple sources. The performance of many runs is thus unknown;
- the tracking errors such as tracking loss and divergence from the actual trajectories are not fully analyzed.

The first one narrows the realistic applications of the algorithms since it is quite common that sources are active dynamically, e.g., one source is active first, and another one joins in and overlaps the former speech for a while. The other two drawbacks make the algorithms incomparable between each other or with other tracking approaches.

Very recently, an updated multisensor multitarget tracking approach, using random finite set (RFS) statistics, has been introduced into the multiple acoustic source tracking problem [1, 15, 16, 72]. A batch of measurements obtained at current time step are used to formulate a combination of TDOA based likelihood, and the states are estimated by a particle filtering similar with that in the single source case [6, 12] without a specific source-measurement assignment. The birth and death process is firstly employed to allow a time-varying number of source during the tracking period. The advantages of this category of algorithms are: 1) the recursively random set approach of estimating the multiple states is tractable and can be done by employing fast online particle filtering; 2) the approaches have decoupled the association and tracking problem, and focus on the locating the acoustic sources but without discriminate the measurements, which is able to sharply reduce the computational complexity. The Bayesian RFS filter [1] is only appropriate for small number of multiple source tracking (say maximum two) since as the

number of sources increases, it becomes expensive to implement. In such cases the probability hypothesis density (PHD) filter [15, 72] is found more tractable because it propagates only the first moment of the multitarget posterior. A brief introduction of Bayesian RFS filter will be presented in Section 3.4.1, on page 70, and a novel RFS based approach for tracking multiple time-varying number of acoustic sources will be fully introduced in Chapter 6.

2.5 Peripheral techniques

Besides the measurement extraction and the tracking approach itself, modelling the source dynamics is also an important component in constructing a complete tracking system. In this section, the choice of the source dynamical model, and peripheral techniques such as voice activity detection which also contributes to the tracking performance will be discussed.

2.5.1 Motion dynamical models

For a general target tracking problem, many different dynamical models have been developed to model the source motion trajectories [73]. An exhausted summary of the motion models can be found in the recent survey [74]. In recent decades, a significant research effort has been devoted to develop the tracking algorithms for acoustic source tracking problem: LI technique in [7]; EKF in [8, 9]; PF in [6, 12]; and more recently, RFS statistics for multiple acoustic source tracking in [1, 15, 16, 72]. Compared to such a broad range of investigations on tracking approaches, the influence of the source dynamical models on the tracking accuracy has received much less attention. Usually researchers assume that the affection of dynamical model is trivial and can be ignored, which is only the case when the source motion is slow-paced and the trajectory is simple. Until very recently, a number of dynamical models for acoustic source tracking have been investigated [75, 76].

The dynamical models investigated in [75] can be divided into two categories: coordinate-uncoupled (CU) models and curvilinear (CL) models. The former uses a Cartesian coordinate to represent the target's velocity. The state vector can be typically defined as

$$\mathbf{x}_k \triangleq [x_k, y_k, \dot{x}_k, \dot{y}_k]^T, \tag{2.41}$$

with x_k and $\dot{x}_k$ denoting the position and the velocity component toward the $x-$coordinate.

This type of dynamical model assumes that the $x-y$ coordinates are uncoupled. In most speaker tracking scenarios, the height of the speaker is assumed to be fixed, and only the $x-y$ two dimensional tracking problem is considered. This assumption is plausible since in practice, the speaker rarely changes the height, and particularly, it is much easier to organise the real recording experiments in $x-y$ plane. For a coordinate-uncoupled model, the expression of the current source position can be given as

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \Delta T \begin{bmatrix} \dot{x}_k \\ \dot{y}_k \end{bmatrix}, \tag{2.42}$$

where x_{k-1} is the x-coordinate at last time step, and $\Delta T \dot{x}_k$ is the displacement during a time period ΔT along the $x-$direction.

The curvilinear model uses a polar coordinate system to represent the target's velocity, which is constructed by a magnitude and an orientation angle. Suppose $\boldsymbol{v}_k$ is the velocity component, and define

$$v_k = |\boldsymbol{v}_k| \qquad \text{and} \qquad \varphi_k = \angle \boldsymbol{v}_k, \tag{2.43}$$

with $|\cdot|$ and $\angle\cdot$ representing the magnitude and angle operation respectively. The formulation of the state vector can be stated as

$$\mathbf{x}_k = [x_k, y_k, v_k, \varphi_k]^T. \tag{2.44}$$

With this model, the updating of the source position can be given as

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} + \Delta T v_k \begin{bmatrix} \cos(\varphi_k) \\ \sin(\varphi_k) \end{bmatrix}. \tag{2.45}$$

The coordinate-uncoupled model (2.42) and the curvilinear model (2.45) are the simplest expressions in each model category since only a velocity component is taken into account. More complicate models can be constructed by further cascading an acceleration component. By considering the different combination with motion velocity or acceleration component, five different transition models whose implementation in acoustic source tracking problem is deemed promising or of some interest, are investigated in [75]. Rewrite them here:

- CU model with random walk velocity (CU-RWV);

- CU model with time-correlated acceleration (CU-TCA);
- CU model with Langevin dynamics (CU-LAN);
- CL model with random walk velocity and random acceleration (CL-RWV-RA);
- CL model with random walk velocity and acceleration (CL-RWV-RWA).

The difference between the last two models is that the acceleration component in CL-RWV-RA is simply modelled by an additive noise term, but in CL-RWV-RWA, it follows a random walk model. The CU-LAN is the most popular models for speaker tracking since it is simple and generally accurate enough to model many different kinds of motion [6, 8, 9, 12].

The parameters in these models are generally optimised by experimental study [75]. The tracking performance does not solely rely on a specific model type, but also the model parameters. The further research in [76] along this investigation presents an online parameter optimisation approach for various models in the particle filtering tracking framework. How to choose a model to optimise the tracking result is out of our research scope in this thesis. In our experiments, the Langevin model is preferably chosen for the following reasons:

1. for a simple motion trajectory, it is sufficient to present a satisfied tracking performance;
2. the Langevin model is the most popular model for the acoustic source tracking problem, and choosing which will reduce the work when comparing the tracking performance between different tracking algorithms.

2.5.2 Voice activity detection

One practical issue when processing the speech signal is that a lot of silence intervals may appear between utterances. Due to the weak source signals in such silence gaps, correct measurements (no matter whether the position estimates or TDOAs) are almost impossible to obtain. The erroneous measurements will misguide the tracking algorithms and make the tracking algorithm unstable. A nature way to alleviate this tracking instability is to fuse a voice activity detector (VAD) into the tracking algorithm.

A study which integrates VAD estimations into the particle filtering framework in a probabilistic sense is proposed by Lehmann et al [28]. The VAD estimations are obtained by thresholding

the instantaneous SNR[3] of the current frame of speech signal. This is based on the assumption that during the active period the signal level is sufficiently higher than the background noise level. The VAD estimations are then fed into the tracking algorithm as a prior to model the probability that the measurement is originated from the real source or not. The general idea behind this method is that if the current frame of the speech signal is active, the measurement should be more reliable and particles should be sampled relying on the current measurements rather than drifting from the source dynamical model, and vice versa. Experiment studies show that integrating the VAD is able to avoid the turbulence from those unvoiced frames and such method is more suitable for real-world implementations than the tracking approaches without a VAD [28, 62].

The authors in [61] also build a VAD into their joint audio and video tracking system. Unlike traditional approaches in which the VAD is operated based upon signal energy, SNR or spectrogram, the speech/silence frames are evaluated based on short-term clustering of the localisation results. This is based on the assumption that location estimates are normally consistent during the speech frames, while location estimates at noisy frames will present high variations over time. Only the cluster lasts more than a predefined period is regarded as an important one and labeled as speech. However, the VAD is included in a hard decision way, not in a full probabilistic sense as used in [28, 62].

The aim of incorporating a VAD into the tracking system is to discriminate the voiced and unvoiced speech signals, and subsequently highlight those reliable measurements generated by voiced speech signals. Voice activity detection in the noisy and reverberant environments itself is still a difficult problem and worth a lot of investigation. It is always preferable that the tracking algorithm is able to differentiate those measurements which are from active sources and which are not. This aim can be achieved by exploiting a data association methods, e.g., a gating technique [73]. Further, VAD may be able to report the exist of source/sources effectively, but not for the number of multiple simultaneously active sources and cannot tell which source is active. In the multiple source scenario, the received signal may be always active since multiple speech signals are mixed together and the received signal is less sparse in the time domain. For all these reasons, the application of VAD in the multiple source tracking problem is very limited and complicated, and thus will not be used in this thesis.

[3]The instantaneous SNR is calculated by the ratio of the power between the current signal and the background noise [28].

2.6 Experiment environment

A method for simulating RIR is of great help for testing the proposed approaches in speech enhancement and source tracking. To fully examine the tracking performance of our tracking approaches, it is desired to implement them in the simulated noisy and reverberant environment as well as in the real room environment. In this section, the image method which is widely used to generate a reverberation will be firstly presented. A virtual room with controllable wall reflexibility is then introduced to simulate different reverberant environments. Finally, the real audio lab and the experiment system will be illustrated explicitly.

2.6.1 Reverberation by image method

The image method has been widely used in many room acoustic simulations. It was first described in [22] to simulate the impulse response of a single microphone in a rectangular room, and then extended for microphone arrays by applying an additional low-pass filter to each impulse response [77].

The theory of image method is that the reflection of the wall can be simulated by placing an image symmetrically on the far side of the wall in three dimensional, as seen in Fig. 2.8. Assuming the source position and the microphone position in the room are $\mathbf{x}$ and $\boldsymbol{p}$ respectively, and the room dimension is $L_x \times L_y \times L_z$ m^3. The room impulse response with the nonrigid

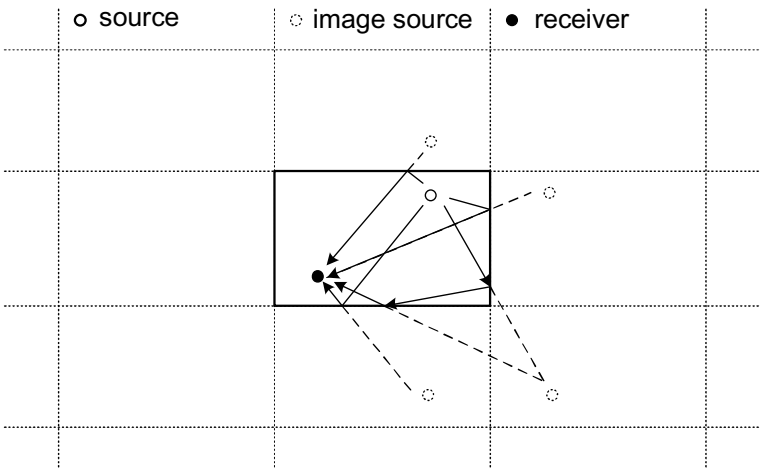


Figure 2.8: *A two dimensional representation of imaging method. The solid rectangular denotes the original room, and reflections are constructed by the direct path of those image sources.*

walls can be expressed as

$$h(\mathbf{x}, \boldsymbol{p}) = \sum_{\mathbf{q}=\mathbf{0}}^{\mathbf{1}} \sum_{\mathbf{r}=-\infty}^{\infty} \rho_{x1}^{|n-i|} \rho_{x2}^{|n|} \rho_{y1}^{|l-j|} \rho_{y2}^{|l|} \rho_{z1}^{|m-k|} \rho_{z2}^{|m|} \frac{\delta[k - |R_{\mathbf{q}} + R_{\mathbf{r}}|/c]}{4\pi |R_{\mathbf{q}} + R_{\mathbf{r}}|},$$

where x, y, z represent the spatial dimensions, and subscript 1 and 2 denote the walls adjacent and opposite respectively. The subscript $\mathbf{q} = (i, j, k)$ and $\mathbf{r} = (n, l, m)$ are in terms of a 3-integer vector. $R_{\mathbf{q}}$ is expressed as

$$R_{\mathbf{q}} = \mathbf{x} - \boldsymbol{p} + 2\mathbf{q}\boldsymbol{p}, \tag{2.46}$$

and $R_{\mathbf{r}}$ is the extension of room dimensions

$$R_{\mathbf{r}} = 2[nL_x,\ lL_y,\ mL_z]. \tag{2.47}$$

The source energy decay is determined by the distance between the image source and the receiver as well as the wall reflection coefficients. For a source fixed at some position, the larger the wall reflection coefficients, the slower the source energy decays, and thus a larger reverberation time T_{60} is expected. In the same room environment, the further the source away from the microphone receiver, the closer the impulse amplitudes will be between the direct path component and the next reflection components. Two example of RIR generated by using the image method is illustrated in Fig. 2.2, on page 16. In this thesis, all the simulated reverberant environments will be obtained by using the image method.

2.6.2 Simulated room environment

Figure 2.9 shows a simulated office room with a dimension of $5 \times 4 \times 3\text{m}^3$. Four microphone pairs each with a separation of 0.5m are organised around the center of the walls. As mentioned in Section 2.5.1, only the two dimensional tracking problem is considered to simplify the experimental study, so that the height of the microphones and sources are assumed to be known and at the same height of 1.7m. The reverberation in the room is simulated using the image method described in the previous section.

The source motion trajectories follow two diagonal lines: one from bottom left to top right; the other from top left to bottom right. Two speakers appearing at different time step will formulate different number of sources. The speed of source moving is set around 0.5m/s (1.8km/h), which

is one third of the pedestrian walking speed of younger individuals (with a velocity ranging from 5.32km/h to 5.43km/h [78]). Considering that the source moving in a room is normally smooth and slow-paced, this experimental speed is reasonable, and also comparable with the source velocities in [1, 6]. To achieve this moving velocity, 50 frames of speech signal with a frame length of 128ms are used, at a sampling frequency of 8000Hz. Since the distance of the trajectory is about 4.3m, this setup will lead to a moving velocity of $4.3/(50 \times 0.128) = 0.67$m/s.

Different wall reflection coefficients are set to generate different reverberant environments. Given the wall reflection coefficients, the corresponding reverberation time T_{60} can be calculated using equation (2.8), on page 17. Fig. 2.10 presents the relationship between the reflection coefficients ρ and the reverberation time T_{60}. To know the exact reverberation time under different reflection coefficients, a series of wall reflection coefficients and the corresponding reverberation time T_{60} are also presented in Table 2.1. The noisy environments are defined by the signal-to-noise ratio (SNR). Different noisy environments are simulated by adding different

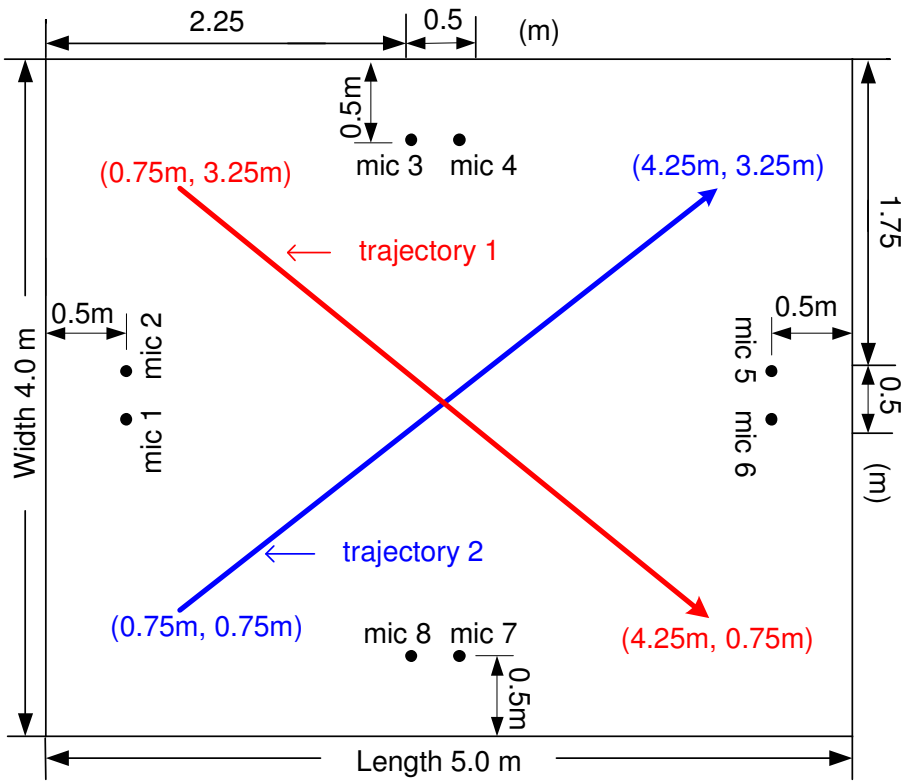


Figure 2.9: *Simulated room environment. Black dots numbered 1 to 8 denote the microphone positions, the solid lines represent the trajectories.*

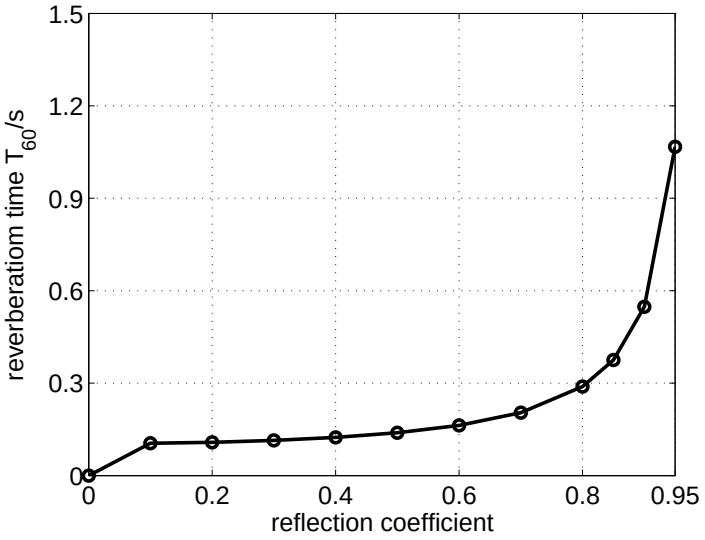


Figure 2.10: *The relationship between the reflection coefficients ρ and the reverberation time T_{60}.*

level of Gaussian white noise (GWN) into the original speech signals; see detailed definition of SNR in Appendix B.1, on page 195.

2.6.3 Real room environment

The algorithms developed in this thesis will also be implemented in a real audio lab environment. Fig. 2.11 shows the sketch of the real room recording environment. The dimension of the lab is $8.1 \times 5.3 \times 3\text{m}^3$. Four microphone arrays each with five microphones are employed to receive the speech signals. 16 microphone pairs are thus available. The separation of each two adjacent microphones is 0.45m. An omni-directional loud speaker is used to generate source signals. Two source trajectories are generated: trajectory 1 from the left bottom corner to up right corner, and trajectory 2 from the up left corner to the right bottom corner. The microphones and the sources are set at a same height of 1.33m.

ρ	0	0.1	0.2	0.3	0.4	0.5
T_{60} (s)	0	0.105	0.108	0.114	0.124	0.139
ρ	0.6	0.7	0.8	0.85	0.90	0.95
T_{60} (s)	0.163	0.204	0.289	0.375	0.548	1.067

Table 2.1: *Different wall reflection coefficients ρ and corresponding reverberation time T_{60}.*

All the received signals are sampled with a system default sampling frequency of 44.1kHz. However, for a tracking problem, a sampling frequency of 8kHz is found enough to cover the potential position information. The signals are thus resampled at 8kHz to achieve a computation efficiency. To better know the ground truth of trajectories, the sources move following diagonal lines, as shown, in Fig. 2.11. The frame length is set to 1024 samples, i.e., $1024/8000 = 0.128$s. We aim to move the source as smooth as possible, and with a velocity around 0.5m/s.

The detailed specifications of recording systems are presented in Appendix A. The microphone response is omni-directional within the frequency range 0 to 4kHz, as shown in Fig. A.1, on page 192. All the microphone gains are set to 0dB. The reverberation time T_{60} of the audio lab is roughly 0.836s, and the noise level is -40dB. The detailed analysis of reverberation level and noise level of the audio lab will be given in Section 4.6.1, on page 110.

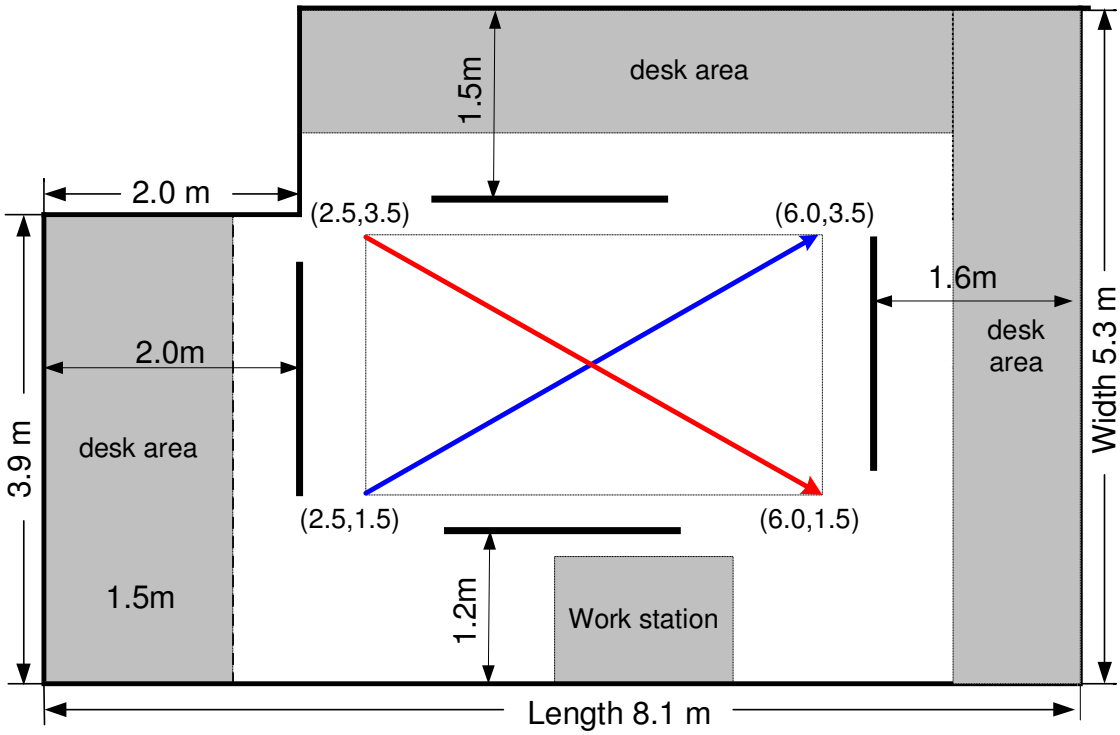


Figure 2.11: *Real audio room environment. Four microphone arrays (thick dark line around the room) each with five microphones are organised in the room to receive the speech signals. The sources are moving like diagonal line trajectories.*

2.7 Chapter summary

In this Chapter, a full range of background knowledge for the room acoustic source localisation and tracking problem have been introduced and discussed. The problem is a challenging one due to a noisy and reverberant environment. An accurate solution to this problem will not only greatly help a series of speech processing applications, but also benefit other tracking scenarios such as radar and sonar target tracking problems.

Since Bayesian filtering has its advantages to incorporate the prior information and ability to find the posterior probabilities automatically, it is naturally an optimum choice to use it to capture the position of a moving source in the room. This thesis will focus on deriving the tracking algorithms from the Bayesian filtering point of view, instead of utilising a localisation approach. In Chapter 3, the concept of the Bayesian filtering, and in particular, a number of its recursive implementations such as Kalman filtering and particle filtering will be presented.

Chapter 3
Sequential Monte Carlo Methods

Chapter 2 introduces the acoustic source tracking problem and reviews a number of tracking techniques. In this chapter, basic concepts and implementation of Bayesian filtering are discussed, which will form a foundation of the tracking algorithms developed in this thesis. The chapter begins with a brief introduction of the Bayes's theorem, and also the recursive Bayesian estimator. Theoretically, the Bayesian filtering is able to extract the posterior probability density function (pdf) exactly. However, in general, the closed form solution can rarely be derived. The Kalman filter is expected to provide an analytical and optimum solution when the system models are linear and noise are Gaussian. For nonlinear models, the extended Kalman filtering, and more sophisticated technique, unscented Kalman filtering can be used.

However, these methods still assume the Gaussianity of the noise term. In many practical problems, the noise term is non-Gaussian together with nonlinear models (may be multimodal). Under such circumstances, particle filtering which is completely appropriate for the nonlinear and non-Gaussian scenario, is a better choice to solve the state posterior estimation problem. For the TDOA measurement based tracking problem, not only the TDOA measurement has a nonlinear relationship with the source position, but also the measurement model has a bi-modal presence: a Gaussian process if the measurement is generated by a real source, and a uniform distribution if it is from clutter. The particle filtering is thus the method to handle the TDOA based acoustic source tracking problem. Section 3.2 and 3.3 will present the detailed derivation of particle filtering and its variants. Discussion about the measurement model will be given in Chapter 5. Since our final aim is to develop an approach to track multiple acoustic sources, the random finite set (RFS) and multiple source Bayesian filtering are briefly addressed in Section 3.4.

3.1 Bayesian estimator

This section introduces the theory of Bayesian filtering. It also presents the classical solutions, the Kalman filter and extended Kalman filter.

3.1.1 Bayes's theorem

Given two random variables $\mathbf{x}$ and $\mathbf{z}$, Bayes's theorem for a posterior pdf is stated as [79, 80]

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{z})}, \tag{3.1}$$

in which all the terms are defined as follows

- $p(\mathbf{z}|\mathbf{x})$: the likelihood of $\mathbf{x}$;
- $p(\mathbf{x})$: the prior probability density of $\mathbf{x}$;
- $p(\mathbf{z})$: the evidence of $\mathbf{z}$;
- $p(\mathbf{x}|\mathbf{z})$: the posterior probability density for $\mathbf{x}$ given $\mathbf{z}$.

In the general applications of Bayes's theorem, $\mathbf{x}$ denotes an unknown and unobserved variable, and $\mathbf{z}$ is the observed information about $\mathbf{x}$. Once the prior information $p(\mathbf{x})$ of the unknown variable $\mathbf{x}$ is available, its contribution to the system state estimation are modified by the likelihood $p(\mathbf{z}|\mathbf{x})$. Since the denominator $p(\mathbf{z})$ is a constant, Bayes's theorem (3.1) can also be interpreted as

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}$$

where $\propto$ denotes proportionality.

3.1.2 The recursive Bayesian filtering

In real-life applications, proper models and noise processes should be defined to obtain the prior information and the likelihood in Bayes's theorem (3.1). Suppose we have following process and measurement equations

$$\mathbf{x}_k = f_k(\mathbf{x}_{k-1}, \mathbf{v}_k), \tag{3.2a}$$

$$\mathbf{z}_k = g_k(\mathbf{x}_k, \mathbf{w}_k), \tag{3.2b}$$

where k is the time index, $\mathbf{x}_k$ and $\mathbf{z}_k$ are the state and the measurement respectively, $\mathbf{v}_k$ and $\mathbf{w}_k$ are the process noise and the measurement noise separately. The functions $f_k(\cdot)$ and $g_k(\cdot)$ are time-varying process equation and measurement equation respectively. The noise sequences $\mathbf{v}_k$ and $\mathbf{w}_k$ are assumed to be independent from each other and with known pdf's. Taking

the tracking problem for example, the state $\mathbf{x}_k$ can be the position and velocity of a target, and the measurement $\mathbf{z}_k$ is the measurement extracted from the received sensor data. The state equation (3.2a) and measurement equation (3.2b) depict the source dynamics and the relationship between the measurement and state respectively. The aim here is to estimate the posterior of state using the measurements and the system model (3.2). This system model can also be depicted by a hidden Markov process [81], in which the state $\mathbf{x}_k$ is an unobserved (hidden) Markov process, and $\mathbf{z}_k$ is the observed measurements of such model, as shown in Fig. 3.1. According to the Markov property assumed in equation (3.2a), the state $\mathbf{x}_k$ is conditionally independent of all earlier states given the immediately previous state, stated as

$$p(\mathbf{x}_k|\mathbf{x}_0, \ldots, \mathbf{x}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}), \tag{3.3}$$

where $\mathbf{x}_0$ is the initial state. Due to the assumed form in equation (3.2b), the measurement at the time step k is dependent only on the current state $\mathbf{x}_k$ and is conditionally independent of all previous states $\mathbf{x}_0, \ldots, \mathbf{x}_{k-1}$, expressed as

$$p(\mathbf{z}_k|\mathbf{x}_0, \ldots, \mathbf{x}_k) = p(\mathbf{z}_k|\mathbf{x}_k). \tag{3.4}$$

Given the observed measurements up to and including the current time step $\mathbf{z}_{1:k} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$, our goal is to estimate the posterior probability $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ of the unknown state $\mathbf{x}$ recursively.

From the time step $k-1$ to k, we first compute the prior conditional pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$, given

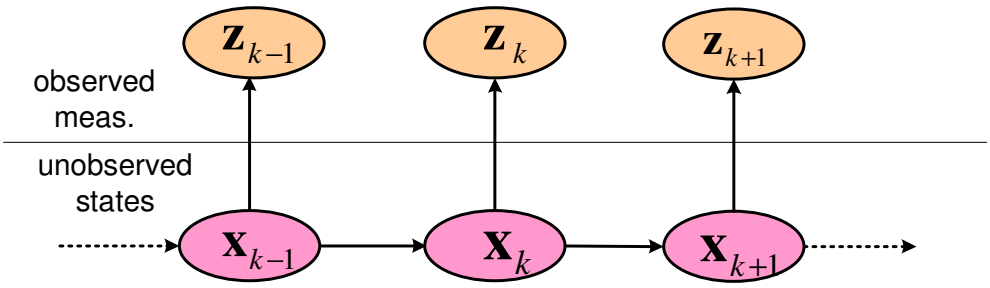


Figure 3.1: *Architecture of the hidden Markov process. The oval shapes at the lower level are hidden states, and these states are represented by the observations at the upper level. Each state follows a Markov property, which means it depends only on the adjacent previous state.*

by

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) &= \int p(\mathbf{x}_k, \mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}; \\ &= \int p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}. \end{aligned} \tag{3.5}$$

This step is the state prediction according to the past measurements and the state transition model. Since $\mathbf{x}_k$ is completely depicted by equation (3.2a), it is determined by the state at previous time step $\mathbf{x}_{k-1}$ and the process noise $\mathbf{v}_k$. The item $\mathbf{z}_{1:k-1}$ in $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k-1})$ can thus be dropped, and equation (3.5) can be written as

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}, \tag{3.6}$$

in which $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, also called transition density, is the pdf given by the process model (3.2a) and the estimate $\mathbf{x}_{k-1}$ at the previous time step, and $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ is simply the posterior calculated at the last time step $k-1$. The posterior distribution at current time step $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ can thus be derived as

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{z}_{1:k}) &= p(\mathbf{x}_k|\mathbf{z}_k, \mathbf{z}_{1:k-1}) \\ &= \frac{p(\mathbf{z}_k|\mathbf{x}_k, \mathbf{z}_{1:k-1}) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\ &= \frac{p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})}. \end{aligned} \tag{3.7}$$

All the pdf's on the right hand side of above equation are available. The likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ can be obtained from the measurement model (3.2b), and $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ is derived by equation (3.6) already. The pdf $p(\mathbf{z}_k|\mathbf{z}_{1:k-1})$ can be derived in a similar way as the derivation of equation (3.5), written as

$$\begin{aligned} p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) &= \int p(\mathbf{z}_k, \mathbf{x}_k|\mathbf{z}_{1:k-1}) d\mathbf{x}_k \\ &= \int p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) d\mathbf{x}_k, \end{aligned} \tag{3.8}$$

Equations (3.6) and (3.7) form the main steps of a recursive Bayesian estimator. Given an initial distribution $p(\mathbf{x}_0)$, i.e., $p(\mathbf{x}_0|\mathbf{z}_0) = p(\mathbf{x}_0)$, and the process and measurement functions $f_k(\cdot)$ and $g_k(\cdot)$ respectively, the recursive Bayesian estimator can be summarised as follows

- Initialisation:

$$p(\mathbf{x}_0|\mathbf{z}_0) = p(\mathbf{x}_0). \tag{3.9}$$

- Predict:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1}. \tag{3.10}$$

- Update:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})}. \tag{3.11}$$

Although the posterior distribution can be recursively derived according to expressions (3.6) and (3.7), the unbiased and closed-form solutions can be achieved only for a few special cases where the state space model is linear and the noise terms are Gaussian.

3.1.3 Kalman filtering

For the case of a linear system with a Gaussian initial distribution $\mathbf{x}_0$ and subject to the Gaussian noise processes $\mathbf{v}$ and $\mathbf{w}$, the posterior distribution can be estimated recursively using the Kalman filter [67, 68, 82]. Suppose the process and measurement equations (3.2a) and (3.2b) are discrete time-varying and have the following linear-Gaussian form

$$\mathbf{x}_k = \mathbf{F}_{k-1}\mathbf{x}_{k-1} + \mathbf{H}_{k-1}\mathbf{u}_k + \mathbf{v}_k, \tag{3.12a}$$

$$\mathbf{z}_k = \mathbf{G}_k\mathbf{x}_k + \mathbf{w}_k. \tag{3.12b}$$

where $\mathbf{u}_k$ is a control vector, and $\mathbf{v}_k$ and $\mathbf{w}_k$ are the process noise and measurement noise with known covariance $\mathbf{Q}_k$ and $\mathbf{R}_k$ respectively. The statistics of the noise processes in (3.12) can be stated as

$$\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{Q}_k), \quad \mathbf{w}_k \sim \mathcal{N}(0, \mathbf{R}_k), \tag{3.13}$$

and the covariance matrix

$$\mathbb{E}\left[\begin{pmatrix}\mathbf{v}_k \\ \mathbf{w}_k\end{pmatrix}\begin{pmatrix}\mathbf{v}_{k'}^T & \mathbf{w}_{k'}^T\end{pmatrix}\right] = \begin{pmatrix}\mathbf{Q}_{k,k'}\delta(k-k') & 0 \\ 0 & \mathbf{R}_{k,k'}\delta(k-k')\end{pmatrix}, \tag{3.14}$$

where $\delta(C)$ and is a dirac function with a value of 1 if $C = 0$ and 0 elsewhere. Given a series of measurements $\mathbf{z}_{1:k}$ and the state estimates at the previous time step $\hat{\mathbf{x}}_{k-1}$, the main steps of implementing a Kalman filtering are summarised as follows

1. Predict:

$$\mathbf{x}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1} + \mathbf{H}_k \mathbf{u}_k, \tag{3.15a}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{P}}_{k-1|k-1} \mathbf{F}_k^T + \mathbf{Q}_k. \tag{3.15b}$$

2. Information gain:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{G}_k^T (\mathbf{G}_k \mathbf{P}_{k|k-1} \mathbf{G}_k^T + \mathbf{R}_k)^{-1}. \tag{3.16}$$

3. Updated:

$$\hat{\mathbf{x}}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{G}_k \mathbf{x}_{k|k-1}), \tag{3.17a}$$

$$\hat{\mathbf{P}}_{k|k} = \mathbf{P}_{k|k-1} - \mathbf{K}_k \mathbf{G}_k \mathbf{P}_{k|k-1}. \tag{3.17b}$$

If the initial distribution $\mathbf{x}_0$ and noise terms $\mathbf{v}$ and $\mathbf{w}$ are Gaussian-distributed, The filtering distribution is thus

$$p(\mathbf{x}_k | \mathbf{z}_{1:k}) = \mathcal{N}(\mathbf{x}_k; \hat{\mathbf{x}}_k, \hat{\mathbf{P}}_k), \tag{3.18}$$

where $\mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \hat{\mathbf{P}})$ denotes a Gaussian distribution for the variable $\mathbf{x}$ with mean $\hat{\mathbf{x}}$ and variance $\hat{\mathbf{P}}$.

The KF algorithm is summarised in Algorithm 1. The Kalman filter can actually be regarded as a filter which whitens the measurements and extracts the maximum possible amount of information from the measurements [68]. In practice, the update of the variance matrix $\hat{\mathbf{P}}_k$ and the information gain $\mathbf{K}_k$ depends only on the model coefficient matrices. The calculation of these items can thus be implemented offline to save the computation load and the memory. Various books and papers derive and present the filter equations in different ways [67, 68, 82]. It is not

Algorithm 1: Kalman filtering algorithm.

Input: measurements $\mathbf{z}_{1:K}$.
Output: state estimates $\hat{\mathbf{x}}_{1:K}$.

Initialisation: $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\hat{\mathbf{P}}_0 \leftarrow \mathbf{P}_0$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
  predict the state $\mathbf{x}_{k|k-1}$ and the corresponding variance matrix $\mathbf{P}_{k|k-1}$ according to (3.15);
  calculate the information gain according to (3.16);
  update the state $\hat{\mathbf{x}}_k$ and variance matrix $\hat{\mathbf{P}}_k$ according to (3.17);
  output the estimates $\hat{\mathbf{x}}_k$.
end

obvious, but these different expressions are, in fact, mathematically equivalent.

3.1.4 Extended Kalman filtering

In practice, utilisation of the KF is limited by the nonlinearity and non-Gaussianity of the physical world. Although some systems are close enough to linear and can be approximated by a linear estimator, the estimation results are no longer satisfied. It thus requires us to explore nonlinear estimators. One of the most popular nonlinear filtering approaches that has been applied in the past few decades is the extended Kalman filter (EKF), in which the core technique is applying a local linearisation to the system equations.

Consider the general model depicted by equations (3.2), but with the functions f and g being nonlinear. To formulate the EKF, first perform a Taylor series expansion of the state equation around the previous state estimation $\hat{\mathbf{x}}_{k-1}$, and ignore the noise term

$$\mathbf{x}_{k|k-1} \approx f_{k-1}(\hat{\mathbf{x}}_{k-1}, u_{k-1}, 0) + \mathbf{F}_{k-1}(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1}) + \mathbf{H}_{k-1}\mathbf{v}_{k-1}, \tag{3.19}$$

where $\mathbf{F}$ and $\mathbf{H}$ are the coefficient matrices of the first order derivation subject to $\mathbf{x}$ and $\mathbf{v}$ respectively, given by

$$\mathbf{F}_{k-1} = \left.\frac{\partial f_{k-1}}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{k-1}} \qquad \mathbf{H}_{k-1} = \left.\frac{\partial f_{k-1}}{\partial \mathbf{v}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{k-1}}. \tag{3.20}$$

Further, rearranging the terms in equation (3.19), yields the linearised form of the state function

$$\mathbf{x}_{k|k-1} \approx \mathbf{F}_{k-1}\mathbf{x}_{k-1} + \bar{\mathbf{v}}_{k-1} + O_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1}), \tag{3.21}$$

where $O_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1})$ is an error term defined as

$$O_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1}) = f_{k-1}(\hat{\mathbf{x}}_{k-1}) - \mathbf{F}_{k-1}\hat{\mathbf{x}}_{k-1} \tag{3.22}$$

, and the new noise term $\bar{\mathbf{v}}_{k-1}$ is

$$\bar{\mathbf{v}}_{k-1} \sim \mathcal{N}(0, \mathbf{H}_{k-1}\mathbf{Q}_{k-1}\mathbf{H}_{k-1}^T). \tag{3.23}$$

The nonlinear state model is thus linearised (has the form $y = ax + b$). This linearisation of the state model (3.19) and (3.22) are also illustrated in Fig. 3.2. The error term $O_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1})$ depicts

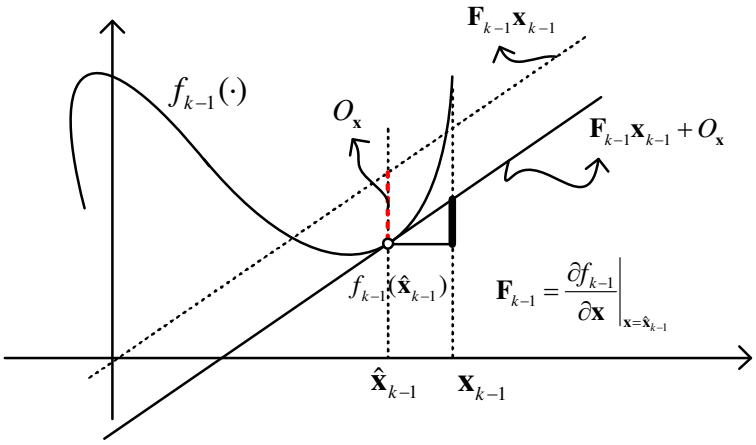


Figure 3.2: *Linearisation of the state model. The partial derivative at $\hat{\mathbf{x}}_{k-1}$ is the slope of the state equation at this point. $O_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1})$ is the error between the linear prediction and the real function value.*

the difference between the linear prediction and the real function value at position $\hat{\mathbf{x}}_{k-1}$.

Similarly, the measurement function can be linearised around $\mathbf{x}_{k|k-1}$. By setting $\mathbf{w}_k = 0$ in the measurement function $g_k(\cdot)$, the first order Taylor expansion can be written as

$$\begin{aligned}\mathbf{z}_k &\approx g(\mathbf{x}_{k|k-1}, 0) + \mathbf{G}_k(\mathbf{x}_k - \mathbf{x}_{k|k-1}) + \mathbf{U}_k\mathbf{w}_k \\ &= \mathbf{G}_k\mathbf{x}_k + \bar{\mathbf{w}}_k + O_{\mathbf{z}}(\mathbf{x}_{k|k-1}),\end{aligned} \tag{3.24}$$

where $\mathbf{G}$ and $\mathbf{U}$ are the coefficients of the first order derivation, given by

$$\mathbf{G}_k = \left.\frac{\partial g_k}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_{k|k-1}}, \qquad \mathbf{U}_k = \left.\frac{\partial g_k}{\partial \mathbf{w}}\right|_{\mathbf{x}=\mathbf{x}_{k|k-1}}, \tag{3.25}$$

and $O_{\mathbf{z}}(\mathbf{x}_{k|k-1})$ and $\bar{\mathbf{w}}_k$ are defined as

$$O_{\mathbf{z}}(\mathbf{x}_{k|k-1}) = g(\mathbf{x}_{k|k-1}, 0) - \mathbf{G}_k\mathbf{x}_{k|k-1} \tag{3.26}$$

$$\bar{\mathbf{w}}_k \sim \mathcal{N}(0, \mathbf{U}_k\mathbf{R}_k\mathbf{U}_k^T), \tag{3.27}$$

Now, given the linear forms of the state equation (3.22) and measurement equation (3.24), the standard Kalman filter can be applied for the state estimation. This results in the following expressions for the EKF:

1. Predict:

$$\mathbf{x}_{k|k-1} = f_{k-1}(\hat{\mathbf{x}}_{k-1}, u_{k-1}, 0) \tag{3.28}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_{k-1}\hat{\mathbf{P}}_{k-1}\mathbf{F}_{k-1}^T + \mathbf{H}_{k-1}\mathbf{Q}_{k-1}\mathbf{H}_{k-1}^T, \tag{3.29}$$

2. Information gain:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{G}_k^T(\mathbf{G}_k\mathbf{P}_{k|k-1}\mathbf{G}_k^T + \mathbf{U}_k\mathbf{R}_k\mathbf{U}_k^T)^{-1}, \tag{3.30}$$

3. Updated:

$$\begin{aligned}\hat{\mathbf{x}}_k &= \mathbf{x}_{k|k-1} + \mathbf{K}_k(\mathbf{z}_k - \mathbf{G}_k\mathbf{x}_{k|k-1} - O_{\mathbf{z}}(\mathbf{x}_{k|k-1})) \\ &= \mathbf{x}_{k|k-1} + \mathbf{K}_k(\mathbf{z}_k - g_k(\mathbf{x}_{k|k-1}, 0)) \qquad (3.31)\end{aligned}$$

$$\hat{\mathbf{P}}_k = \mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{G}_k\mathbf{P}_{k|k-1}. \tag{3.32}$$

The EKF algorithm is summarised in Algorithm 2. The main steps of EKF and KF are almost the same except that the EKF needs to calculate the first order expansion coefficient matrices at the beginning of each recursion. Note that here, only the first order Taylor expansion has been employed. Higher order expansions can also be employed in a similar way to formulate the EKF, such as second-order Kalman filtering, iterated Kalman filtering, and grid-based Kalman filtering [68, 83, 84]. These approaches are able to enhance the estimation performance due to a more accurate approximation of system model, but at the cost of higher complexity and

Algorithm 2: Extended Kalman filtering algorithm.

Input: Measurements $\mathbf{z}_{1:K}$.
Output: State estimates $\hat{\mathbf{x}}_{1:K}$.

Initialisation: $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\hat{\mathbf{P}}_0 \leftarrow \mathbf{P}_0$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
  compute the partial derivative matrices according to (3.20) and (3.25);
  predict the state $\mathbf{x}_{k|k-1}$ and the corresponding variance matrix $\mathbf{P}_{k|k-1}$ according to (3.28) and (3.29);
  calculate the information gain according to (3.30);
  update the state $\hat{\mathbf{x}}_k$ and variance matrix $\hat{\mathbf{P}}_k$ according to (3.31) and (3.32);
  output the estimates $\hat{\mathbf{x}}_k$.
end

computation burden. Due to the nonlinearity of the TDOA measurement function, the EKF will be employed to linearise it and estimate the position state in Chapter 5 and Chapter 6.

3.2 Particle filtering

When the nonlinearities of the systems are severe, the performance of the EKF will degrade and unreliable state estimates will be presented. This is because the linearalisation accuracy of the state space model is limited by the first-order Taylor expansion. Although more accurate filters such as higher-order Kalman filtering and unscented Kalman filtering (UKF) [68] are proposed to solve this problem, the divergence of state estimation still inevitably occurs since all of these approximations are not sufficient in such a case. In this section, the particle filtering, a nonlinear estimator which is completely appropriate for nonlinear system model and non-Gaussian noise process will be fully presented. The original work of this section and the application of the PF in the tracking problem can be found in [85] and [10] respectively.

3.2.1 Monte Carlo approximation

Monte Carlo methods use statistical sampling and estimation technique to evaluate the solutions to the mathematical problems. Consider the estimation of a Lebesque-Stieltjes integral

$$I(f) = \mathbb{E}\left[f(\mathbf{x})\right] = \int f(\mathbf{x})P(d\mathbf{x}), \tag{3.33}$$

where $f(\mathbf{x})$ is an integrable function in a measurable space, and P stands for the distribution. Monte Carlo approximation is using a number of independent random samples $\mathbf{x}^{(i)}; i = 1, \ldots, N$ to represent the distribution P, given by

$$P(d\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}^{(i)}}(d\mathbf{x}), \tag{3.34}$$

where $\delta_{\mathbf{x}^{(i)}}(d\mathbf{x})$ is the delta-Dirac mass only with value of one at $\mathbf{x}^{(i)}$ and 0 elsewhere. One can thus obtain the expectation (3.33) by a numerical integration, stated as

$$I_N(f) \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)}), \tag{3.35}$$

From the strong law of large numbers (SLLN), the average value of samples converges almost surely to the expected value, as infinite trials are performed. We thus have

$$\lim_{N\to+\infty} I_N(f) \overset{a.s.}{\to} I(f), \tag{3.36}$$

where $\overset{a.s.}{\to}$ denotes that converges almost surely. This means the Monte Carlo approximation converges to the real distribution as the number of samples increase to infinity. If the posterior variance of $f(\mathbf{x})$ satisfies

$$\sigma_f^2 = \text{var}[I_N(f)] < +\infty. \tag{3.37}$$

A central limit theorem (CLT) then holds

$$\lim_{N\to+\infty} \sqrt{N}\,[I_N(f) - I(f)] \Longrightarrow \mathcal{N}(0, \sigma_f^2), \tag{3.38}$$

where $\Longrightarrow$ denotes convergence in distribution. The crucial property of Monte Carlo approximation is then clear: the estimation accuracy and the speed of convergence are independent on the dimensionality of the state space but depend only on the number of particles N. Unfortunately, it is usually very difficult to sample efficiently from the distribution $P(\mathbf{x})$. One fundamental problem is thus arising in Monte Carlo sampling approach: how to draw the random samples $\mathbf{x}^{(i)}$ to approximate a probability distribution $P(d\mathbf{x})$?

3.2.2 Importance sampling

Due to the multivariate and non-standard distribution of the state, it is often impossible to sample directly from the posterior distribution. One solution is to use the importance sampling (IS) scheme: draw the samples from a probability distribution $q(\mathbf{x})$ by which the samples can be easily sampled, rather than from the true distribution $p(\mathbf{x})$. Since the distribution of the important region is more interested, the objective of the IS is to sample the distribution around this area to achieve a computational efficiency.

Suppose the support of $q(\mathbf{x})$ is able to cover that of $p(\mathbf{x})$. The integration (3.33) can be rewritten

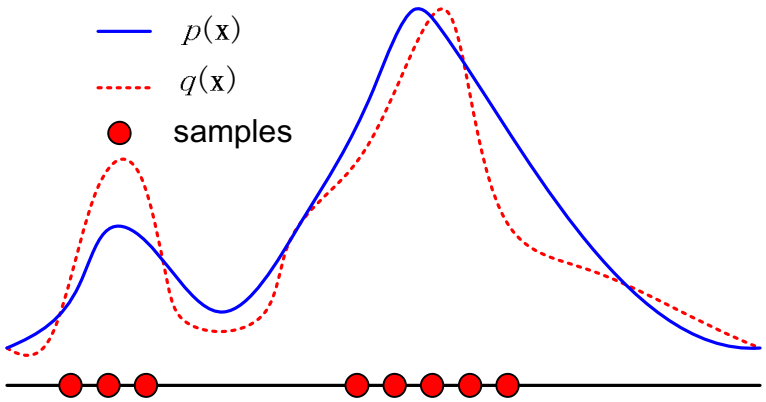


Figure 3.3: *Illustration of importance sampling. $p(\mathbf{x})$ is the actual distribution; $q(\mathbf{x})$ is the proposed distribution.*

as

$$\begin{aligned} I(f) &= \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{3.39}$$

Suppose a number (say N) of i.i.d samples are drawn from the distribution function $q(\cdot)$, as shown in Fig. 3.3. One can approximate the expectation I as follows

$$\hat{I}_N(f) = \frac{1}{N}\sum_{i=1}^{N} w^{(i)} f(\mathbf{x}^{(i)}), \tag{3.40}$$

where

$$w^{(i)} = \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}, \tag{3.41}$$

is called the importance weight. We can further normalise these importance weights

$$\tilde{w}^{(i)} = \frac{w^{(i)}}{\sum_{i=1}^{N} w^{(i)}} = \frac{1}{N} w^{(i)}. \tag{3.42}$$

The expression (3.40) can then be written as

$$\hat{I}_N(f) = \sum_{i=1}^{N} \tilde{w}^{(i)} f(\mathbf{x}^{(i)}). \tag{3.43}$$

For the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$, suppose we have the importance function $q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$

in hand, we can approximate it by

$$dP(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = \sum_{i=1}^{N} \tilde{w}(\mathbf{x}_{0:k}^{(i)}) \delta_{\mathbf{x}_{0:k}^{(i)}}(d\mathbf{x}_{0:k}), \tag{3.44}$$

with

$$w(\mathbf{x}_{0:k}^{(i)}) = \frac{p(\mathbf{x}_{0:k}^{(i)}|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}^{(i)}|\mathbf{z}_{1:k})} \quad \text{and} \quad \tilde{w}(\mathbf{x}_{0:k}^{(i)}) = \frac{w(\mathbf{x}_{0:k}^{(i)})}{\sum_{i=1}^{N} w(\mathbf{x}_{0:k}^{(i)})}, \tag{3.45}$$

and thus the expected estimate is

$$\hat{I}_N\left(f(\mathbf{x}_{0:k})\right) = \sum_{i=1}^{N} \tilde{w}(\mathbf{x}_{0:k}^{(i)}) f(\mathbf{x}_{0:k}^{(i)}). \tag{3.46}$$

The estimation of $I_N\left(f(\mathbf{x}_{0:k})\right)$ in (3.46) is unbiased and converges $a.s.$ to the ground truth $I\left(f(\mathbf{x}_{0:k})\right)$ as $N \to +\infty$ [10, 11]. However, it is cumbersome to store all the samples for all time steps, as the computational complexity will increase as more measurements are available.

3.2.3 Sequential importance sampling

The key step in Monte Carlo simulation is designing an efficient proposal distribution. A natural way to achieve the computation efficiency of IS is to construct proposal distribution sequentially. According to the chain rule of probability, we have

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = p(\mathbf{x}_0) \prod_{n=1}^{k} p(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \mathbf{z}_{1:n}), \tag{3.47}$$

$$q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) = q(\mathbf{x}_0) \prod_{n=1}^{k} q(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \mathbf{z}_{1:n}). \tag{3.48}$$

Following the derivation of (3.7), the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ can be written as [10]

$$\begin{aligned} p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) &= p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1}, \mathbf{z}_k) \\ &= \frac{p(\mathbf{z}_k|\mathbf{x}_{0:k}, \mathbf{z}_{1:k-1})p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\ &= \frac{p(\mathbf{z}_k|\mathbf{x}_{0:k-1}, \mathbf{x}_k, \mathbf{z}_{1:k-1})p(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k-1})p(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\ &= p(\mathbf{x}_{0:k-1}|\mathbf{z}_{1:k-1})\frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})}. \end{aligned} \tag{3.49}$$

The derivation at last step is according to the state space model (3.2) and equations (3.3) and (3.4). Hence the importance weight $w_k^{(i)}$ can be constructed as

$$\begin{aligned} w_k^{(i)} &= \frac{p(\mathbf{x}_{0:k}^{(i)}|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}^{(i)}|\mathbf{z}_{1:k})} \\ &= \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})p(\mathbf{x}_{0:k-1}^{(i)}|\mathbf{z}_{1:k-1})}{q(\mathbf{x}_{0:k-1}^{(i)}|\mathbf{z}_{1:k-1})q(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathbf{z}_{1:k-1})p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\ &\propto w_{k-1}^{(i)}\frac{p(\mathbf{z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_{1:k-1})}. \end{aligned} \tag{3.50}$$

Note that the last step is proportional to the right hand side of the expression since $p(\mathbf{z}_k|\mathbf{z}_{1:k-1})$ is a constant. One advantage of SIS is that it only stores the replicates sampled from the previous importance sampler other than all the particles at each step, which consequently improves the efficiency. However, the variance of the importance weights increases over time, which leads to the so-called particle 'degeneracy problem' [11, 85]. That is, after a few iterations of the algorithm, only several or even just one particle with a nonzero weight. Most of the computation are wasted to update those unimportant particles.

In order to alleviate the particle degeneracy, resampling schemes are suggested to be employed after the weight normalisation. The general idea is that replicate the particles with high normalised weights, and discard those with low normalised weights. Typical resampling schemes are multinomial resampling [86], residual resampling [87], and systematic resampling [88]. Since these resampling techniques are now standard modules in implementing a particle filter, detailed algorithms can be easily found in the corresponding publications, and will not be presented here.

The effective sample size N_{eff} is introduced to measure the degeneracy, stated as [89]

$$N_{\text{eff}} = \frac{N}{1 + \text{var}_{q(\cdot|\mathbf{z}_{1:k})}\left[w(\mathbf{x}_{0:k})\right]}. \tag{3.51}$$

In practice, $\text{var}_{q(\cdot|\mathbf{z}_{1:k})}\left[w(\mathbf{x}_{0:k})\right]$ is not available [89]. An alternative estimation is given by

$$\hat{N}_{\text{eff}} = 1 \Big/ \sum_{i=1}^{N} (\tilde{w}_k^{(i)})^2, \tag{3.52}$$

when $\hat{N}_{\text{eff}}$ is below a predefined threshold N_T (say $N/2$ or $N/4$), the resampling scheme is performed. The core steps of implementing a SIS filter with a resampling scheme are summarised in Algorithm 3. The advantage of the additional resampling step is that it is able to alleviate the particle degeneracy problem.

Algorithm 3: Sequential importance sampling filter with resampling.

Input: Particles and the corresponding weights $\{\mathbf{x}_{k-1}^{(i)}, \tilde{w}_{k-1}^{(i)}\}, \qquad 1 \leq i \leq N.$

Output: Resampled particles and the corresponding weights $\{\mathbf{x}_{k}^{(i)}, \tilde{w}_{k}^{(i)}\}, \qquad 1 \leq i \leq N.$

for $i = 1, \ldots, N$ **do**

  sample $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_{1:k})$, and set $\mathbf{x}_{0:k}^{(i)} = \left(\mathbf{x}_k^{(i)}, \mathbf{x}_{0:k-1}^{(i)}\right)$.

end

for $i = 1, \ldots, N$ **do**

  evaluate the importance weights according to equation (3.50).

end

for $i = 1, \ldots, N$ **do**

  normalise the importance weights: $\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{N} w_k^{(i)}}$.

end

Calculate $\hat{N}_{\text{eff}}$ according to equation (3.52).

if $\hat{N}_{\text{eff}} > N_T$ **then**

  return.

else

  resampling: multiply/discard the samples $\mathbf{x}_{0:k}^{(i)}$ with high/low importance weight $\tilde{w}_k^{(i)}$, respectively, to obtain N new particles.

end

3.2.4 Bootstrap/SIR filter

The Bootstrap filter [90, 91] and the sampling importance resampling (SIR) filter [87] are very close in spirit. Here we regard them as a same category for discussion, and the generic algo-

rithm is summarised in Algorithm 4. The sampling step is exactly the same as that depicted in SIS method, with the only difference on the resampling step: in SIR filter the resampling scheme is usually performed at each recursion; whereas in SIS filter resampling scheme is only implemented if necessary.

Algorithm 4: Sequential importance resampling filter with resampling.

Input: Particles and the corresponding weights $\{\mathbf{x}_{k-1}^{(i)}, \tilde{w}_{k-1}^{(i)}\}, \qquad 1 \leq i \leq N.$

Output: Resampled particles and the corresponding weights $\{\mathbf{x}_{k}^{(i)}, \tilde{w}_{k}^{(i)}\}, \qquad 1 \leq i \leq N.$

for $i = 1, \ldots, N$ **do**

  sample $\mathbf{x}_k^{(i)} \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_{1:k})$, and set $\mathbf{x}_{0:k}^{(i)} = \left(\mathbf{x}_k^{(i)}, \mathbf{x}_{0:k-1}^{(i)}\right)$.

end

for $i = 1, \ldots, N$ **do**

  evaluate the importance weights according to equation (3.50).

end

for $i = 1, \ldots, N$ **do**

  normalise the importance weights: $\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{N} w_k^{(i)}}$.

end

Resampling: multiply/discard the samples $\mathbf{x}_{0:k}^{(i)}$ with high/low importance weight $\tilde{w}_k^{(i)}$, respectively, to obtain N new particles.

For the implementation of SIS and SIR filtering, the choice of importance function plays a crucial role in the state estimation performance. Different choices of importance function have been extensively discussed in [11, 85]. It is suggested that the miminum variance can be achieved by employing the optimal importance function [85].

Figure 3.4 shows a complete iteration of a particle filter. The target distribution are firstly sampled according to the importance function. The likelihood combining with the transition model are then employed to evaluate the importance (weight) of each particles. The particles with high/low weights are replicated/discarded. Consequently, the new particles represent the current posterior distribution.

3.3 Rao-Blackwellised paticle filtering

Rao-Blackwellisation, motivated by the Rao-Blackwell theorem, is a kind of marginalization technique. Because of its intrinsic property of variance reduction, it has been employed in PFs to improve estimation performance [92–94]. In this section, the RBPF formulation for state

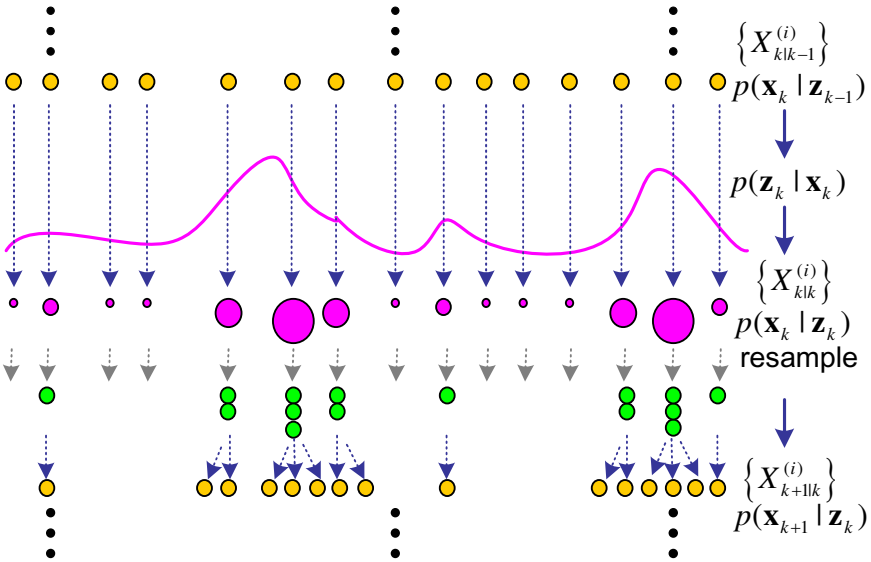


Figure 3.4: *Illustration of the particle filtering. The samples are drawn according to the importance function, and the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ is used to correct the distribution of the particles.*

decomposition is presented. It is the theoretical basis in constructing the multiple acoustic source tracking approach developed in Chapter 6.

3.3.1 Rao-Blackwellised model decomposition

For the dynamic state space model depicted by equations (3.2), suppose that the state $\mathbf{x}_k$ can be split into two parts $(\mathbf{x}^1_k, \mathbf{x}^2_k)$. The basic principle of Rao-Blackwellisation is to decompose the model structure: marginalise one part by exploiting an analytical solution, and estimate the other part by employing an SMC approach.

To prove this advantage, following observations proposed in [85, 95] are represented here. Given the new states $\mathbf{x}_k = (\mathbf{x}^1_k, \mathbf{x}^2_k)$, assume the marginal density $p(\mathbf{x}^2_k|\mathbf{x}^1_k)$ is analytically tractable. The expectation of $f(\mathbf{x}_{0:k})$ with respect to the posterior $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ can be rewritten

as

$$
\begin{aligned}
I(f(\mathbf{x}_{0:k})) = \mathbb{E}(f) &= \int f(\mathbf{x}_{0:k}) p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) d\mathbf{x}_{0:k} \\
&= \frac{\int f(\mathbf{x}_{0:k}) p(\mathbf{z}_{1:k}|\mathbf{x}_{0:k}) p(\mathbf{x}_{0:k}) d\mathbf{x}_{0:k}}{\int p(\mathbf{z}_{1:k}|\mathbf{x}_{0:k}) p(\mathbf{x}_{0:k}) d\mathbf{x}_{0:k}} \\
&= \frac{\int \left[\int f(\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}) p(\mathbf{z}_{1:k}|\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}) p(\mathbf{x}^2_{0:k}|\mathbf{x}^1_{0:k}) d\mathbf{x}^2_{0:k}\right] p(\mathbf{x}^1_{0:k}) d\mathbf{x}^1_{0:k}}{\int \left[\int p(\mathbf{z}_{1:k}|\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}) p(\mathbf{x}^2_{0:k}|\mathbf{x}^1_{0:k}) d\mathbf{x}^2_{0:k}\right] p(\mathbf{x}^1_{0:k}) d\mathbf{x}^1_{0:k}}.
\end{aligned}
\tag{3.53}
$$

Since the second part $\mathbf{x}^2_{0:k}$ is assumed to be able to analytically integrate out, above expression can be written as

$$
I(f(\mathbf{x}_{0:k})) = \frac{\int \phi(\mathbf{x}^1_{0:k}) p(\mathbf{x}^1_{0:k}) d\mathbf{x}^1_{0:k}}{\int p(\mathbf{z}_{1:k}|\mathbf{x}^1_{0:k}) p(\mathbf{x}^1_{0:k}) d\mathbf{x}^1_{0:k}}, \tag{3.54}
$$

where

$$
\phi(\mathbf{x}^1_{0:k}) = \int f(\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}) p(\mathbf{z}_{1:k}|\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}) p(\mathbf{x}^2_{0:k}|\mathbf{x}^1_{0:k}) d\mathbf{x}^2_{0:k}. \tag{3.55}
$$

The Rao-Blackwellised Monte Carlo approximation $\hat{I}_{\text{RB}}$ of equation 3.54 can be given as

$$
\hat{I}_{\text{RB}}(f(\mathbf{x}_{0:k})) = \frac{\sum_{i=1}^N \phi(\mathbf{x}^{1,(i)}_{0:k}) w(\mathbf{x}^{1,(i)}_{0:k})}{\sum_{i=1}^N w(\mathbf{x}^{1,(i)}_{0:k})}. \tag{3.56}
$$

Simply applying the importance function $q(\mathbf{x}^1_{0:k}, \mathbf{x}^2_{0:k}|\mathbf{z}_{1:k})$, the Monte Carlo approximation $\hat{I}$ can be written as

$$
\hat{I}(f(\mathbf{x}_{0:k})) = \frac{\sum_{i=1}^N f(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k}) w(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k})}{\sum_{i=1}^N w(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k})}, \tag{3.57}
$$

where the importance weight $w(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k})$ is

$$
w(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k}) = \frac{p(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k}|\mathbf{z}_{1:k})}{q(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^{2,(i)}_{0:k}|\mathbf{z}_{1:k})}. \tag{3.58}
$$

The importance weight $w(\mathbf{x}^{1,(i)}_{0:k})$ for Rao-Blackwellised Monte Carlo approximation (3.56) is obtained by integrating out the state part $\mathbf{x}^2_{0:k}$, given as

$$
w(\mathbf{x}^{1,(i)}_{0:k}) = \frac{p(\mathbf{x}^{1,(i)}_{0:k}|\mathbf{z}_{1:k})}{\int q(\mathbf{x}^{1,(i)}_{0:k}, \mathbf{x}^2_{0:k}|\mathbf{z}_{1:k}) d\mathbf{x}^2_{0:k}}. \tag{3.59}
$$

After the Rao-Blackwellisation step, part of the state $\mathbf{x}_{0:k}^2$ is integrated out analytically, and the other part of the state $\mathbf{x}_{0:k}^1$ is needed to be handled by Monte Carlo approximation.

According to the formula for variance decomposition: if $\mathbf{x}_{0:k}^1$ and $\mathbf{x}_{0:k}^2$ are two random variables, and the variance of f_k exists, then

$$\text{Var}\left[f_k\right] = \text{Var}\left[\mathbb{E}\left[f(\mathbf{x}_{0:k}^1, \mathbf{x}_{0:k}^2)|\mathbf{x}_{0:k}^1\right]\right] + \mathbb{E}\left[\text{Var}\left[f(\mathbf{x}_{0:k}^1, \mathbf{x}_{0:k}^2)|\mathbf{x}_{0:k}^1\right]\right], \tag{3.60}$$

where $\mathbb{E}\left[f(\mathbf{x}_{0:k}^1, \mathbf{x}_{0:k}^2)|\mathbf{x}_{0:k}^1\right]$ is the conditional expectation of f given $\mathbf{x}_{0:k}^1$, and $\text{Var}\left[f(\mathbf{x}_{0:k}^1, \mathbf{x}_{0:k}^2)|\mathbf{x}_{0:k}^1\right]$ is the conditional variance of f given $\mathbf{x}_{0:k}^1$. Since both the terms at the right hand side of equation (3.60) are greater than 0, we have

$$\text{Var}\left[f_k\right] \geq \text{Var}\left[\mathbb{E}\left[f(\mathbf{x}_{0:k}^1, \mathbf{x}_{0:k}^2)|\mathbf{x}_{0:k}^1\right]\right]. \tag{3.61}$$

The equations (3.60) and (3.61) illustrate that the variance of whole state estimation is not less than the variance of the state estimation conditional on part of states $\mathbf{x}_{0:k}^1$.

Similarly, the variance decomposition for the importance weights can be stated as

$$\begin{aligned}\text{Var}\left[\frac{p(\mathbf{x}_k^1, \mathbf{x}_k^2)}{q(\mathbf{x}_k^1, \mathbf{x}_k^2)}\right] &= \text{Var}\left[\frac{\int p(\mathbf{x}_k^1, \mathbf{x}_k^2) d\mathbf{x}_k^2}{\int q(\mathbf{x}_k^1, \mathbf{x}_k^2) d\mathbf{x}_k^2}\right] + \mathbb{E}\left[\text{Var}\left[\frac{p(\mathbf{x}_k^1, \mathbf{x}_k^2)}{q(\mathbf{x}_k^1, \mathbf{x}_k^2)}\middle| \mathbf{x}_k^1\right]\right] \\ &\geq \text{Var}\left[\mathbb{E}\left[\frac{p(\mathbf{x}_k^1, \mathbf{x}_k^2)}{q(\mathbf{x}_k^1, \mathbf{x}_k^2)}\middle| \mathbf{x}_k^1\right]\right],\end{aligned} \tag{3.62}$$

with

$$\mathbb{E}\left[\frac{p(\mathbf{x}_k^1, \mathbf{x}_k^2)}{q(\mathbf{x}_k^1, \mathbf{x}_k^2)}\middle| \mathbf{x}_k^1\right] = \frac{\int p(\mathbf{x}_k^1, \mathbf{x}_k^2) d\mathbf{x}_k^2}{\int q(\mathbf{x}_k^1, \mathbf{x}_k^2) d\mathbf{x}_k^2}. \tag{3.63}$$

Hence, the variance of the importance weights can be reduced via Rao-Blackwellisation step. This proof can be originally found in [85, 95]. The advantage of Rao-Blackwellised Monte Carlo simulation is shown by equations (3.56), (3.57) and (3.62). By taking the Rao-Blackwellisation model decomposition, a lower variance of the marginalised state estimate, as well as that of the importance weight can be achieved.

3.3.2 Rao-Blackwellised particle filtering

Consider the following state-space model

$$\boldsymbol{\theta}_k = h(\boldsymbol{\theta}_{k-1}, \mathbf{e}_k) \tag{3.64a}$$

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}, \boldsymbol{\theta}_k, \mathbf{v}_k) \tag{3.64b}$$

$$\mathbf{z}_k = g(\mathbf{x}_k, \boldsymbol{\theta}_k, \mathbf{w}_k), \tag{3.64c}$$

where $\boldsymbol{\theta}_k$ is a latent variable. The new state Θ_k to be estimated is an extension of the original state $\mathbf{x}_k$: $\Theta_k = (\mathbf{x}_k, \boldsymbol{\theta}_k)$. The Bayesian network of such model is illustrated in Fig. 3.5. By the model decomposing, we have

$$\begin{aligned} p(\Theta_k|\Theta_{k-1}) &= p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta}_{k-1}) \\ &= p(\mathbf{x}_k|\mathbf{x}_{k-1}, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}), \end{aligned} \tag{3.65}$$

and the corresponding posterior distribution $p(\mathbf{x}_{0:k}, \boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k})$ is given by

$$p(\mathbf{x}_{0:k}, \boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k}) = p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}, \boldsymbol{\theta}_{0:k}) p(\boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k}). \tag{3.66}$$

Here, assume that conditional on $\boldsymbol{\theta}_{0:k}$, the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}, \boldsymbol{\theta}_{0:k})$ is analytically tractable. That means we can integrate out $\mathbf{x}_{0:k}$ from the posterior, and only need to implement a particle filtering method on estimating $p(\boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k})$. Following the derivation of Bayesian recursion (3.49), the posterior pdf $p(\boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k})$ in (3.66) can be written as

$$p(\boldsymbol{\theta}_{0:k}|\mathbf{z}_{1:k}) = p(\boldsymbol{\theta}_{0:k-1}|\mathbf{z}_{1:k-1}) \frac{p(\mathbf{z}_k|\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})}. \tag{3.67}$$

In this manner, a part of states are computed in closed form, which can be regarded as using an infinite number of samples instead of a limited number of samples by a Monte Carlo approximation. According to the Rao-Blackwell theorem, better results are expected than applying the sampling method to the whole state estimation.

Given the particles $\boldsymbol{\theta}_{k-1}^{(i)}$, and the importance weight $w_{k-1}^{(i)}$, $i = 1, \ldots, N$ at time step $k-1$, the importance weight at time step k can be recursively updated as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{z}_k|\boldsymbol{\theta}_k^{(i)}, \mathbf{z}_{1:k-1}) p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)})}{q(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, \mathbf{z}_{1:k})}, \tag{3.68}$$

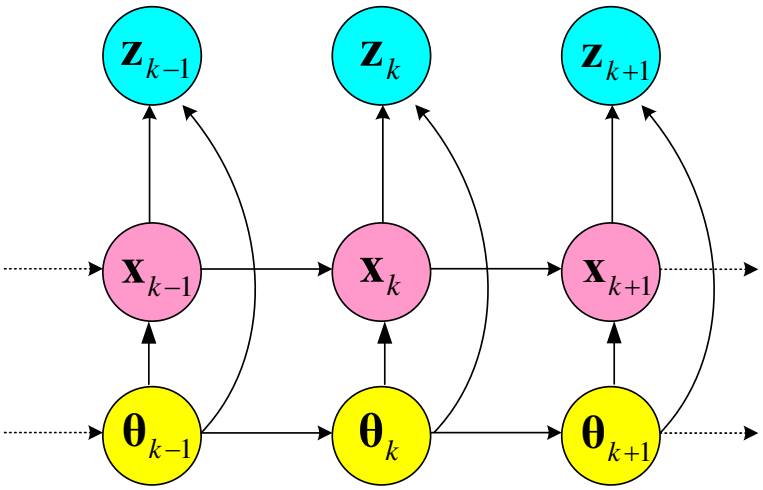


Figure 3.5: *Bayesian network of the state-space model with a latent variable.*

in which each component can be calculated in the same way as we derived for a regular particle filtering. The filtered density is obtained by

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k|\mathbf{z}_k) = \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k) p(\mathbf{x}_k|\boldsymbol{\theta}_k^{(i)}, \mathbf{z}_{1:k}), \tag{3.69}$$

where $\tilde{w}_k^{(i)}$ is the normalised weight, and $\delta(\cdot)$ is a delta-Dirac mass as defined in equation (3.34), on page 57.

For example, the state space model (3.64) has the following expression

$$\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}), \tag{3.70a}$$

$$\mathbf{x}_k = \mathbf{F}_{k-1}(\boldsymbol{\theta}_k)\mathbf{x}_{k-1} + \mathbf{H}_{k-1}(\boldsymbol{\theta}_k)\mathbf{u}_k + \mathbf{v}_k(\boldsymbol{\theta}_k), \tag{3.70b}$$

$$\mathbf{z}_k = \mathbf{G}_k(\boldsymbol{\theta}_k)\mathbf{x}_k + \mathbf{w}_k(\boldsymbol{\theta}_k). \tag{3.70c}$$

where $\mathbf{x}_k$ and $\boldsymbol{\theta}_k$ are the state of the system, and conditional on $\boldsymbol{\theta}_k$ the state space model for $\mathbf{x}_k$ and $\mathbf{z}_k$ is linear and Gaussian. The coefficient matrices $\mathbf{F}$, $\mathbf{H}$, $\mathbf{G}$ and noise term $\mathbf{v}$ and $\mathbf{w}$ are dependent on $\boldsymbol{\theta}_k$. This means that the state $\mathbf{x}_k$ can be estimated from the measurements $\mathbf{z}_k$ using a Kalman filter, and only the state $\boldsymbol{\theta}_k$ is needed to be estimated by particle filtering. Suppose for ith, $i = 1, \ldots, N$ particle, the posterior distribution of $\mathbf{x}_k^{(i)}$ based on KF estimation is

$$p(\mathbf{x}_k^{(i)}|\boldsymbol{\theta}_k^{(i)}, \mathbf{z}_{1:k}) = \mathcal{N}(\mathbf{x}_k^{(i)}|\hat{\mathbf{x}}_k^{(i)}, \hat{\mathbf{P}}_k^{(i)}). \tag{3.71}$$

where $\hat{\mathbf{x}}_k^{(i)}$ and $\hat{\mathbf{P}}_k^{(i)}$ are the estimated mean vector and covariance matrix from Kalman filtering.

The joint posterior distribution of whole state $p(\mathbf{x}_k, \boldsymbol{\theta}_k|\mathbf{z}_{1:k})$ will be

$$p(\mathbf{x}_k, \boldsymbol{\theta}_k|\mathbf{z}_{1:k}) = \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k) \mathcal{N}(\hat{\mathbf{x}}_k^{(i)}|\hat{\mathbf{x}}_k^{(i)}, \hat{\mathbf{P}}_k^{(i)}). \tag{3.72}$$

The Kalman filtering here is exactly the same as it described in Section 3.1.3, on page 52. If either or both of the state-space equations f and g are nonlinear, $p(\mathbf{x}_k^{(i)}|\boldsymbol{\theta}_k^{(i)}, \mathbf{z}_{1:k})$ can be obtained by using an EKF presented in Section 3.1.4, on page 54, or a UKF [68]. The complete Rao-Blackwellised particle filtering with a Kalman filtering marginalisation is shown in Algorithm 5.

Algorithm 5: Rao-Blackwellised particle filtering.

Input: Particles $\{\mathbf{x}_{k-1}^{(i)}, \boldsymbol{\theta}_{k-1}^{(i)}\}$ and the corresponding weights $\tilde{w}_{k-1}^{(i)}$, $1 \leq i \leq N$
Output: Estimated states.

for $i = 1, \ldots, N$ **do**
  sample $\boldsymbol{\theta}_k^{(i)} \sim q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, \mathbf{z}_{1:k})$, and set $\boldsymbol{\theta}_{0:k}^{(i)} = \left(\boldsymbol{\theta}_k^{(i)}, \boldsymbol{\theta}_{0:k-1}^{(i)}\right)$.
end
for $i = 1, \ldots, N$ **do**
  evaluate the importance weights according to (3.68);
  perform Kalman filtering according to equations from (3.15) to (3.17), on page 53.
end
for $i = 1, \ldots, N$ **do**
  normalise the importance weights: $\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{N} w_k^{(i)}}$.
end
Resampling if necessary: multiply/discard the samples $\boldsymbol{\theta}_{0:k}^{(i)}$ with high/low importance weight $\tilde{w}_k^{(i)}$ respectively, to obtain N new particles.
Output the estimated states.

3.4 Multiple source Bayesian filtering

The Bayesian filtering and its implementations introduced in the previous sections 3.1 and 3.2 focus on a single state (the state may be multi-dimensional) estimation problem, e.g., there is always a state and all the measurements are the observations of this state. For the single acoustic source tracking problem, the particle filtering allows the presence of clutter due to the using of a reverberation measurement model. The details of such model will be introduced in chapter 5. In the multiple source filtering scenario, both the states and the measurements are with set value, which employs a different structure from those in the single source case. In this section,

a random finite set (RFS) formulation of multiple source filtering problem will be introduced. Multiple source Bayesian filtering and its SMC implementation are also briefly presented.

3.4.1 Random finite set formulation

In essence, a RFS $\mathcal{X}$ is a finite set-valued variable where not only each element in the variable is random but also the number of the elements (i.e., the cardinality of $\mathcal{X}$) is random [96–98]. Its cardinality randomness is usually described by a discrete probability distribution and the joint distribution of its elements is depicted by an appropriate density. The RFS theory provides an elegant framework to the multiple source tracking problem since in the scenario of multiple sources, the number of sources varies over time due to the appearance or disappearance of the sources, and also the number of measurements is time-varying and usually not the same as the number of sources due to the presence of clutter.

In the multiple source tracking problem, the number of sources, M_k, and the states of each source at time step k can be summarised in a RFS, given as

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{M_k,k}\}, \tag{3.73}$$

where $\mathbf{x}_{m,k} \in \mathbb{R}^{n_\mathbf{x}}$, $m = 1, \ldots, M_k$ represents the mth state vector with $n_\mathbf{x}$ denoting the dimension of the state vector (the dimension of the state vector is fixed), and $M_k = |\mathcal{X}_k|$, where $|\cdot|$ stands for the cardinality, denotes the number of the sources. Assume that $\mathbf{z}_k^\ell$ are the measurements obtained from the ℓth, $\ell = 1, \ldots, L$, sensor, expressed as

$$\mathbf{z}_k^\ell = \{z_{1,k}^\ell, \ldots, z_{n_k^\ell,k}^\ell\}, \tag{3.74}$$

where the number of the measurements is n_k^ℓ, and $z_{n,k}^\ell \in \mathbb{R}^{n_\mathrm{z}}$ is usually a singleton corresponding to a measurement (i.e., $n_\mathrm{z} = 1$). The complete measurement set can be written as

$$\mathcal{Z}_k = \bigcup_{\ell=1}^{L} \mathbf{z}_k^\ell. \tag{3.75}$$

The cardinality of the measurement set is thus $|\mathcal{Z}_k| = \sum_\ell n_k^\ell$.

Equation (3.73) and (3.74) define a finite set based system model, in which the multiple source state space $\mathcal{F}(\mathcal{X})$ is spanned by the state sets $\mathcal{X}$, and the measurement space $\mathcal{F}(\mathcal{Z})$ is con-

structed by the finite measurement set $\mathcal{Z}$. The aim of the multiple source tracking is to estimate the state sets $\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{M_k,k}\}$ based on the available observation sets $\mathcal{Z}_k$ for all time steps $k = 1, \ldots, K$. From time step $k-1$ to time step k, the randomness in the system consist of:

- sources may die;
- sources may survive and evolve to a new state;
- new source may appear;
- each source may or may not generate a measurement, and the measurements can be obscured by false alarms or clutter.

Fig. 3.6 gives an illustration about this randomness in the multiple source tracking problem. In the target space, the source may die, survive or evolve to a new state, and new source may appear. Source detections, as well as false alarms are presented in the measurement space.

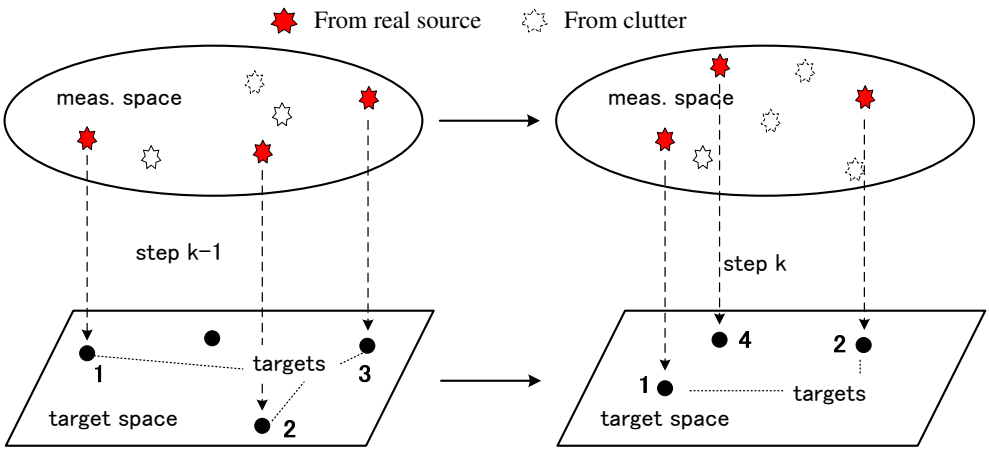


Figure 3.6: *RFS formulation for multiple sources tracking. The source may die, survive or evolve to a new state, and new source may appear. False alarms may be presented in the measurement space.*

3.4.2 Multiple source Bayesian filtering

The task of multiple source filtering is to estimate the number of sources and the corresponding states from the measurement set jointly. In brief, let $\mathcal{Z}_{1:k} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_k\}$ be a time sequence

of the measurement sets. Assume that the multi-source RFS transition density is

$$p_{k|k-1}(\mathcal{X}_k|\mathcal{X}_{k-1}), \tag{3.76}$$

and the multi-source RFS likelihood function

$$p_k(\mathcal{Z}_k|\mathcal{X}_k). \tag{3.77}$$

The multisource Bayesian filter then has the form

- predict:

$$p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{Z}_{1:k-1}) = \int p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{X}_{k-1}) p_{k-1|k-1}(\mathcal{X}_{k-1}|\mathcal{Z}_{1:k-1}) \delta\mathcal{X}_{k-1}, \tag{3.78}$$

- update:

$$p_{k|k}(\mathcal{X}_k|\mathcal{Z}_{1:k}) = \frac{p_k(\mathcal{Z}_k|\mathcal{X}_{k|k-1}) p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{Z}_{1:k-1})}{p_k(\mathcal{Z}_k|\mathcal{Z}_{1:k-1})}, \tag{3.79}$$

where $\int f(\mathcal{X})\delta\mathcal{X}$ is a set integration which is calculated over the set space $\mathcal{F}(\mathcal{X})$, and $p_k(\mathcal{Z}_k|\mathcal{Z}_{1:k-1})$ is a Bayes normalization factor, given as

$$p_k(\mathcal{Z}_k|\mathcal{Z}_{1:k-1}) = \int p_k(\mathcal{Z}_k|\mathcal{X}_{k|k-1}) p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{Z}_{1:k-1}) \delta\mathcal{X}_{k|k-1}. \tag{3.80}$$

The derivations of the RFS transition density (3.76) and the likelihood (3.77) depend on the exact source dynamical models and the measurement models, and will thus be given in Chapter 6 where all of these models are clearly defined. In the case that only one source exists, i.e., a singleton $\{\mathbf{x}_k\}$, and no source birth or death happens, the above RFS Bayesian predict and update steps will be simplified to a standard Bayesian filter in the single source scenario as described in Section 3.1.2.

3.4.3 Sequential Monte Carlo implementation

Sequential Monte Carlo approaches have been shown to be effective in handling the nonlinear system in the single source tracking scenario [6, 12]. Similar to the single Bayesian filtering scenario, the multi-source Bayesian recursion in (3.78) and (3.79) can also be computed by using a SMC approximation [66, 98]. However, in the context of multi-source filtering, each

particle is constructed by a finite set and particles themselves have varying dimensions. In this section, the RFS SIR particle filtering method is briefly introduced as follows [98].

Assume that at time step k, a set of particles and the corresponding weights are available: $\{\mathcal{X}_k^{(i)}, w_k^{(i)}\}_{i=1}^N$. The multiple source posterior pdf $p(\mathcal{X}_k|\mathcal{Z}_{1:k})$ can be approximated as

$$p_{k|k}(\mathcal{X}_k|\mathcal{Z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta_{\mathcal{X}_k^{(i)}}(\mathcal{X}_k), \tag{3.81}$$

where $\delta_{\mathcal{X}_k^{(i)}}(\cdot)$ is a set-valued Dirac delta function with value 1 at set region $\mathcal{X}_k^{(i)}$ and 0 otherwise. Suppose the particles are sampled according to

$$\mathcal{X}_k^{(i)} \sim q_k(\cdot|\mathcal{X}_{k-1}^{(i)}, \mathcal{Z}_k). \tag{3.82}$$

Following the importance weight derivation (3.50) in the single source scenario, one can obtain the RFS weight updating similarly

$$w_k^{(i)} = \tilde{w}_{k-1}^{(i)} \frac{p_k(\mathcal{Z}_k|\mathcal{X}_k^{(i)}) p_{k|k-1}(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)})}{q_k(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)}, \mathcal{Z}_k)}, \tag{3.83}$$

where $\tilde{w}_{k-1}^{(i)}$ is the normalised weight term such that $\sum_{i=1}^N \tilde{w}_{k-1}^{(i)} = 1$. The Bayesian recursion of (3.81) can be written as

$$p_{k|k}(\mathcal{X}_k|\mathcal{Z}_{1:k}) \approx \sum_{i=1}^{N} \tilde{w}_{k-1}^{(i)} \frac{p_k(\mathcal{Z}_k|\mathcal{X}_k^{(i)}) p_{k|k-1}(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)})}{q_k(\mathcal{X}_k^{(i)}|\mathcal{X}_{k-1}^{(i)}, \mathcal{Z}_k)} \delta_{\mathcal{X}_k^{(i)}}(\mathcal{X}_k). \tag{3.84}$$

The same as a standard SIR particle filtering, the main practical problem with the RFS SIR particle filtering is difficult to find an efficient importance density. A naive choice is generating the particles according to the transition density, by which the importance function can be written as

$$q_k(\cdot|\mathcal{X}_{k-1}^{(i)}, \mathcal{Z}_k) = p_{k|k-1}(\cdot|\mathcal{X}_{k-1}^{(i)}), \tag{3.85}$$

and the weight can be simplified as

$$w_k^{(i)} = \tilde{w}_{k-1}^{(i)} p_k(\mathcal{Z}_k|\mathcal{X}_k^{(i)}). \tag{3.86}$$

The essential ingredient for the RFS SIR particle filtering is thus formulating the likelihood

$p_k(\mathcal{Z}_k|\mathcal{X}_k^{(i)})$. It is worth pointing out that the computations of $p_k(\mathcal{Z}_k|\mathcal{X}_k^{(i)})$ are exponential in the cardinality of the state set $|\mathcal{X}_k^{(i)}|$ [1]. Moreover, as for the single source SMC scenario, using the transition density as the importance function will lead to a decrease in efficiency, which is typically exponentially with the number of sources given the fixed number of particles. Further explanation of these merits and drawbacks of implementing a RFS SIR particle filtering will be given in Chapter 6 when formulating the multiple acoustic source tracking problem.

3.5 Chapter summary

In this Chapter, the background knowledge for Bayesian filtering is presented. The Kalam filter and its variants to approximate the moderate nonlinear systems are introduced. The more advanced approaches, sequential Monte Carlo methods, which are completely appropriate for the nonlinear systems are also addressed. Section 3.2 is essentially important for deriving the particle filtering approaches for the single/nonconcurrent multiple acoustic source tracking problem in Chapter 5.

Further, by recognising a special state estimation problem in which the states can be split into two parts so that one of them can be marginalised out analytically and only the rest of the states need a Monte Carlo approximation, the Rao-Blackwellised particle filtering is fully illustrated in that it is able to reduce the estimation variance in such a case. Subsequently, a brief introduction of the multiple source Bayesian filtering and its SMC implementation are presented. Section 3.3 and Section 3.4 are the basic materials for developing a particle filtering based approach for multiple time-varying number of acoustic source tracking in Chapter 6.

Chapter 4
Experimental studying of the TDOA measurements

Time-delay of arrival is found an attractive measurement in acoustic source tracking problem, in that it is simple and can easily be estimated in a number of speech applications. To construct a tracking system, it is important to fully examine the TDOA measurement extraction methods in different noisy and reverberant environments. In this chapter, a series of experimental parameters, including the signal to reverberation ratio (SRR) and signal to noise ratio (SNR), and the probabilities of detection and false alarm, are first defined. The phase-transform weighted generalised cross correlation (PHAT-GCC) method is then introduced and some preliminary studies are presented. Knowing that speech mixtures are window-disjoint-orthogonal (WDO) in the time-frequency domain, the degenerate unmixing estimation technique (DUET) is employed to separate the mixed speech spectrogram and a DUET-GCC method is proposed to extract the TDOA measurements for multiple simultaneously active sources. The performance of these two methods are subsequently investigated under different simulated noisy and reverberant environments. Finally, the real audio lab environment is studied and the TDOA measurements extracted by using the PHAT-GCC and DUET-GCC methods based on the real recorded signals are analysed.

4.1 Parameter definition

Several experimental parameters are defined in this section to investigate the performance of the TDOA measurements. The simulated adverse environments are generated by setting different levels of SNR and SRR. Different SNRs are used to depict the noisy environments, and SRRs are used to generate the reverberant environments for a static source. When a source is moving in the room, the geometry between the source and microphone receivers are time-varying, and a broad range of SRR will appear. The reflection coefficients ρ and the corresponding reverberation time T_{60} are thus used to describe the room reverberant environment. The probabilities

of detection and false alarm are also defined here for evaluating the performance of the TDOA measurements given an experiment.

4.1.1 Noisy and reverberant environments

Different noisy environments can easily be simulated by setting different signal-to-noise ratio; see Appendix B for the detailed definitions. It is achieved by adding white Gaussian noise (WGN) with different energy level into the received sinal. The reverberation time T_{60} is usually used to evaluate the reverberation in the room environment. For the tracking problem, the distance between the source and the microphones can change drastically due to the movement of the source. The performance of TDOA estimates thus varies even in the same T_{60} environment, as shown, in Fig. 4.1(a). The source moving trajectory and microphone positions are illustrated in Fig. 2.9, on page 44. The source follows the trajectory 2, and the signals received at the first microphone pair (microphone 1 and microphone 2) are used to estimate the TDOAs. The reflection coefficient of the walls is 0.8, which leads to a reverberation time T_{60} of 0.289s. The TDOA estimation is better at the lower left corner where the source is close to the microphone receiver. However, at the far-end area, the upper right corner, the TDOA estimation is deteriorated significantly.

The effect of distance and reflection coefficient on reverberation can be summarised by one parameter: the signal to reverberation ratio. Given the reflection coefficient ρ and the distance r between the source and the sensor, the SRR can be defined as [99]

$$\text{SRR} = 10\log_{10}\left(\frac{\mathcal{A}(1-\rho^2)}{16\pi r^2\rho^2}\right), \tag{4.1}$$

where $\mathcal{A}$ is the whole wall reflection area. Detailed derivation of equation (4.1) can be found in Appendix B. The expression (4.1) indicates that the TDOA performance are affected by the room dimension, reflection coefficients and the distance between the source and the microphones as well. Fig. 4.1(b) presents the SRRs of the simulated room environment (Fig. 2.9 on page 44) versus the distance r schematically. It shows that in the same room environment, SRR increases quadratically as the decreasing of the distance between the source and the sensor. Better TDOAs are thus likely to be obtained when the source is located at the close-end, and vice versa.

In the following sections, different SNRs and SRRs will be used to generate the simulated noisy

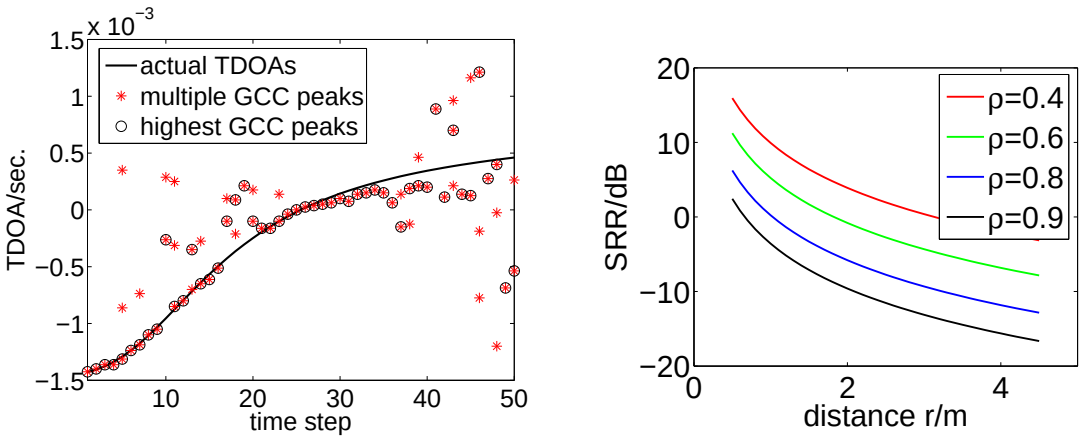


Figure 4.1: *Preliminary experimental study of the simulated room environment (Fig. 2.9 in Section 2.6.2 on page 44). (a) TDOA measurements from the first microphone pair (microphone 1 and microphone 2). The Source is close to the microphone receiver at lower left corner, and far away from the microphone at the upper right corner (following trajectory 2). The reverberation time T_{60} in the room is 0.289s; (b) SRR* vs. *the distance between the source and the microphone receiver under different reflection coefficients.*

and reverberant environments respectively. We will fully examine how much the TDOA measurements will be distorted under these predefined adverse environments, and how the probabilities of detection and false alarm are affected.

4.1.2 Anomaly TDOA estimates

Given a TDOA estimate $\hat{\tau}$, whether it is a detection or not is dependent on how far it diverges from the ground truth τ. The authors in [38, 100] present a definition of anomaly TDOA estimation, which regards any TDOA estimates diverging further than $T_c/2$ away from the ground truth as anomaly ones. This definition is also used in [101, 102] to evaluate the performance of the TDOA estimation.

The signal correlation time T_c is computed as the bandwidth of 3dB degradation of the main lobe in the autocorrelation function [38, 100]. Fig. 4.2 shows the autocorrelation calculated by a long-time average of 64 speech signals from TIMIT database [103], which leads to 1962 frames. The autocorrelation function is calculated according to equation (2.17), on page 21. The 3dB degradation results a T_c of 2 samples roughly, which is the same as the T_c estimation in

[101,102]. However, this definition of anomaly TDOA estimation does not take the microphone separation into account, and will result in an inconsistent anomaly error. For example: given a sampling frequency of 8kHz, the signal correlation time T_c will be 0.25ms; if the microphone separation is 1m, the maximum TDOA would be $1/343 \approx 3$ms, and among all the TDOA estimates, only a small portion of TDOAs are regarded as within the detection range; however, if the microphone separation is 0.1m, the maximum TDOA will be $0.1/343 \approx 0.3$ms, and this leads to a detection range 10 times as that of 1m microphone separation. To keep the consistency of the probabilities of detection and false alarm over all the possible microphone separations, the anomaly error can be normalised as follows

$$\epsilon = \frac{T_c}{2}\frac{d}{d_{\text{ref}}}, \tag{4.2}$$

where d is the microphone pair separation, and d_{ref} is a reference separation where the standard correlation error $T_c/2$ will apply. Usually $d_{\text{ref}} = 1$m is found reasonable and sufficient to analyse the anomaly percentage [38, 104]. For all the TDOA measurements collected from a microphone pair, at most one (if it is located within the anomaly range ϵ from the ground truth) is regarded as a detection for a source.

The definition of anomaly error can only be used to evaluate the percentage of the correct TDOA estimation. For a tracking problem, the final result is determined by both the probability of correct estimation (probability of detection) and the probability of false estimates (probability of false alarm). In the next section, the probabilities of detection and false alarm are thus presented.

4.1.3 Probability of detection and false alarms

In practice, a set of peaks from GCC function is picked to include the potential TDOA measurements as complete as possible. Given a threshold value R_{TH}, the measurement set obtained across the ℓth microphone pair can be written as

$$\mathbf{z}_k^\ell = \{\tau : R_\ell(k,\tau) > R_{\text{TH}}\}, \tag{4.3}$$

where $R_\ell(k,\tau)$ is the GCC function calculated from equation (2.28) in Section 2.28 on page 25. If there is no such a peak larger than the threshold, the largest peak in the GCC function will be picked. The measurement set contains detections as well as false alarms. It is desired to know

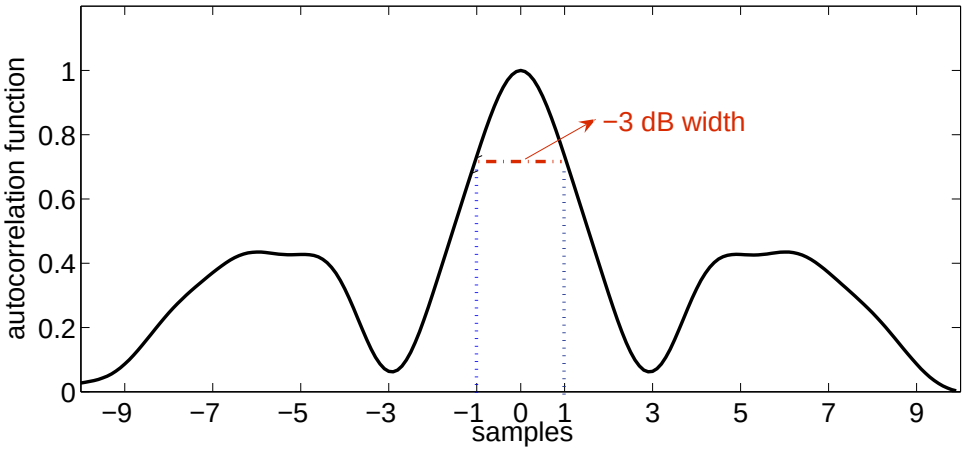


Figure 4.2: *Autocorrelation calculated by a long-time average of 1962 frames; -3 dB decay below the peak leads to 2 samples width of the main lobe.*

exactly the probability of detection and the probability of false alarms in different scenarios, since these two parameters are what the final tracking performance are relying on.

For any TDOA measurement $\hat{\tau}$, if it is located within the range of anomaly error ϵ from the ground truth τ, it is regarded as a detection of the true TDOA estimate. Suppose that $\kappa = 1$ denotes a detection, and 0 otherwise. This definition of a detection can be written as

$$\kappa = \begin{cases} 1, & |\hat{\tau} - \tau| < \epsilon; \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Suppose n_k^ℓ measurements are picked from the GCC function at the ℓth microphone pair, given as

$$\mathbf{z}_k^\ell = \{\hat{\tau}_{1,k}^\ell, \ldots, \hat{\tau}_{n_k^\ell,k}^\ell\}, \tag{4.5}$$

and $|\mathbf{z}_k^\ell| = n_k^\ell$. Following the expression in (4.4), the corresponding detection indicator $\kappa_{i,k}^\ell$, $i = 1, \ldots, n_k^\ell$ can be obtained for each TDOA measurement, i.e., $\kappa_{i,k}^\ell = 1$ if the corresponding TDOA estimation is a detection, and $\kappa_{i,k}^\ell = 0$ otherwise. The probability of detection is thus

$$P_{\mathrm{D}} = \frac{\sum_{i,\ell,k} \kappa_{i,k}^\ell}{LK} \times 100\%, \tag{4.6}$$

and the probability of false alarm is calculated as

$$P_{\mathrm{F}} = \frac{\sum_{i,\ell,k}(1-\kappa_{i,k}^{\ell})}{\sum_{\ell,k} n_k^{\ell}} \times 100\%, \tag{4.7}$$

where L is the total number of microphone pairs and K is the number of time steps. The probability of detection will be its maximum value of 1 if the TDOA is always detected, and 0 if no TDOA detection can be collected. The more peaks are picked and the fewer of them are actual TDOAs, the higher the probability of false alarm will be, and vice versa. Obviously, the probability of detection P_{D} and the probability of false alarm P_{F} depend on the value of the threshold R_{TH}. A lower threshold will lead a higher probability of detection, but the probability of false alarm increases as well. It thus requires that the following tracker is more capable of discriminating those false alarms. A higher threshold can always exclude the false alarms, but may cause a number of miss detections. In such cases, the tracker should be able to smooth the trajectory to compensate the measurement missing. In Chapter 6, a detailed discussion about how the algorithm is affected by the false alarm and the miss detection will be given.

4.1.4 ROC curve

The final probabilities of detection and false alarm depend on the threshold and the distribution of the detections and the false alarms. Fig. 4.3 illustrates the relations between the probability of detection and false alarm and the distributions of the detections and false alarms. In Fig.

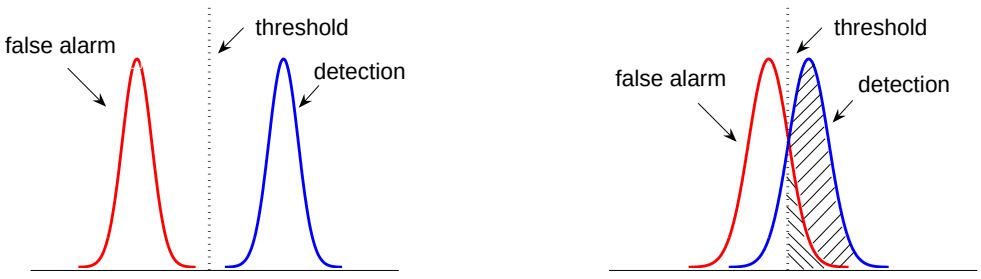


Figure 4.3: *Different distribution of detections and false alarms; (a) the source detections and false alarms are perfectly separated; (b) the overlap presence in the source detections and false alarms. As the threshold decreases, the better detection can be achieved. However, the false alarm rate will increase as well.*

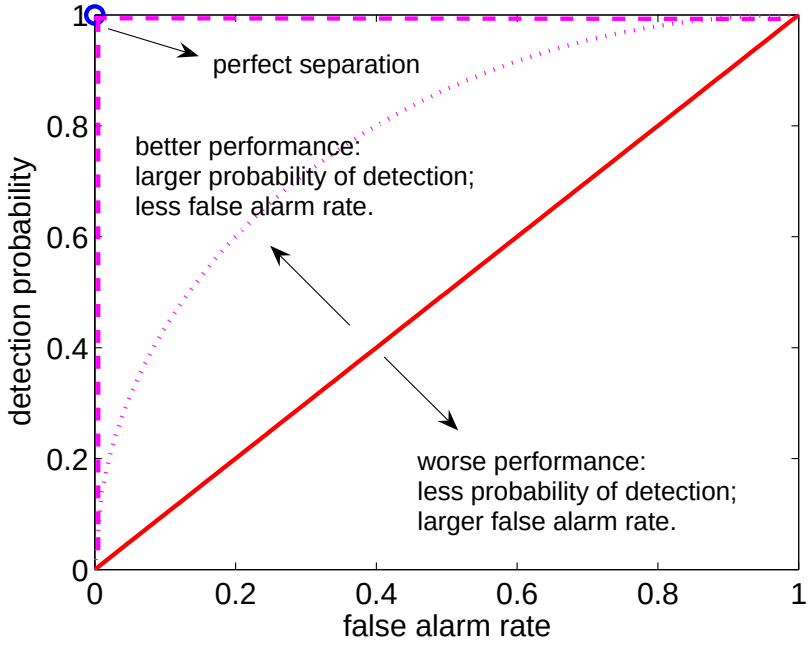


Figure 4.4: *ROC curve interpretation of the probability of detection and false alarm. Perfect separation is achieved at top left corner, where the probability of detection is one, and false alarm rate is zero.*

4.3(a), the detections and false alarms are perfectly separated. The threshold is thus easily to set to pick up all the detections and meanwhile exclude the false alarms completely. However, when the large overlap presents in the detections and false alarms, as shown in Fig. 4.3(b), the perfect threshold is impossible to achieve. To gain a higher probability of detection, the threshold has to be lower, and this will lead to a higher false alarm rate simultaneously. On the other hand, to exclude the false alarms perfectly, the threshold has to be relatively high, and this may cause a significant detection missing.

In signal detection theory, the receiver operating characteristic (ROC) curve is used to represent the probability of detection vs. the false alarm rate (or the correct rejection rate vs. the probability of detection missing) for a binary classifier system as the threshold varies [105, 106]. Fig. 4.4 gives an illustration of the ROC curve. Any coordinate closer to the upper left corner represents a higher probability of detection and lower false alarm rate, and vice versa. The perfect separation is achieved when the ROC point is located exactly at the top left corner, where the probability of detection is 1 and the false alarm rate is zero. The red line indicates that the distribution of the detections and false alarms completely overlapped, and for all threshold values, the probability of detection and false alarm are equal. The ROC curve for the distribution

of detections and false alarms as Fig. 4.3(a) is the dashed vertical-horizontal line illustrated in the Fig. 4.4. As the threshold decreases, the probability of detection become less, and the false alarm will appear when the threshold below a certain level. The ROC curve for Fig. 4.3(b) is an arch line similar as the dotted line in Fig. 4.4, where there is a large section that the false alarms and detections simultaneously exist. As this ROC plot gives an explicit illustration of the probabilities of detection and false alarm, it will be used in this thesis to evaluate the performance of TDOA measurements.

4.2 PHAT-GCC based TDOA measurements

Among all the TDOA estimation approaches, the generalised cross correlation method is very popular and widely used due to its simplicity and robustness in moderate noisy and reverberant environments. A number of pre-filtering weighting terms have been developed in GCC method to emphasise the peak from the true delay and suppress the effects from the speech signal itself and the noise in estimating the cross-correlation function [31]. It is observed that the phase transform (PHAT) weight performs more consistently than many other GCC weighting terms when the statistics of the source signal is unavailable *a priori* and varying in time [107, 108]. In this section, the PHAT-GCC method is studied in detail.

4.2.1 PHAT-GCC

The PHAT-GCC method employs the amplitude component at each frequency to normalise the cross-correlation function. The PHAT weight can be addressed as

$$\begin{aligned}\Phi_\ell(k,\omega) &= \frac{1}{|G_\ell(k,\omega)|} \\ &= \frac{1}{|Z_{\ell,1}(k,\omega)Z^*_{\ell,2}(k,\omega)|},\end{aligned} \tag{4.8}$$

where G is the cross spectral and Z is Fourier transform of the received frame signal; see Section 2.2.3 on page 20 for detailed definitions. Suppose the attenuations to the two microphones of a given microphone pair ℓ are $a_{\ell,1}$ and $a_{\ell,2}$ respectively. The GCC function in (2.28) on

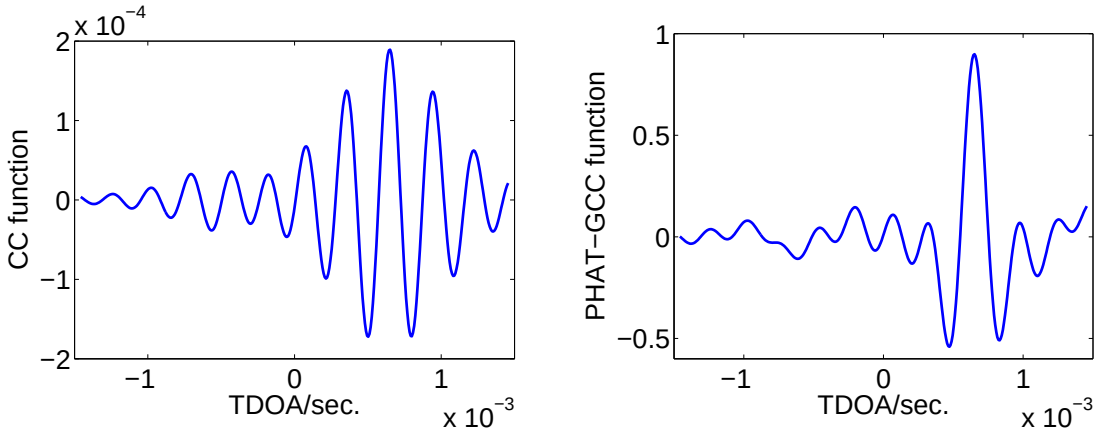


Figure 4.5: *(a) CC function and (b) PHAT-GCC function for the same frame of speech signal. The largest peak corresponds the ground truth TDOA. The periodical peaks in CC function are well removed by the PHAT pre-filtering, and a clearer peak is exhibited by PHAT-GCC function.*

page 25, is thus

$$\begin{aligned} R_\ell(k,\tau) &= \int_\Omega \frac{G_\ell(k,\omega)}{|G_\ell(k,\omega)|} e^{j\omega\tau} d\omega \\ &= \int_\Omega \frac{a_{\ell,1}a_{\ell,2}G_{ss}(k,\omega)e^{-j\omega\tau_\ell(k)}}{|a_{\ell,1}a_{\ell,2}G_{ss}(k,\omega)e^{-j\omega\tau_\ell(k)}|} e^{j\omega\tau} d\omega \\ &= \int_\Omega e^{-j\omega\tau_\ell(k)} e^{j\omega\tau} d\omega \\ &= \delta_{\tau_\ell(k)}(\tau). \end{aligned} \tag{4.9}$$

where $G_{ss}(k,\omega)$ is the power spectral density of the source signal. Since the cross spectral is weighted by the reciprocal of its modulus, the amplitude of each frequency component is reduced to one, and only the phase information $\omega\tau_\ell(k)$ left. The GCC function finally turns out to be a dirac function $\delta(\cdot)$ with value one at $\tau = \tau_\ell(k)$ and zero elsewhere. Rather than the CC function, which is affected by the attenuations and the source signal itself, the PHAT-GCC method exhibits a pure delay pulse and the affect from the speech signal and attenuation are canceled.

Voiced speech frames are always dominated by one or several principal frequencies, and very weak at other frequency components. For the CC method, the cross-correlation function in the time domain is thus periodically generated by these principal frequencies, rather than by

all the frequency components together. This phenomenon is clearly shown in Fig. 4.5(a), in which several periodical peaks are exhibited other than the peak from the ground truth TDOA. The PHAT-GCC method has the advantage to emphasise the true TDOA and suppress those periodical peaks. Since it discards the amplitude information of each frequency component, all the frequency components are whitened and contribute to the final CC function equally, as shown, in Fig. 4.5(b).

However, since the modulus of all the frequency components are equalised to one, the PHAT-GCC method may also amplify the erratic phase information from those frequency components which have little energy, e.g., the frequency components at higher frequency band. Sometimes these frequency components are poor in speech signal (may be purely noise) and with very small modulus. The modulus cancelation at these frequency components may degrade the TDOA estimation accuracy. A maximum likelihood (ML) based weighting term, which roughly equals to the SNR at each frequency component, is presented in [31]. Under the assumptions of uncorrelated, stationary Gaussian signal and noise, it is an ideal weighting function since theoretical variance bound can be achieved. It should be pointed out that this approach requires an estimation of the coherence function to achieve the optimal performance, and thus needs a large sample space at each time step. Moreover, the assumptions for the ML weight are very strict, and always violated in a real speech environment. For a number of speech applications, the PHAT-GCC remains one of the most popular TDOA estimation approaches due to its easy implementation and simplicity.

4.2.2 Preliminary studying of the PHAT-GCC amplitude

In an anechoic environment, there is merely a time-delay between two received signals. In such a case, a sharp peak will be exhibited in the PHAT-GCC function representing the TDOA measurement. However, real room acoustic environments are always challenging, with unexpected noise and multipath components. In such scenarios, the peak of PHAT-GCC function may be distorted, and incorrect TDOA estimates will present. Fig. 4.6 shows the PHAT-GCC function estimated from microphone pair 1 under an adverse environment. The source position is $(2.5,\ 3.0)$m, which leads to a TDOA of 0.64ms. The SNR is 0dB, and the reflection coefficient is set to 0.8. Not only the actual peak is distorted, but a number of false peaks exhibit, and some of them are as high as or even higher than the actual peak. This makes estimating the TDOA extremely difficult.

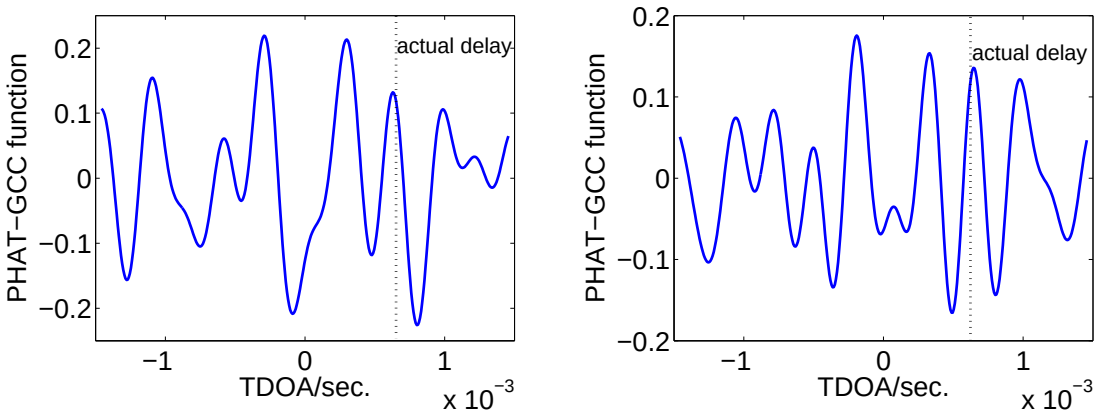


Figure 4.6: *PHAT-GCC function under (a) reverberant environment (ρ=0.8); and (b) noisy environment (SNR=0dB). The ground truth of TDOA is 0.64ms. The actual peak is distorted by the noise and reverberation, and even worse, false peaks are presented – some of these false peaks are even higher than the peak corresponds to the actual TDOA.*

Since we will threshold the peaks in the PHAT-GCC function, the studying of the amplitude of GCC function is naturally interesting and necessary. Given a set of TDOA estimates $\{\hat{\tau}^{\ell}_{1,k}, \ldots, \hat{\tau}^{\ell}_{n^{\ell}_k,k}\}$ which are estimated according to equation (4.3), suppose the corresponding PHAT-GCC amplitudes of these TDOA estimates are $\{\hat{a}^{\ell}_{1,k}, \ldots, \hat{a}^{\ell}_{n^{\ell}_k,k}\}$ at time step k. The root-mean-square (RMS) amplitude of a batch of GCC peaks is calculated as

$$\bar{a}^{\ell}_k = \sqrt{\frac{1}{n^{\ell}_k} \sum_{i=1}^{n^{\ell}_k} (\hat{a}^{\ell}_{i,k})^2}. \tag{4.10}$$

Following the definition of the detection rate and false alarm rate, the amplitude from the source and clutter can also be determined. That is

$$\hat{a}^{\ell}_{i,k} \text{ is generated by } \begin{cases} \text{a source}, & |\hat{\tau}^{\ell}_{i,k} - \tau_{\ell}(k)| < \epsilon; \\ \text{a clutter}, & \text{otherwise}, \end{cases} \tag{4.11}$$

where ϵ is already defined in Section 4.1.3. Given this classification, the RMS amplitude from the source and the clutter can be obtained: simply take the summation in (4.10) over the collection of source generated amplitudes and clutter generated amplitudes respectively.

Figure 4.7 shows the RMS amplitude under different noisy and reverberant environments. For

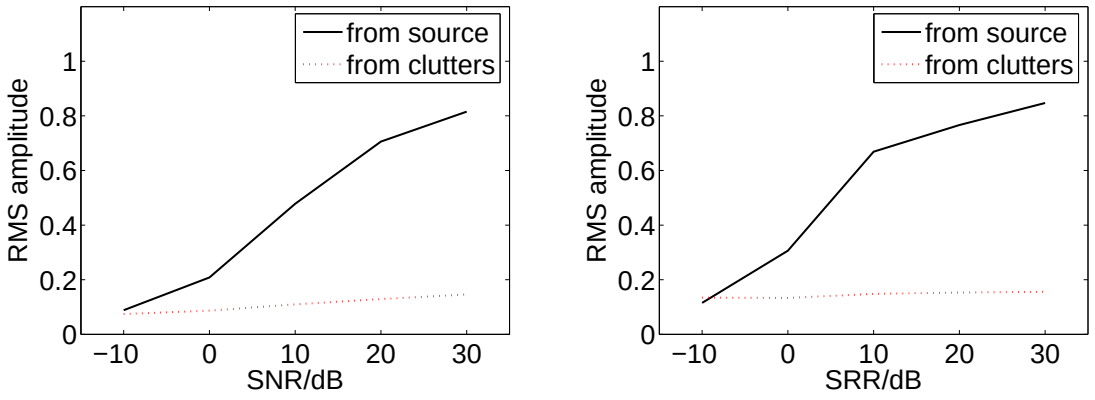


Figure 4.7: *The RMS amplitudes generated by the source and the clutter; (a) under different SNR environments; (b) under different SRR environments.*

different SNRs, the source is located at $(2.5,\ 3.0)$m, and the reflection coefficients are set to zero. The parameters (source position and wall reflection coefficients) to generate different SRRs are illustrated in Table 4.1. For those peaks generated by a source, the corresponding RMS amplitude is higher than those generated by the clutters in the moderate adverse environments. It is thus possible to threshold the peaks from the PHAT-GCC function to obtain the TDOA measurements. However, when the SRR or SNR is very low, the RMS amplitudes generated by the source and clutter will be close, which means the peaks generated by the clutter may be as high as, or even higher than the peaks generated by the real source. In such cases, detecting the TDOAs will be very difficult. Since the RMS amplitudes change in different SNR and SRR environments, an appropriate threshold should be carefully picked to balance the probability of detection and false alarm rate.

SRR(dB)	-10	0	10	20	30
ρ	0.8	0.8	0.6	0.6	0.1
$x-y$ position	$(3.6, 3.01)$	$(0.7, 3.01)$	$(0.8, 1.51)$	$(0.8, 1.91)$	$(0.9, 2.11)$

Table 4.1: *Corresponding SRRs generated by different combinations of the source positions and wall reflections.*

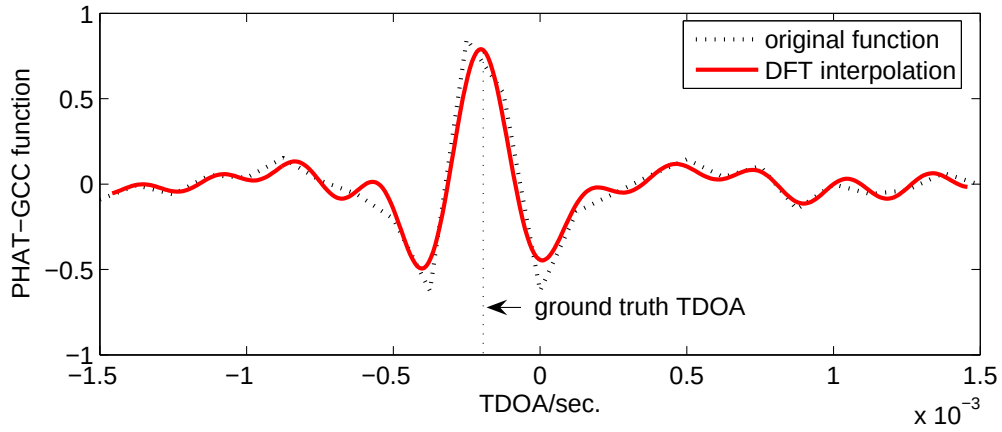


Figure 4.8: *DFT interpolation for the PHAT-GCC function. The ground truth TDOA is -0.2ms. After the interpolation, the peak is more smooth and is able to indicate an accurate TDOA estimation.*

4.2.3 Microphone separation and TDOA resolution

Apart from the noise and reverberation, the performance of the GCC is also affected by the interpolation method. GCC method usually requires a high time-delay resolution to discriminate the peaks and to match these peaks with the corresponding TDOAs accurately. In all of our experiments, the GCC function are manipulated by a 16-point DFT interpolation. This is achieved by calculating the Fourier coefficients at all required discrete frequency components. The value at an interpolation point is then obtained by taking an inverse Fourier transform with corresponding number of points. The TDOA resolution after 16-point interpolation is found accurate enough to present correct TDOA measurements. Fig. 4.8 gives a comparison between the original PHAT-GCC function and the PHAT-GCC function after DFT interpolation. The ground truth TDOA is -0.2ms. It shows that after the interpolation, the peak is more sharp and is able to indicate a TDOA estimation accurately.

Obviously, the microphone separation d has a significant impact on the time-delay resolution. In this experiment, the TDOA performance of the PHAT-GCC method under different microphone separations is studied. Two dynamical sources moving with a trajectory depicted in Fig. 2.9 is simulated. The source signals are from the TIMIT speech database [103]. The frame length is set to 1024 samples, and the sampling frequency is 8kHz. The microphone separation varies from 0.1m to 0.9m, with an increment of 0.2m. Figure 4.9 presents the ROC curve versus different microphone separations. The ROC curve is construct by setting different thresholds, from 0.1 to 0.9 with an increment of 0.1, and also 0 and 0.99. The microphone distance ob-

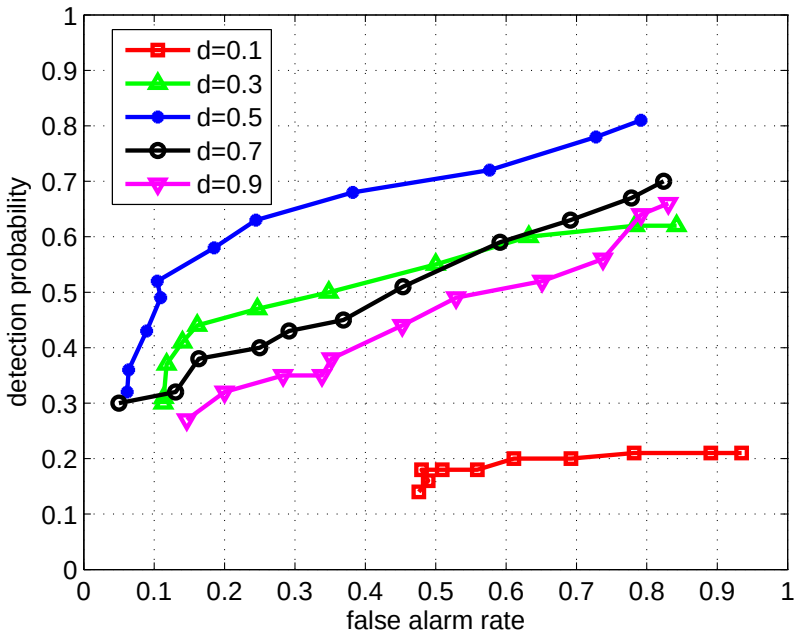


Figure 4.9: *ROC curve interpretation of the probabilities of detection and false alarm under different microphone separations for the PHAT-GCC method.*

viously effects the probability of detection and false alarm. Among all these separations, the best TDOA performance is achieved at a microphone distance of 0.5m. For the separation less than 0.5m, especially with a separation of 0.1m, the probability of detection is low since the TDOA resolution is not satisfied. On the other hand, an overly large microphone separation will destroy the correlation property between two received signals, and thus decrease the TDOA performance.

The probability of detection can be enhanced by setting a low threshold. For the microphone separation of 0.5m, the probability of detection, P_D, can be as large as 0.8 when the threshold is less than 0.5. However, the false alarm rate increases significantly; even larger than that of the detection probability. This is usually unacceptable for the subsequent tracking algorithm. Such a high false alarm rate makes it impossible for the tracker to differentiate the detections and the false alarms, and thus difficult to filter the source states. This study is of great interest when setting up the experiment system. In our next experiments, a microphone separation around 0.5m will be used throughout to simulate the speech signals for the PHAT-GCC method.

4.3 TF masking based multiple source TDOA estimation

The GCC method is robust for the single source case and is widely used as the measurement extraction approach in the tracking problem. However, in the scenario that multiple speakers exist and especially when the speakers are simultaneously active, the performance of the GCC based TDOA measurements become severely degraded in that:

1. Cross-correlation based TDOA estimation is not appropriate for multiple speech sources since it theoretically assumes only one impinging wavefront arrives [31]. The multiple speech source signals are not always independent in the time domain, and thus the correlation function may not yield sharp peaks for each actual time-delay.

2. The cross-correlation function always has a time resolution problem. As the moving speakers are closely spaced in the room, it is almost impossible to report multiple TDOAs accurately [41].

In order to circumvent this problem, the degenerate unmixing estimation technique (DUET) [54, 109] which is a powerful tool for separating the sources from the received mixtures is introduced here. It assumes that the source mixtures are window-disjoint-orthogonal (WDO) on the TF domain. Fig. 4.14(a) gives an illustration on how the speech mixtures are separated on the TF representation. The DUET is achieved by clustering the TF spectrogram bins of each source to form a two dimensional (2-D) time-delay and gain-ratio histogram (as shown, for example, in Fig. 4.14(b)). Although the original purpose of WDO is for source separation, it is also employed to localise multiple sources in [110, 111]. The authors in [110] developed a TDOA estimation approach for multiple sources based on WDO assumption and a steered response power (SRP) PHAT method, and use a KF to localise the direction of arrival. Although the TDOA measurement extraction approach is similar to the DUET-GCC method developed in this chapter, it does not take the phase ambiguity problem into account and can only be applied for the arrays with small microphone separations (4cm in [110]). Mandel et al [111] also build a probabilistic model of interaural level and phase differences and use an expectation-maximization (EM) algorithm to find the TDOAs for multiple sources. Both of these methods need a burn-in period to converge to the TDOA estimates, and are thus only appropriate for the localisation problem.

When the WDO assumption is satisfied, DUET based TDOAs are more reliable than the GCC based ones in that:

1. The TF masking method is appropriate for the measurement extraction of multiple sources, even when these sources are simultaneously active.

2. By clustering the TF bins, we obtain the measurements from the source signal but get rid of those bins generated by noise. For this reason, calculating the TDOAs using TF masks has an advantage by yielding robust estimates even at low SNRs.

3. Since the TDOAs are calculated from the phase term directly, it also has the advantage of achieving higher accuracy than the TDOAs estimated by the GCC method.

In this section, the DUET separation based TDOA measurement extraction approach is introduced.

4.3.1 WDO assumption

Assuming that $S_i(k,\omega)$ and $S_j(k,\omega)$ denote the short time Fourier transform of speech frames from two different source signals $\mathbf{s}_i(k)$ and $\mathbf{s}_j(k)$. The WDO property of speech signal can be stated as

$$S_i(k,\omega)S_j(k,\omega) = 0, \quad \forall\ (k,\omega) \text{ and } i \neq j;\ i,j = 1,\ldots,M_k. \tag{4.12}$$

where M_k is the number of sources. Equation (4.12) states that the spectrograms of multiple sources do not overlap in the time-frequency domain. The WDO property is clearly shown in Fig. 4.10, where the spectrograms of speech mixtures are sparse and disjoint. For two speech signals, the product of the corresponding spectrograms is zero at the most area on the TF domain. For each TF bin, the interfering signal can be defined as

$$Y_i(k,\omega) = \sum_{\substack{k=1 \\ k\neq i}}^{M_k} S_k(k,\omega). \tag{4.13}$$

The idea binary mask $\Lambda_i(k,\omega)$ which indicates whether the source is active in this TF bin can thus be formulated as

$$\Lambda_i(k,\omega) = \begin{cases} 1, & 20\log\left(\frac{|S_i(k,\omega)|}{|Y_i(k,\omega)|}\right) \geq \eta; \\ 0, & \text{otherwise}, \end{cases} \tag{4.14}$$

where η is a threshold, with a value of 0dB is always used to regard the TF bins in which the source energy is larger than the energy of all other interfering signals as the active source bins. Suppose that the source signals are known, the masks of each source can thus be evaluated

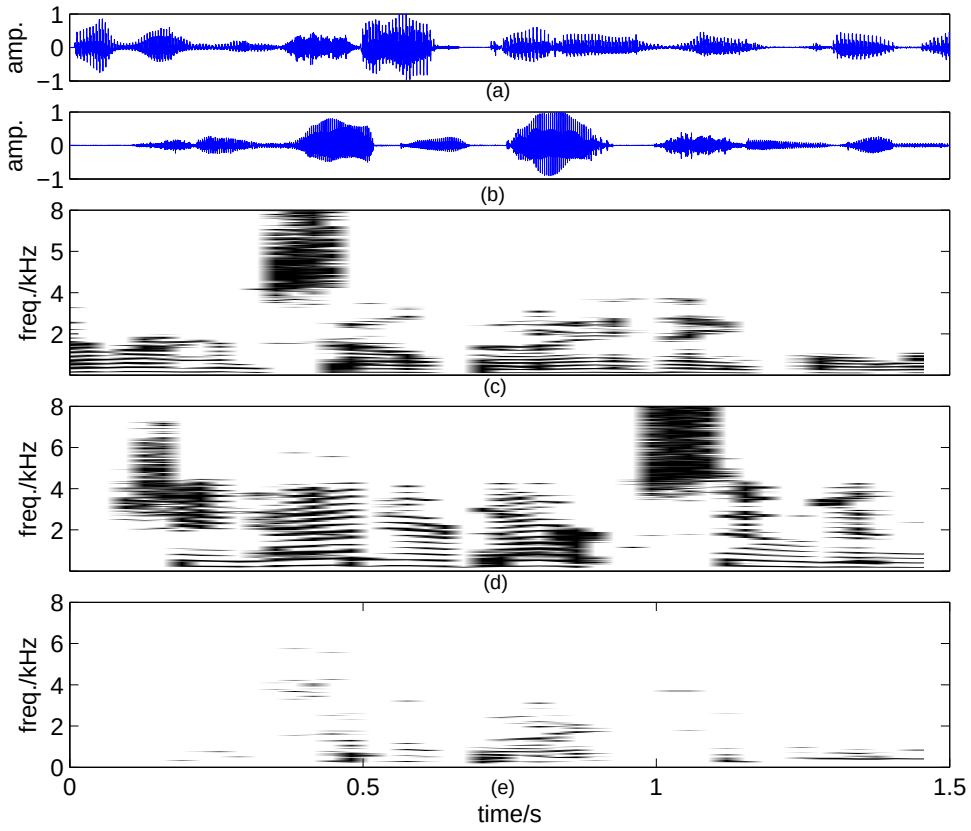


Figure 4.10: *W-disjoint orthogonality of two speech signals. Original speech signal (a)* $s_1(t)$ *and (b)* $s_2(t)$*; corresponding STFT spectrogram of the source signal (c)* $|s_1(k,\omega)|$ *and (d)* $|s_2(k,\omega)|$*; (e) product of the two spectrogram* $|s_1(k,\omega)s_2(k,\omega)|$*. The corresponding discrete time step* k *is from* 1 *to* 11*.*

according to equation (4.14). This mask indicator will be used in the following formulation of the WDO measure given a speech mixture.

To evaluate the WDO performance of the different mixtures, Yilmaz et al. [54] proposed a measure to calculate the WDO, which is based on two criteria, the preserved signal ratio (PSR), and the signal-to-interference ratio (SIR). The PSR represents the ability to preserve the source of interest, given as

$$\text{PSR}_i = \frac{\|\Lambda_i(k,\omega)S_i(k,\omega)\|^2}{\|S_i(k,\omega)\|^2}, \tag{4.15}$$

while the SIR denotes the suppression of the interfering sources, defined as

$$\text{SIR}_i = \frac{\|\Lambda_i(k,\omega)S_i(k,\omega)\|^2}{\|\Lambda_i(k,\omega)Y_i(k,\omega)\|^2}. \tag{4.16}$$

The measure of WDO is calculated as [54]

$$\text{WDO}_i = \text{PSR}_i - \frac{\text{PSR}_i}{\text{SIR}_i} \tag{4.17}$$

$$= \frac{\|\Lambda_i(k,\omega)S_i(k,\omega)\|^2 - \|\Lambda_i(k,\omega)Y_i(k,\omega)\|^2}{\|\Lambda_i(k,\omega)S_i(k,\omega)\|^2}. \tag{4.18}$$

For speech signals which are perfectly orthogonal in the TF domain, one can have that PSR $= 1$, and SIR $= \infty$. The maximum WDO can thus be obtained, i.e., $\text{WDO}_i = 1$. A minimum value zero ($\text{WDO}_i = 0$) achieves when the mask kills all the interest source energy or results in equal energy for the source and interferences.

4.3.2 W-disjoint orthogonality in the adverse environment

Fully analysis of the WDO property of the speech mixtures can be found in [54]. The WDO for two speech mixture can be as high as 0.90. However, all the results in [54] are obtained in the anechoic and speech only environment. Clearly, the WDO assumption will be violated in the noisy and reverberant environment. The TF spectrogram for a speech signal is very clear in an anechoic environment, as shown, in Fig. 4.11(a). However, in the reverberant and noisy environment, the spectrogram is blurred and the sparsity is destroyed. Fig. 4.11(b) and (c) shows a speech signal under a reverberant and noisy environment respectively. Due to the adverse environment, the signal energy on the TF plane are smeared. In such a case, the WDO

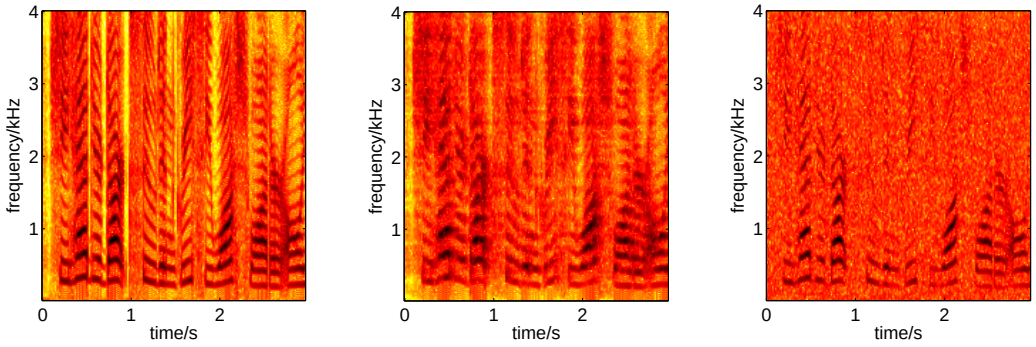
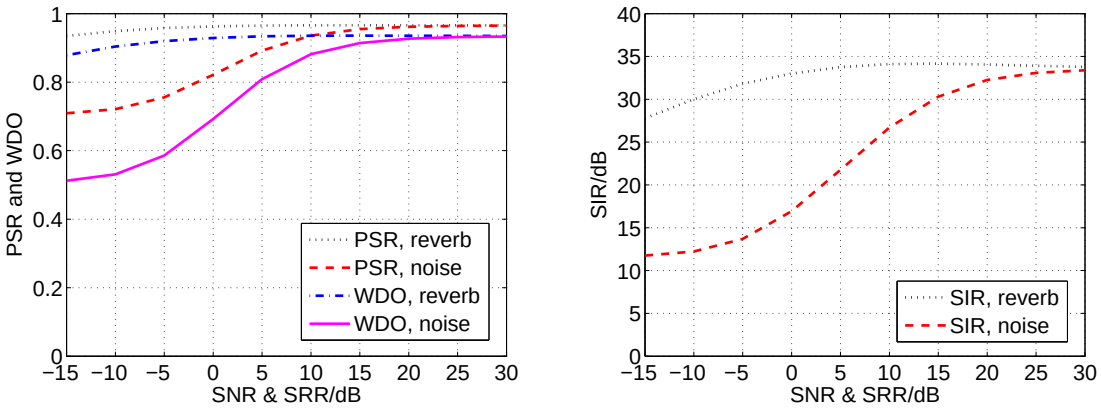


Figure 4.11: *(a) TF spectrogram in the anechoic environment; (b) TF spectrogram in the reverberant environment; and (c) TF spectrogram in the noisy environment. The spectrogram is very clear in the anechoic environment but smeared around by the reverberation and noise. The corresponding discrete time step k is from* 1 *to* 22.



(a) (b)

Figure 4.12: *(a) PSR and WDO for two sources under different noisy and reverberant environments; (b) SIR (in dB) for two sources under different noisy and reverberant environments.*

will be degraded since the reverberant and noisy TF components fill those empty TF bins and interfere with those bins from sources.

Our interest here is to investigate the influence of the noise and the reverberation on the WDO of the speech mixtures. The number of simultaneously active speakers considered here is small (say two or three). This is the case for the realistic room acoustic tracking problem, in which it assumes that the number of simultaneously active speaker is small. The WDO values of mixtures from two sources or three sources are then calculated. 32 speech signals from the TIMIT database are used to generate the mixtures, which lead to $32 \times 31 = 992$ mixtures for two sources. For the mixtures of three sources, 992 mixtures are also randomly picked to evaluate the WDO assumption. The simulated SRRs, the corresponding reflection coefficients ρ, and reverberation times, T_{60}, are given in Table 4.2.

Figure 4.12 shows the WDO, PSR, and SIR of mixtures of two sources under different rever-

SRR(dB)	-15	-10	-5	0	5	10	15	20	25	30
ρ	0.97	0.95	0.91	0.85	0.78	0.68	0.57	0.46	0.36	0.28
$T_{60}(s)$	0.99	0.60	0.38	0.24	0.16	0.12	0.09	0.07	0.06	0.05

Table 4.2: *Relations among reflection coefficients ρ, reverberation time T_{60} and SRR. Room dimension $5 \times 5 \times 3m^3$; microphone position $(0.1\ 2.5\ 1.5)m$; and source position $(2.5\ 1.0\ 1.5)m$.*

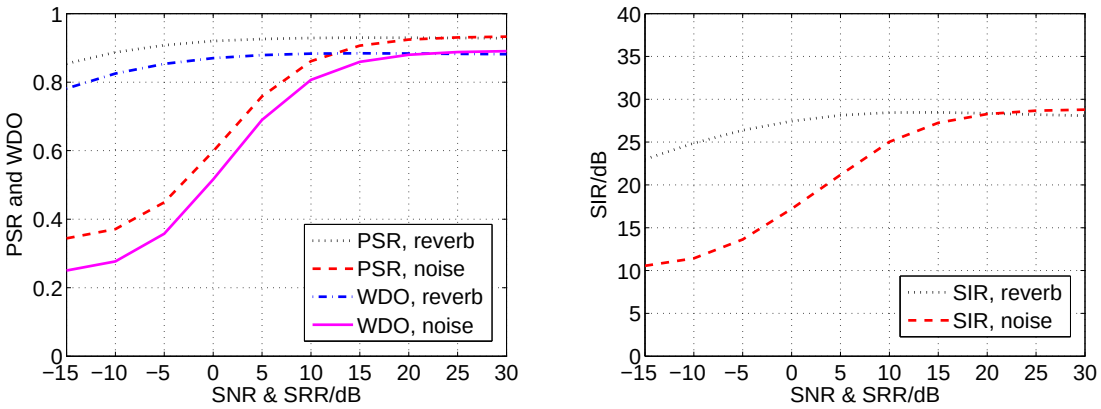


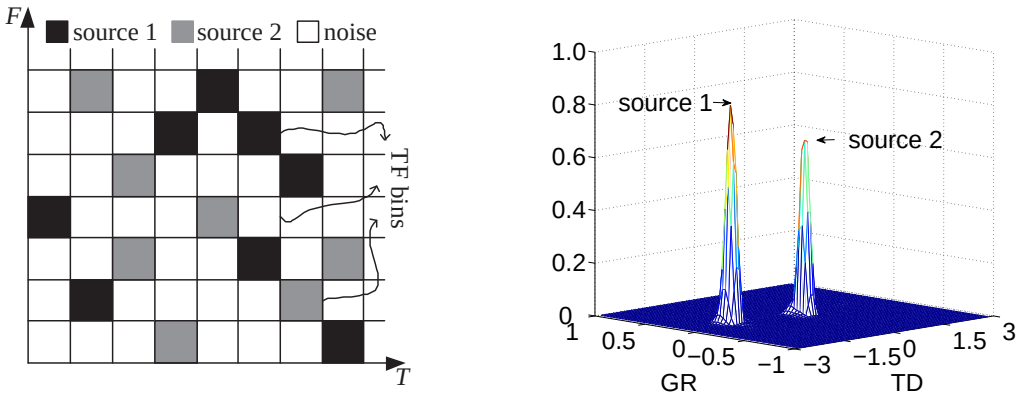
Figure 4.13: *(a) PSR and WDO for three sources under different noisy and reverberant environments; (b) SIR (in dB) for three sources under different noisy and reverberant environments.*

berant and noisy environments. Due to the overlap of the reflection components, the WDO decreases as the increasing of the reverberation. To consider the WDO assumption in the noisy environment, similar experiments are also implemented under different SNR environments. Gaussian white noise is added to the pure speech signal to generate different SNRs. Since the energy of the GWN is uniformly distributed on the TF domain, a number of empty TF bins are filled by the GWN, as shown, in Fig. 4.11(c). The WDO performance thus degrades sharply as the algorithm simply takes all of these noise bins into consideration.

To illustrate the effect of different number of sources, similar experiments are also implemented for three sources. Figure 4.13 gives the WDO, PSR, and SIR of mixtures of three sources under different adverse environments. The signals from multiple sources interfere each other, and decrease the sparsity of the spectrogram of the mixtures. The WDO thus deteriorates with increasing number of simultaneously active sources. Generally, the higher SNR/SRR and the fewer the number of talkers, the better the WDO can be achieved.

4.3.3 TDOA estimation via DUET

As discussed in the previous section, the mixture of multiple speech signals can be regarded as WDO in the TF domain. It would be a great interest to investigate an approach that separate the speech signals in the TF domain first, and the TDOA of each source can then be estimated from the TF bins of the corresponding speech signal. This section introduces a TF spectrogram



(a) (b)

Figure 4.14: *(a) Illustration of disjoint TF spectrogram, each TF bin is either dominated by a single source or noise; (b) 2-D histogram of two sources in the anechoic environment.*

separation based TDOA estimation approach by using DUET.

According to the WDO assumption, the source energy is much higher than the observation noise when the source is active in a TF bin (k, ω), given as

$$|S_m(k,\omega)| \gg |V_{m,i}(k,\omega)|\,. \tag{4.19}$$

Let $a^m_{\ell,i}$ represent the attenuation of the mth source signal at the ith microphone of ℓth microphone pair, i.e., $a^m_{\ell,i}(k) = 1/4\pi r^m_{\ell,i}(k)$. Ignoring the effect of noise and following the STFT in equation (2.14) on page 21, the signal model in the TF domain can thus be simplified as

$$Z_{\ell,i}(k,\omega) = \sum_{m=1}^{M_k} a^m_{\ell,i}(k) e^{-j\omega\tau^m_{\ell,i}(k)} S_m(k,\omega). \tag{4.20}$$

Figure 4.14(a) illustrates the disjoint TF spectrogram of different speech signals. Since all these TF bins are disjoint, each TF bin thus either carries the time-delay and attenuation information for a single source or has no meaningful information as a noise bin. The ratio of a given TF bin across a microphone pair can be defined as

$$R_\ell(k,\omega) = Z_{\ell,1}(k,\omega)/Z_{\ell,2}(k,\omega). \tag{4.21}$$

Given a TF bin at (k, ω), suppose that the mth source is active (the contribution of other sources

on this TF bin is thus ignored). The expression (4.21) can be written as

$$\begin{aligned} R_\ell(k,\omega) &= \frac{a^m_{\ell,1}(k)e^{-j\omega\tau^m_{\ell,1}(k)}S_m(k,\omega)}{a^m_{\ell,2}(k)e^{-j\omega\tau^m_{\ell,2}(k)}S_m(k,\omega)} \\ &= \frac{a^m_{\ell,1}(k)}{a^m_{\ell,2}(k)}e^{-j\omega(\tau^m_{\ell,1}(k)-\tau^m_{\ell,2}(k))} \\ &= \frac{a^m_{\ell,1}(k)}{a^m_{\ell,2}(k)}e^{-j\omega\tau^m_{\ell}(k)}, \end{aligned} \tag{4.22}$$

from which the gain-ratio (GR) and time-delay (TD) estimates for each time-frequency bin can easily be obtained as

$$\begin{aligned} a^\ell_{k,\omega} &= |R_\ell(k,\omega)| = \frac{a^m_{\ell,1}(k)}{a^m_{\ell,2}(k)}, \\ \tau^\ell_{k,\omega} &= -1/\omega \angle R_\ell(k,\omega) = \tau^m_\ell(k), \end{aligned} \tag{4.23}$$

with $|\cdot|$ and $\angle\cdot$ denoting the amplitude and the phase of the estimates respectively, and $a^\ell_{k,\omega}$ and $\tau^\ell_{k,\omega}$ are the GR and TD for the TF bin (k,ω) separately, which are the GR and TD information particularly for source m.

Given a GR resolution parameter A and a TD resolution parameter, D, define an indicator function such that

$$\Lambda^\ell_{A,D}(k,\omega) = \begin{cases} 1, & \text{if } |a^\ell_{k,\omega} - \zeta A| \le A \text{ and } |\tau^\ell_{k,\omega} - \eta D| \le D; \\ 0, & \text{otherwise,} \end{cases}.$$

where ζ and η are any integers which lead $(\zeta A, \eta D)$ to cover a GR and TD range completely. The function Λ indicates whether the GR and TD of a TF bin locate around a given parameter $(\zeta A, \eta D)$. Based on this indicator, a 2-D histogram for different integers ζ and η can be constructed as

$$h_\ell(\zeta A, \eta D) = \sum_{k,\omega} \Lambda^\ell_{A,D}(k,\omega)\left|Z_{\ell,1}(k,\omega)Z_{\ell,2}(k,\omega)\right|^\gamma, \tag{4.24}$$

where $|Z_{\ell,1}(k,\omega)Z_{\ell,2}(k,\omega)|^\gamma$ is a weighting term for some γ. For a source with mixing parameter pairs (ζ,η), a large portion of the source TF bins will be captured by setting a large dimension parameters A and D around its parameter pair. Detailed discussion about the different choices of γ can be found in [54]. Here $\gamma = 0$ is picked to equalise the importance of all the TF bins, by which the TDOA information is emphasised, and the effect of signal energy is

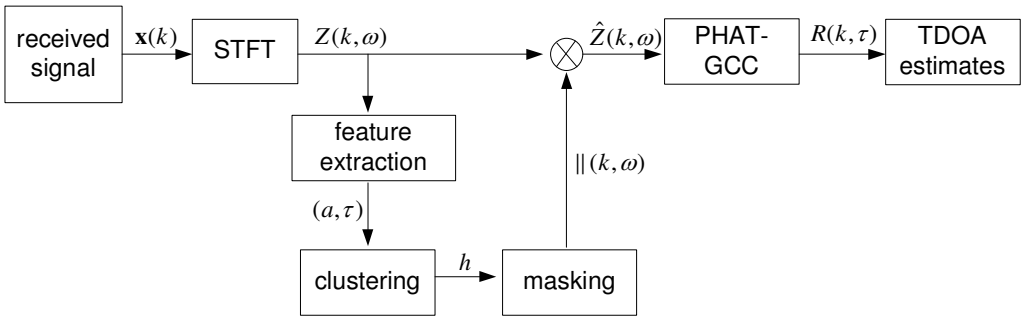


Figure 4.15: *Flow diagram of the DUET-GCC approach. Basically, the speech mixtures are separated by using the DUET in the TF domain, and the PHAT-GCC is then employed for the spectrogram of each source to estimate the TDOAs.*

reduced. Fig. 4.14(b) gives an example of 2D histogram of two source mixtures in the anechoic environment.

Figure 4.15 shows the flow chart of the DUET-GCC approach. At the DUET step, the GR and TD features are extracted by using equation (4.23). Following the calculation of the 2-D histogram $h_\ell(\zeta A, \eta D)$, the mixing parameters can be estimated by selecting the local maxima within the range of interest of pairs. Supposing that n_k^ℓ local maxima can be obtained from a 2-D histogram, the source TF bin indicator function can thus be obtained as

$$\mathbb{I}_n^\ell(k,\omega) = \begin{cases} 1, & \text{if } \left|a_{k,\omega}^\ell - \hat{a}_n^\ell\right| < A \text{ and } \left|\tau_{k,\omega}^\ell - \hat{\tau}_n^\ell\right| < D; \\ 0, & \text{otherwise,} \end{cases} \tag{4.25}$$

where $(\hat{a}_n^\ell, \hat{\tau}_n^\ell)$, for $n = 1, \ldots, n_k^\ell$ is the peak location obtained from the 2-D histogram in equation (4.24). The TDOA for the nth peak in the 2-D histogram, h_ℓ, can thus be estimated as

$$\hat{\tau}_{k,n}^\ell = \mathbb{E}\left(\mathbb{I}_n^\ell(k,\omega)\tau_{k,\omega}^\ell\right), \tag{4.26}$$

where $\mathbb{E}(\cdot)$ denotes the expectation.

The DUET based TDOA measurement extraction is a separation-based approach. Unlike the PHAT-GCC function which is unable to differentiate between all the different source signals, the TDOAs obtained by TF masking approach are estimated from the TF bins of individual source signals. It is thus able to give the TDOAs for multiple sources, even when these sources are simultaneously active. However, in the noise and reverberation environments, the TF spectrogram is smeared and blurred, and the WDO assumption is violated. The TDOA estimation

will thus be degraded due to the decrease of the WDO of the speech mixtures as well as the phase distortion. In particular, the final expectation step (4.26) is very sensitive to the TD and GR parameters (A, D). Such parameters always require an extensive experimental studying and depend on different noisy and reverberant environments. All these factors make the final TDOA estimation diverge from the ground truth TDOA, i.e., $\hat{\tau}_{n,k}^{\ell} \neq \tau_{n,k}^{\ell}$. In this thesis, this expectation based TDOA estimation is thus not used. In the next section, a DUET based GCC method will be developed to obtain robust TDOA estimation.

4.4 Multiple source TDOA estimation via DUET-GCC

In this section, a DUET-GCC method is proposed to estimate TDOAs for multiple simultaneously active sources. The DUET is used to separate the source in the TF domain. The TF spectrogram for each source is then employed to generate the GCC function. The maximum peak in the GCC function thus represents the TDOA for each source individually.

4.4.1 DUET-GCC

The degenerate unmixing estimation technique provides a remarkable way to separate the spectrogram of multiple concurrent speech signals into sets of TF bins, each containing the spectrogram of a single source. Since the PHAT-GCC method is found robust in the noisy and reverberant environments, it is expected to combine PHAT-GCC to estimate the TDOAs based on the TF bin set of each source extracted by DUET.

Suppose that n_k^{ℓ} peaks (i.e., n_k^{ℓ} sources) can be enumerated from a 2-D histogram, and each peak with the indicator $\mathbb{I}_n^{\ell}(k,\omega)$ defined by (4.25). The STFT in equation (2.14) on page 21 can be formed for the received signal from the nth source individually, given by

$$\hat{Z}_{\ell,i}^{n}(k,\omega) = \mathbb{I}_n^{\ell}(k,\omega) Z_{\ell,i}(k,\omega), \tag{4.27}$$

for all $n = 1, \ldots, n_k^{\ell}$, where $\hat{Z}_{\ell,i}^{n}(k,\omega)$ is the TF bin set for each potential source n. Following the GCC equation in (2.28) on page 25, the GCC function for each individual source can thus be written as

$$R_{\ell}^{n}(k,\tau) = \int_{\Omega} \Phi_{\ell}^{n}(k,\omega) \hat{Z}_{\ell,1}^{n}(k,\omega) \hat{Z}_{\ell,2}^{n*}(k,\omega) e^{j\omega\tau} d\omega, \tag{4.28}$$

where

$$\Phi_{\ell}^{n}(k,\omega) = \frac{1}{|\hat{Z}_{\ell,1}^{n}(k,\omega)\hat{Z}_{\ell,2}^{n*}(k,\omega)|}, \tag{4.29}$$

is the PHAT weighting term. The same as in equation (2.27) on page 25 the TDOA for each source can be obtained via exploring an one-dimensional search over the GCC function, given as

$$\hat{\tau}_{k,n}^{\ell} = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_{\ell}^{n}(k,\tau). \tag{4.30}$$

The TDOA estimation for multiple sources is thus achieved. It is worth pointing out that the DUET-GCC estimation steps (4.27) to (4.30) are the same as the GCC estimation procedure introduced in Section 4.2.1, but the spectrogram of each individual source extracted by using DUET is employed to replace the STFT of whole received signal in the traditional PHAT-GCC approach.

In practice, when the frame length is larger than the window length of STFT, several STFT will be operated to obtain the full spectrogram. Assume that J times of STFT are operated at each frame. For example, in this thesis, the frame length for each processing is 1024 samples, and the window length of STFT is set to 512 to obtain a better time resolution. To use the received signal effectively, the STFT is implemented with a half window overlap. This thus leads to $J = 3$. In such a case, the TF spectrogram for each source can be estimated by implementing a spectrogram averaging, given as

$$\hat{Z}_{\ell,i}^{n}(k,\omega) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}_{n}^{\ell}(j,\omega) Z_{\ell,i}(j,\omega). \tag{4.31}$$

Equation (4.27) is actually a special case of (4.31). In (4.27), the block of data collected at a frame are used to generate the STFT spectrogram, i.e., $J = 1$. Equation (4.31) further splits the frame signal by adding a window, and more detailed instantaneous information can be obtained from the STFT.

Since the TDOA estimation of the speech sources are handled separately, the interference between the source signals is naturally decreased. The DUET-GCC method is thus more appropriate for the TDOA estimation of multiple simultaneously active sources than the traditional PHAT-GCC method. Also due to the capability of suppressing the reverberation and noise by PHAT weighting term, the TDOA estimation performance via DUET-GCC approach is better than simply taking the expectation of the TDOA information from all the TF bins in equation

(4.26). Fig. 4.16 shows the GCC function extracted from DUET-GCC method and PHAT-GCC method respectively. Two largest local extremals in DUET 2-D histogram are used to represent the TDOAs, and thus two GCC functions. The DUET-GCC method presents accurate TDOA estimates for two sources. However, the PHAT-GCC is only able to present one source effectively, and fails to produce a sharp peak for the other one.

4.4.2 Phase ambiguity and unwrapping

Due to the periodicity of the complex exponential function, the phase of the complex TF bin will be only taken between $-\pi$ and π. The means that the phase term yields a meaningful TDOA estimate at any TF bin and for any microphone pair, only if

$$|\omega\tau| < \pi \tag{4.32}$$

where ω and τ are the radial frequency and the delay of the corresponding TF bin. The maximum TDOA will be achieved when the source lies on an extended line which links up with the microphone pair. The possible TDOA range is thus

$$|\tau| \leq \frac{d}{c}, \tag{4.33}$$

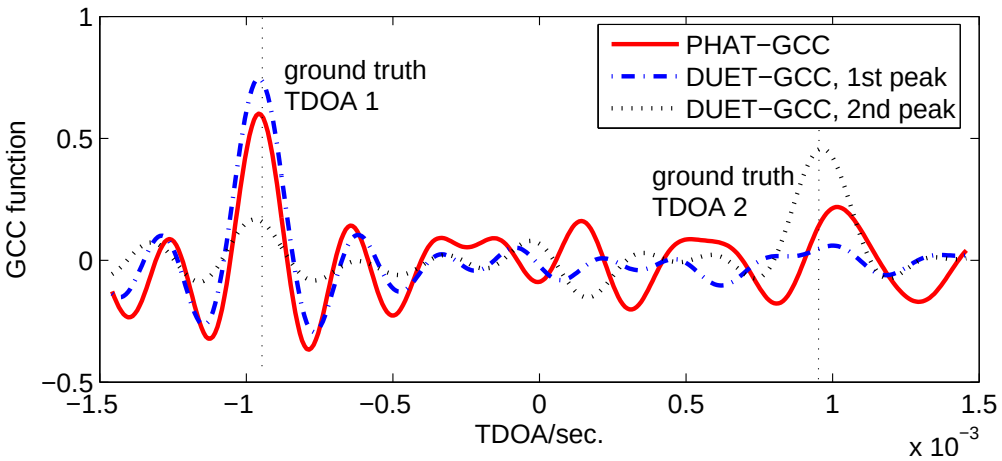


Figure 4.16: *GCC function from DUET approach and traditional PHAT weighting. Two sources are located at (1.4, 1.2)m and (1.4, 2.8)m respectively. The GCC function is estimated from the first microphone pair (microphone 1 and microphone 2), as shown, in Fig. 2.9. The ground truth TDOAs are* $\pm$ *0.95ms.*

with d denoting the microphone separation and c is the sound velocity. The allowable frequency range is

$$|\omega| \leq 2\pi f_{\text{max}}, \tag{4.34}$$

where f_{max} is the maximum frequency which is half of the Nyquist sampling frequency f_{s}, i.e., $f_{\text{max}} = f_{\text{s}}/2$. The microphone separation d by which the phase term will not be wrapped is thus addressed as

$$d < d_{\text{max}} = \frac{c}{2f_{\text{max}}} = \frac{c}{f_{\text{s}}}. \tag{4.35}$$

To obtain the unwrapped phase term, a proper microphone distance should be chosen. For example, as mentioned in Section 2.3.2.2, on page 27, given a sampling frequency of 8kHz, the maximum microphone distance for unwrapped phase term will be roughly 4.2cm. This microphone separation is too small to apply the PHAT-GCC method; due to a limited TDOA resolution, the performance of TDOA estimation will be very low with this microphone separation; even worse than the ROC plot for 10cm microphone separation in Fig. 4.9, on page 88.

In many applications, the microphone separation will be larger than such a small d_{max}, a phase unwrapping method should be adopted to unwrap the phase term at the higher frequency band ($f > c/(2d_{\text{max}})$). In this thesis, the unwrapping approach depicted in Section 2.3.2 on page 24 will be used. The unwrapped TF bins at the lower frequency band ($f < c/(2d_{\text{max}})$) is used to initially histogram the TDOAs $\hat{\tau}_{n,k}^{\ell}$ for multiple sources by using the TF masking method, given by equation (4.26). The estimated TDOAs are then used to predict the phase term at the higher frequency band and unwrap them. Finally, all the TF bins (including those at higher frequency band which previously are wrapped) are employed to form the indicator function (4.25) again.

4.4.3 More practical issues

In practice, ideal 2-D TD and GR histograms can rarely be achieved because of the following two reasons. First is due to the outliers of the GR $a_{k,\omega}^{\ell}$. Although for tracking problem, $a_{k,\omega}^{\ell}$ should be with a reasonable value, some extreme values may present in it. These outliers make the TD feature not be clustered even though the TD components are correct. The second reason is that the TD and GR are with different units. Compared to the GR feature, the TD feature is too small. It is thus very difficult to give a meaningful parameter pair (A, D) to cluster them. The detailed studying of TD and GR features can be found in [112].

One way to solve this problem is normalising these two features [112], and thus make the

	original (a, τ)	normalised $(\bar{a}, \bar{\tau})$
source 1	$(0.91, -0.20 \times 10^{-3})$	$(0.68, 0.20)$
source 2	$(1.00, 0.10 \times 10^{-3})$	$(0.66, -0.44)$

Table 4.3: *The ground truth of GR and TD for original features and normalised features respectively.*

parameter studying of (A, D) controllable. Since the allowable range of the TD feature is $[-\tau_{\max}, \tau_{\max}]$, the TD feature can thus be normalised as

$$\bar{\tau}^{\ell}_{k,\omega} = \frac{\tau^{\ell}_{k,\omega}}{2\tau_{\max}}, \tag{4.36}$$

This normalisation guarantees the TD feature is within the range $-1/2 \leq \bar{\tau}^{\ell}_{k,\omega} \leq 1/2$. The nomalisation of GR feature can be addressed as

$$\bar{a}^{\ell}_{k,\omega} = \frac{|Z_{\ell,i}(k,\omega)|}{\sqrt{\sum_{i=1}^{2} |Z_{\ell,i}(k,\omega)|^2}}. \tag{4.37}$$

This normalised GR feature has the property of $0 \leq \bar{a}^{\ell}_{k,\omega} \leq 1$. According to the WDO assumption, each TF bin will be dominated by at most one source. The minimum value zero is achieved when the source is extremely close to the jth ($j \neq i$) microphone, and maximum value one when the source is extremely close to the other microphone.

Figure 4.17 gives an example of 2-D histogram using the original GR and TD features and normalised features respectively. Table 4.3 gives the ground truth of GR and TD features for the received signal. Although the TD histogram presents two TDOA peaks in (e), the peak cannot be detected in the 2-D histogram (a) and contour plot (f) since the GR feature cannot be clustered. In contrast, using the normalised features is able to solve this problem; two peaks can be found in the 2-D histogram (d) and contour plot (g).

Now both the TD and GR features are normalised with a value range of one. The value of parameters A and D are thus comparable with each other. As defined in section 4.1.3, the TD feature is regarded as a correct estimation of the real TDOA if it is located in the admissible range of anomaly error ϵ. The spacing parameter D is thus picked as a normalised version of anomaly error ϵ, given as

$$D = \frac{\epsilon}{2\tau_{\max}} = \frac{cT_c}{4d_{\text{ref}}}. \tag{4.38}$$

The effect of the microphone separation is reduced. The d_{ref} is always chosen as 1m, and T_c

is a two samples time interval. For a sampling frequency of 8kHz, the normalised TD spacing parameter is thus about 0.02. Since the GR feature is also normalised within a range of one, we can simply set the GR parameter the same as TD parameter, i.e., $A = D = 0.02$.

The microphone separation will still affect the final TDOA estimation performance since it is an important parameter for the DUET step: the larger the microphone separation, the fewer phase terms are unwrapped. Fig. 4.18 presents the ROC curve interpretation of the probabilities of detection and false alarm under different microphone separations. The experiment setup is exactly the same as described in Section 4.2.3. Each curve is constructed by setting different thresholds for the 2-D histogram, from 0.1 to 0.9 with an increment of 0.1, and 0 and 0.99.

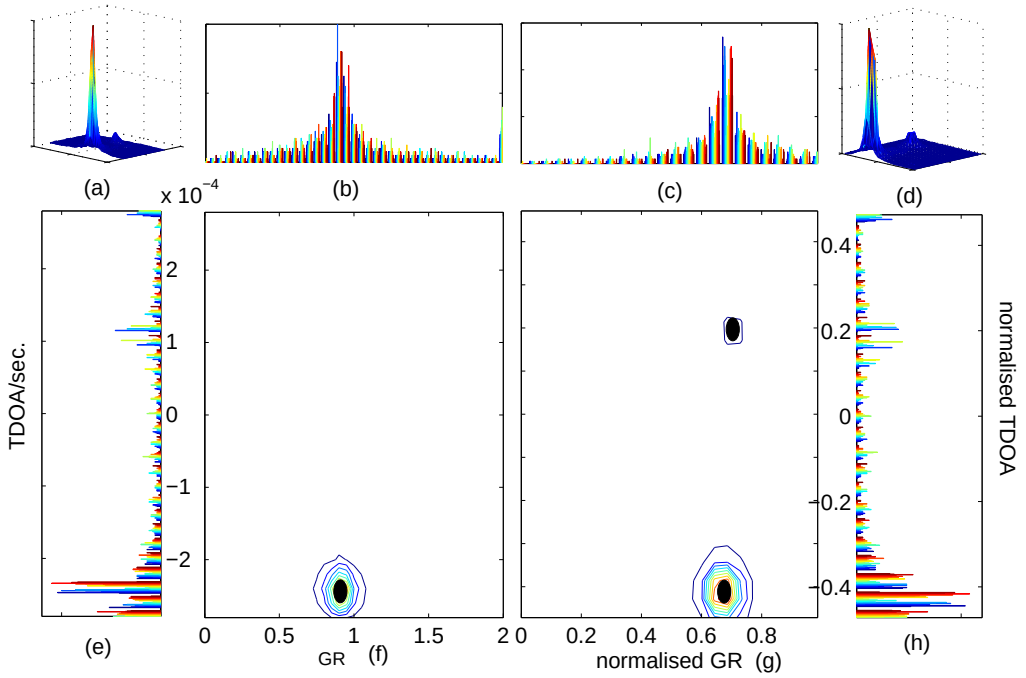


Figure 4.17: *GR and TD histogram with original features and normalised features. (a) 2-D histogram with original features; (b) original GR histogram (c) normalised GR histogram; (d) 2-D histogram with normalised features; (e) original TD histogram; (f) contour presentation of 2-D histogram with original features; (g) contour presentation of 2-D histogram with normalised features; (h) normalised TD histogram. Although the TD histogram presents two TDOA peaks in (e), the peak cannot be detected in the 2-D histogram (a) and contour plot(f) since the GR feature cannot be clustered. In contrast, using the normalised features is able to solve this problem; two peaks can be found in the 2-D histogram (d) and contour plot(g).*

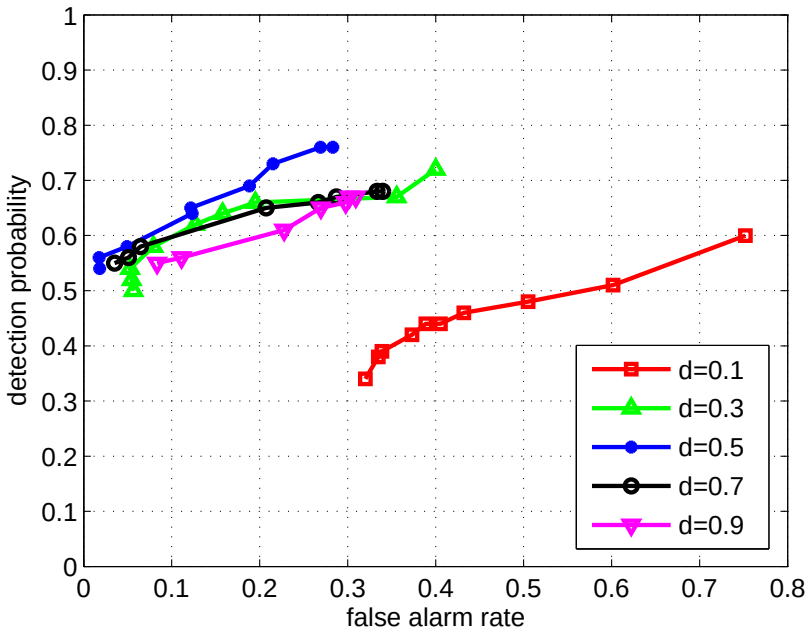


Figure 4.18: *ROC curve interpretation of the probabilities of detection and false alarm under different microphone separations by DUET-GCC method.*

The best TDOA performance can be achieved with a microphone separation of 0.5m. For a separation of 0.3m and 0.7m, very similar results can be obtained.

4.5 Performance in the adverse environment

Other than the microphone separation and a bunch of parameters in the algorithm, the final probabilities of detection and false alarm are affected by the acoustic environment: the noise level SNR and the reverberation level SRR. In this section, the performance of PHAT-GCC and the proposed DUET-GCC method for a static source are first investigated under different SNRs and SRRs. Since the aim is to track dynamical source/sources, the TDOA estimation performance for dynamical sources is also explored.

4.5.1 Static source scenario

In this experiment, the TDOA estimation performance for a stationary source under different SNRs and SRRs is considered. To generate different SRRs, the source is located at different positions in the room together with different wall reflection coefficients. The microphone pair

SRR(dB)	-5	0	5	10	20
$x-y$ position	$(3.5, 3.3)$	$(2.0, 3.0)$	$(1.5, 2.2)$	$(1.0, 1.8)$	$(0.6, 1.7)$

Table 4.4: *Corresponding SRRs generated by different source positions. The wall reflections are set to 0.6.*

is located at $(0.5, 1.75)$ and $(0.5, 2.25)$ respectively, and the simulated room environment is depicted in Fig. 2.9, on page 44. The different SNRs are generated for $[-5, 0, 5, 10, 20]$dB, and the SRR is set to 30dB. For the different simulated reverberant environments, the SRR is also $[-5, 0, 5, 10, 20]$dB, and the background noise level is set to 30dB. Table 4.4 gives the corresponding SRRs generated by different combination of the source positions and wall reflection coefficients, calculated by using equation (4.1), on page 76.

PHAT-GCC performance. The ROC plot of the probabilities of detection and false alarm for PHAT-GCC method under different simulated noisy and reverberant environments is given in Fig. 4.19(b) and Fig. 4.20(b) respectively. Although better probability of detection can be achieved with a threshold value less than 0.5, the false alarm rate is very large under such thresholds and make the following tracking algorithm impossible to filter the source states. A higher threshold value is able to reject the false alarms well (and thus with the lower probability of false alarm), but at a cost of reducing the probability of detection. It is also observed that the thresholds above 0.7 preserve the probability of detection relatively well and reduce the false alarms effectively. It is thus reasonable to set the threshold value within this range.

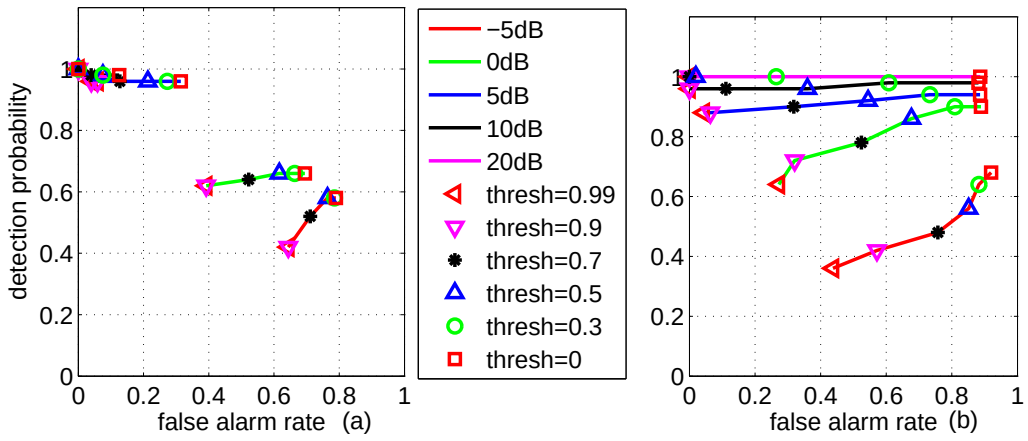


Figure 4.19: *ROC plot of the probabilities of detection and false alarm under different noisy environments (a) for DUET-GCC method; (b) for PHAT-GCC method.*

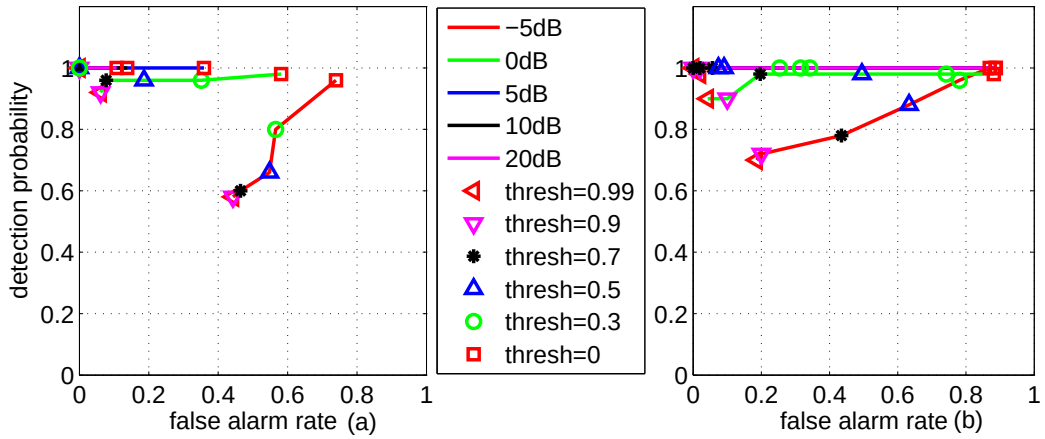


Figure 4.20: *ROC plot of the probabilities of detection and false alarm under different reverberant environments (a) for DUET-GCC method; (b) for PHAT-GCC method.*

DUET-GCC performance. The ROC plot for DUET-GCC method under different noisy and reverberant environments are shown in Fig. 4.19(a) and Fig. 4.20(a) respectively. The affect of the threshold on the ROC plot is very similar as that of the PHAT-GCC method. The DUET-GCC method does better in rejecting the false alarms in the anechoic environment and moderate noisy and reverberant environments. This can be seen from the two figures: when the SRR is larger than 5dB (includes 5dB), the probability of detection is near 1 and the false alarm rate can be limited less than 0.4 with almost all the thresholds. Generally, for the single source scenario, the threshold value above 0.5 is able to present a good probability of detection with limited false alarms.

Low SNR and SRR environments. In the low SNR and SRR environments (e.g., 0dB and -5dB), the TDOA estimation of DUET-PHAT is worse than that of the PHAT-GCC method. This is mainly because the DUET step is sensitive to the noise and reverberations. In particular, under a very low SNR or SRR (e.g., -5 dB) environment, it will be extremely difficult to report the detections and the measurements are always with a high false alarm rate. Both the methods are likely collapsed and thus the probability detection is very low under a reasonable threshold value. This phenomenon is the same as that depicted in [38] that the anomaly percentage increases abruptly when a certain reverberation reaches (as shown in [38], $T_{60} \geq 0.5$s). The main reason is that the free filed model is seriously violated by the noise and the reverberation. The reflections can no longer be depicted by an uncorrelated noise term but they are more like one or several correlated sources which emit signals from the image positions. A further illustration of both methods failing in the low SRR environment is given in Appendix C. In

heavy reverberant environments, more sophisticated TDOA extraction approach, or more data at each time step should be employed to enhance the probability of detection.

4.5.2 Single dynamical source scenario

As shown in Fig. 4.1(b), a dynamical source moving in the room can employ a large range of SRRs even in the same room environment. This section investigates the probabilities of detection and false alarm of PHAT-GCC and DUET-GCC methods when the source is dynamical. The experiment setup is shown in section 2.6.2, on page 43. The speech signal with 50 frames are used to generate a diagonal line source trajectory. Different reflection coefficients are used to generate various reverberant environments. The corresponding reverberation time T_{60} can be found in Table 2.1, on page 45.

The ROC plot of the probabilities of detection and false alarm for DUET-GCC and PHAT-GCC methods are shown in Fig. 4.21(a) and Fig. 4.21(b) respectively. The probability of detection falls down sharply with increasing wall reflection coefficients, and the false alarm rate increases significantly at the same time. In the less reflective environment, all the threshold values can present a good probability of detection. However, the false alarm is very heavy with a low threshold for the PHAT-GCC method, and thus makes the low threshold values (e.g., 0.1 or 0.3) not appropriate for the TDOA extraction. For the PHAT-GCC method, the thresholds larger than 0.7 is able to excludes the false alarm best and preserves the probability of detection well.

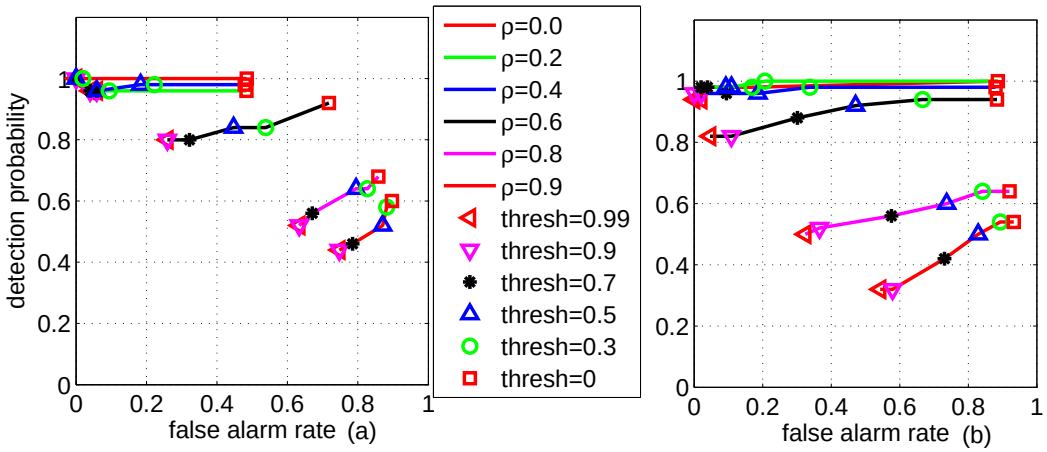


Figure 4.21: *ROC plot of the probabilities of detection and false alarm of a single dynamic source under different reverberant environments (a) for DUET-GCC method; (b) for PHAT-GCC method.*

For the DUET-GCC method, all the thresholds can give a satisfactory probabilities of detection and false alarm. This is because the DUET-GCC method is able to cluster the source signal and estimate the TDOA of the source specifically. When the reflection is strong, both of the methods fail to report the TDOA estimation effectively.

4.5.3 Multiple dynamical source scenario

In this section, the TDOA performance of two simultaneously active sources is investigated. The experiment setup is the same as the previous experiments in this chapter, except that two diagonal line trajectories are organised, each with 50 frames. Similar to the single dynamical source scenario, different reflection coefficients are used to generate various reverberant environments. The PHAT-GCC method is originally employed for extracting the TDOAs of a single source. However, it can be used to estimate the TDOAs for multiple sources by picking multiple peaks in practice. In contrast, the DUET-GCC method is proposed for the TDOA extraction of multiple simultaneously active sources.

Figure 4.22(a) and Fig. 4.22(b) show the ROC plot of the probabilities of detection and false alarm for DUET-GCC and PHAT-GCC methods respectively. Due to the interference between the sources, the probability of detection for the two sources which are simultaneously active is generally lower than that in the single source scenario. In the less reverberant environment, the DUET-GCC is able to extract the TDOAs for the individual sources and thus presents better

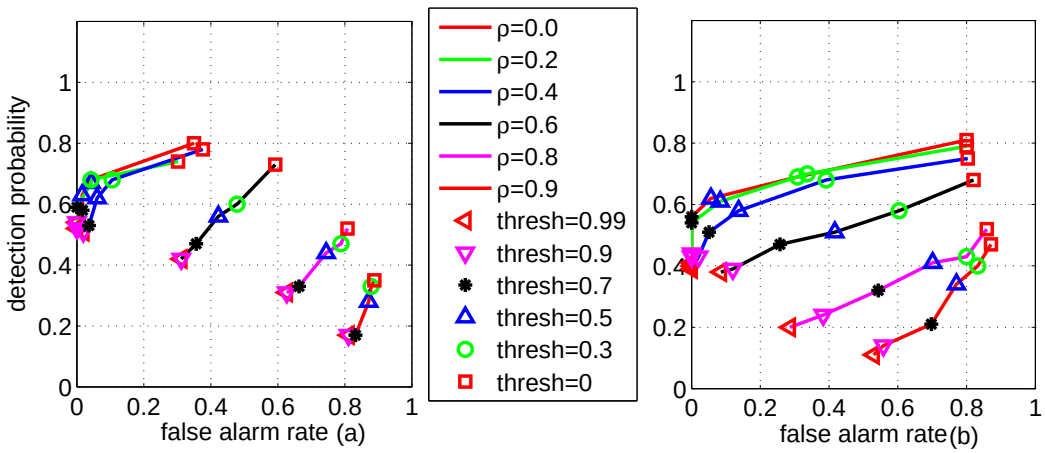


Figure 4.22: *ROC plot of the probabilities of detection and false alarm of two simultaneously active sources under different reverberant environments (a) for DUET-GCC method; (b) for PHAT-GCC method.*

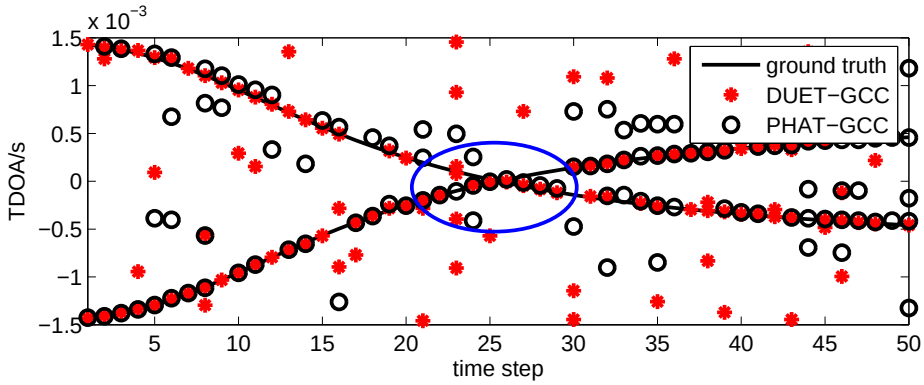


Figure 4.23: *Both the DUET-GCC and PHAT-GCC methods fail to extract TDOAs at the cross area which is marked with an ellipse in the figure. The signals received from microphone pair 1 in the simulated room environment is used to generate the TDOAs.*

ROC plots. However, it also has a resolution problem in clustering the 2-D histogram. When the two sources are very closely spaced (e.g., at the cross area as shown in Fig. 4.23), the TDOA estimates are emerged and can not be detected by using both methods.

The TDOA extraction for multiple sources in the strong reverberant environment is an extremely difficult problem in that no matter what a threshold is chosen, either a large detection missing or a large false alarm rate will appear. In the multiple source tracking scenario, the optimum threshold can only be achieved by taking the source number, and noise and reverberation level into account. However, this information is unknown *a priori*. Generally, to make the balance between the probabilities of detection and false alarm, and also between different number of sources (one or two), a threshold above 0.7 for PHAT-GCC method and 0.5 for DUET-GCC method will provide satisfactory TDOA measurements. Particularly, the probabilities of detection and false alarm of DUET-GCC method are almost the same for all the thresholds above 0.5. The final choice of the threshold is also determined by the tracking algorithm, and depends on how much can the algorithm tolerate the detection missing/false alarms. Since the

method	nonconcurrent source	multiple concurrent source
DUET-GCC	-	0.7
PHAT-GCC	0.7	0.9

Table 4.5: *Threshold choices for DUET-GCC method and PHAT-GCC method under nonconcurrent multiple source tracking and time-varying number of multiple source tracking scenarios.*

nonconcurrent tracking approach in Chapter 5 is actually dealing with one source at each time step, the threshold of 0.7 will be chosen to enhance the probability of detection while with a reasonable false alarm rate. For the tracking approach developed in Chapter 6, a threshold of 0.9 for PHAT-GCC method and 0.7 for DUET-GCC is exploited to exclude the false alarm efficiently to avoid an exhaustive data associations. The values of threshold for different tracking scenarios are summarised in Table 4.5.

4.6 Experiments with real room recordings

To investigate the measurement performance in the real recording environment, the DUET-GCC and PHAT-GCC are also implemented in the audio lab described in Fig. 2.11, on page 46. In this section, the real room recording environment is studied first. The TDOA measurement in the real room environment is then illustrated.

4.6.1 Room environment study

4.6.1.1 Reverberation level

The room reverberation time T_{60} is measured as the time needed for the sound pressure level decaying by 60 dB of its original level after it has been switched off. To measure the T_{60} of the audio lab, a white Gaussian noise placed 1m before a microphone receiver is used as the source signal. Fig. 4.24 shows the received signal. The signal decays fast after the source signal switched off. The reference energy power is calculated as the average energy of a short period

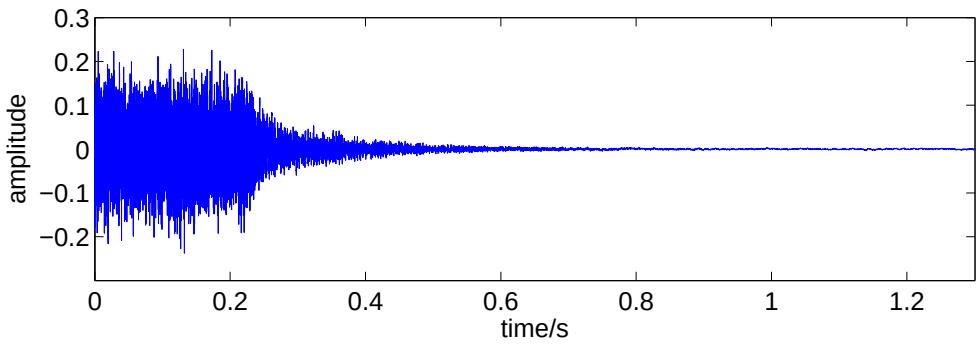


Figure 4.24: *Received white Gaussian noise.*

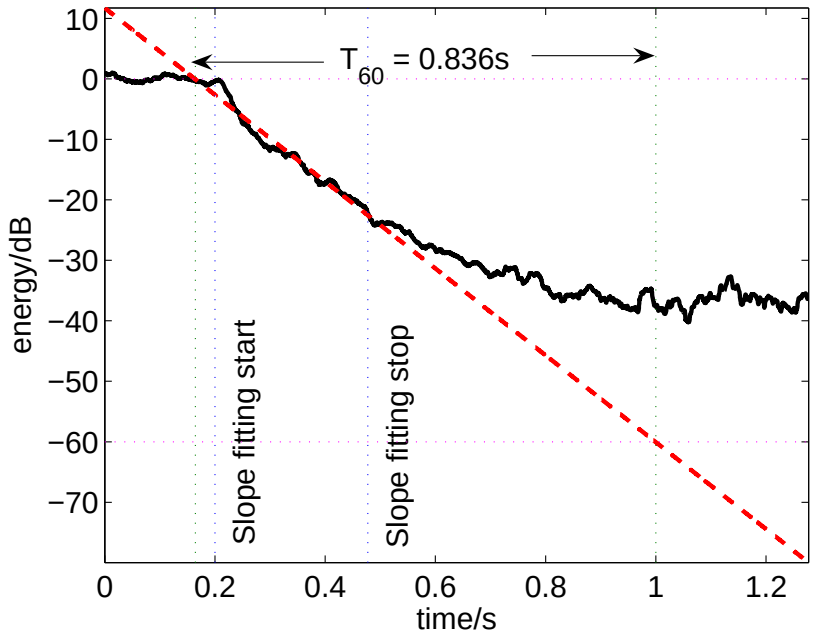


Figure 4.25: *Reverberation time T_{60} calculation of the real audio recording environment.*

T_0 before the received signal $z(t)$ decayed, given as

$$P_0 = \frac{1}{T_0}\int_0^{T_0} z^2(t)dt. \tag{4.39}$$

The signal power at any interval $t - T$ and t can be calculated as

$$P_t = \frac{1}{T}\int_{t-T}^{t} z^2(t')dt'. \tag{4.40}$$

The energy decay in dB is thus defined as

$$D(t) = 10\log\frac{P_t}{P_0}. \tag{4.41}$$

T_{60} is then the time needed for $D(t)$ decaying from 0dB to -60dB.

In the calculation of T_{60}, the reference period T_0 is chosen as 1024 data points before the decay happens. The window length used to calculate the decayed signal is 1024 data points and with a 10 points lag for each calculation. Fig. 4.25 shows the measurement of the T_{60} in the real recording environment. Since the received signal only decays to -40dB after it is switched off, we have to obtain the slope of the decay line and employ the extrapolation to get the exact T_{60}.

The energy decay line between 0.2s and 0.48s is used to fit the line slope. The T_{60} obtained by the final extrapolation is 0.836s, as shown, in Fig. 4.25.

4.6.1.2 Noise level

The noise in the room is usually generated by the electronic equipments (e.g., computer fan and recording systems), and the footstep of the source. To calculate the noise level, a signal with length of T samples is recorded. The source is silent but moving around in the room during the recording. The noise power is

$$P = \frac{1}{T}\int_0^T z^2(t')dt', \tag{4.42}$$

where T is number of samples for a discrete system. The final noise level is calculated as

$$N_{\text{dB}} = 10\log P. \tag{4.43}$$

In this experiment, 5 second unvoiced signal is recorded, which leads to the length of signal $T = 5 \times 8000$ under the sampling frequency of 8kHz. The estimated noise level in the room is about -40dB.

4.6.2 Measurements from real room recordings

The TDOA estimations from real room recording signals are presented in this section. A sketch of the experiment setup is presented in Fig. 2.11 on page 46. A picture for the real room environment and experiment setup is also presented in Fig. A.2 on page 194. Fig. 4.26 presents the recorded signal at microphone 5 and microphone 15 respectively. The motion of the source follows the trajectory 2 marked in Fig. 2.11. The effect of the movement on the received signal energy can be found in the figure: as the source moves away from microphone 5, the amplitude of the received acoustic signal becomes smaller, and vice versa for microphone 15.

Due to a lack of motion capture system, it is difficult to tell the position of the source at a particular time step in the room exactly. The ground truth TDOA is thus unknown. However, the ground truth at each time step can be roughly estimated by assuming that the movement of the source is at an even speed and follows the marked trajectory.

Fig. 4.27 gives the TDOA extraction for a time-varying number of acoustic sources with a threshold of 0.5, 0.7 and 0.9 respectively. The signal for multiple simultaneously active sources

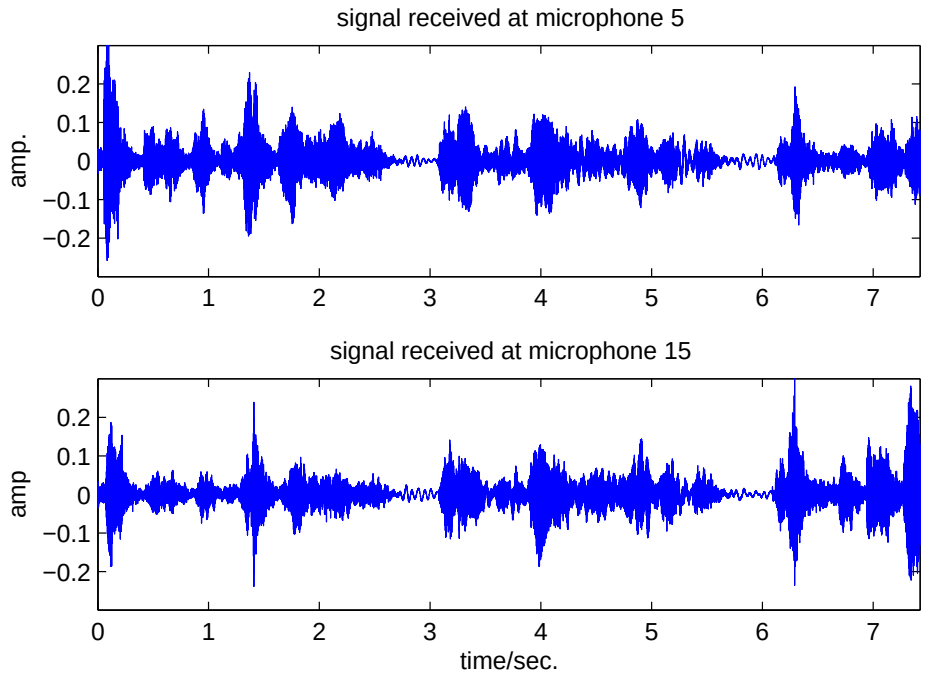


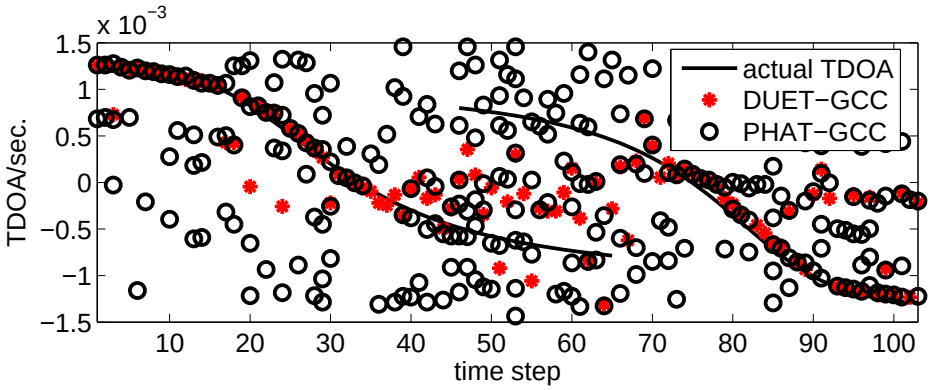
Figure 4.26: *Real recorded signals from microphone 5 and microphone 15 respectively. The motion of the source follows the trajectory 2 marked in Fig. 2.11. The source moves away from the microphone 5, and gets closer to the microphone 15.*

is generated by overlapping the recorded signal from two single sources via post-processing. One source is active from time step 1 to time step 65, and the other from time step 45 to time step 103. Since the ground truth cannot be obtained, the probabilities of detection and false alarm are unknown. As shown in the figure, the threshold of 0.9 for the PHAT-GCC method presents the best TDOA estimates. For DUET-GCC approach, all the TDOA estimates with thresholds are almost the same.

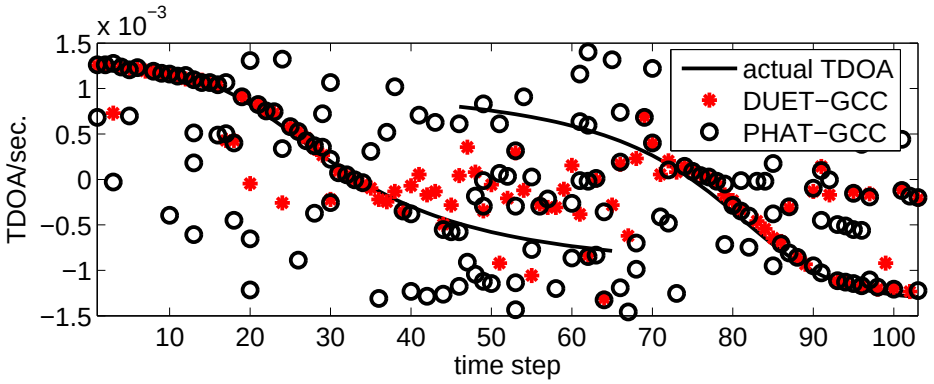
4.7 Chapter summary

In this chapter, PHAT-GCC method is introduced to extract TDOA measurements for multiple speech signals received from microphone pairs. Particularly, based on the WDO assumption of speech mixtures, the DUET-GCC method is proposed to extract TDOA measurements for multiple simultaneously active sources. Further, based on the definition of the probabilities of detection and false alarms, a number of parameters are examined.

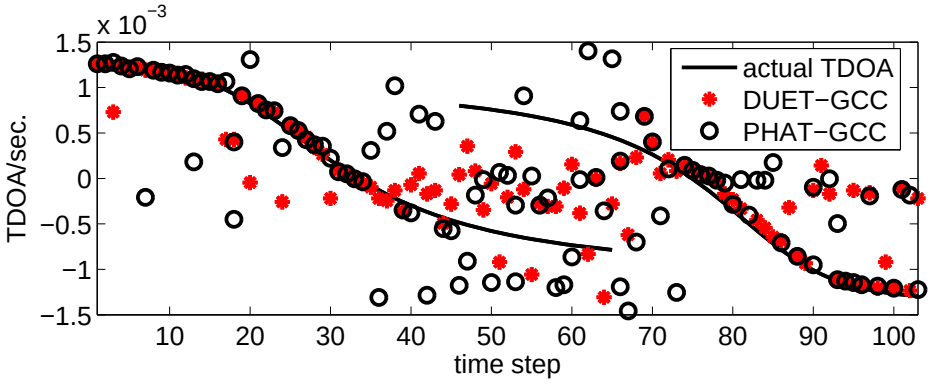
The preliminary investigations show that the microphone separation has a significant effect



(a)



(b)



(c)

Figure 4.27: *TDOA measurement extracted from the real recorded signals. The threshold is set to (a) threshold value 0.5; (b) threshold value 0.7; (c) threshold value 0.9 for both the DUET-GCC and PHAT-GCC approaches. For DUET-GCC approach, the threshold larger 0.5 presents almost the same TDOA estimates; for PHAT-GCC approach, a threshold of 0.9 is able to present best TDOA estimates since large false alarm can be excluded.*

on the TDOA measurement performance. For both the PHAT-GCC method and DUET-GCC method, the best TDOA measurement performance can be achieved when set the microphone

separation around 0.5m. A small separation (e.g., 0.1m) will reduce the resolution of GCC function in PHAT-GCC method and 2-D histogram in DUET-GCC method, and make the TDOA extraction very difficult. On the other hand, an overly large microphone separation will destroy the correlation property between two received signals, and thus decrease the TDOA performance.

Further experiments on different thresholds illustrate that an appropriate choice of the threshold is very important in determining the probabilities of detection and false alarm. Although the final choice of the threshold is dependent on the acoustic environments and experimental studies, our extensive experiments shows that a satisfied TDOA performance can be achieved by setting the thresholds with values around 0.7 and 0.9 for DUET-GCC and PHAT-GCC methods respectively.

Finally, the performance of PHAT-GCC and DUET-GCC approaches is also investigated under different noisy and reverberant environments. Generally, DUET-GCC performs better than PHAT-GCC method when extracting the TDOA measurements for multiple simultaneously active sources under the anechoic and moderate noisy and reverberant environments. However, in the heavy noisy and reverberant environments, the WDO assumption is significantly deteriorated. The performance of DUET-GCC is thus degraded sharply.

Since the PHAT-GCC approach is simple and robust in extracting the TDOA measurements for a single source, it will be used for measurement extraction in nonconcurrent multiple acoustic source tracking in Chapter 5. The DUET-GCC method has its advantage of extracting the TDOA measurements for multiple simultaneously active sources. It will be used as a measurement extraction approach in Chapter 6 for tracking a time-varying number of acoustic sources.

Chapter 5
Nonconcurrent multiple acoustic source tracking

In Chapter 2 and 3, the acoustic source tracking problem and Bayesian source tracking approaches have been fully illustrated and reviewed. The PHAT-GCC based TDOA measurement extraction method is extensively studied in the simulated room environment as well as real room recordings in Chapter 4. Based on all these basic materials, the EKF and PF tracking approaches are fully investigated in this chapter. We further note that, in practice, multiple speakers may have following presence in the room environment: one speaker is active for a period, and then another follows. This special case requires the algorithm to capture the sharp change in position and lock on the new speaker quickly. It is shown in this chapter that the general sequential importance resampling particle filtering (SIR-PF) approach fails to do so since the importance function only uses the prior information and the particles are not drawn effectively. Our contribution in this Chapter is developing an approach that combines the EKF and PF, namely extended Kalman particle filtering (EKPF) to track the nonconcurrent multiple sources. The core idea is that by employing an EKF, the optimum importance function can be approximated, and the particles are thus sampled in a more relevant area based on this importance function rather than a prior density function used in SIR-PF.

5.1 Introduction

The general single source tracking problem is formulated in Section 2.4.1, on page 34. Due to a nonlinear relationship between the TDOA measurements and the source position, the measurement function is usually approximated by a first-order linearisation and the EKF is widely employed [8,9]. The EKF usually takes one TDOA measurement (corresponding to the highest peak in the PHAT-GCC function) from each microphone pair as the detection of the source. If the source is within an anechoic and noise free environment, satisfied performance can be obtained using the EKF for the TDOA based tracking problem since high probability of detection can be achieved and the false alarms are trivial. However, the studies in Chapter 4 on the

TDOA measurements have shown that in the room environments, heavy false alarms are always presented and high probability of detection can rarely be obtained. Therefore, it is desired to develop an approach which is able to model the TDOA measurement generated by the source as well as by the clutter separately.

The sequential importance resampling particle filtering (SIR-PF) is such an approach that allows a bimodal measurement likelihood, and is able to build a likelihood for the combination of the source and clutter generated measurements. Particle filtering has a specific advantage that it can be directly applied to a nonlinear measurement function, by which the Taylor expansion used in EKF can be avoided. The only drawback when employing it for the acoustic source tracking problem is that the optimal importance function can rarely be obtained. The general SIR-PF developed in [6, 12] simply uses a prior distribution as the importance function. Since the prior importance function does not take the current measurements into account, it will cause a tracking lag or even a tracking loss in following a sharp change of the position. The details of this tracking approach will be introduced in Section 5.4. Table 5.1 gives a summary of pros and cons when using EKF and SIR-PF for TDOA based on acoustic source tracking.

	pros	cons
EKF	better state update model	single measurement
SIR-PF	multiple TDOA measurements	poor model for state update

Table 5.1: *Pros and cons of EKF and SIR-PF for TDOA based on acoustic source tracking.*

The SIR-PF is able to track a single source with a slow-paced movement (usually with a speed less than 1m/s) in the room environment well. Generally, due to the incorporation of a bimodal measurement likelihood and a large sample space, the SIR-PF is more robust than the traditional EKF approaches in the noisy and reverberant environments. In this Chapter, we aim to develop a tracking system that is able to track nonconcurrent multiple sources. In such scenario, the source positions may switch at some time steps. It requires the tracking approach robust to the room environment as well as sensitive to the sharp changing of the source position. In our work, the EKF and PF are combined to construct an EKPF [113] to handle this problem. The EKF is employed to filter the samples coarsely according to the current measurement. During each iteration, samples are thus relocated according to both the knowledge of the former state estimates and the current measurements. The particle filtering is then used to further resample these particles and refine the estimates.

Since multiple TDOA measurements are collected across each microphone pair, the EKF cannot be applied directly. Two methods are developed to employ the EKF in this chapter. The first only uses the TDOA from the highest peak in PHAT-GCC function as the measurement at each microphone pair. This is reasonable since the TDOAs from the highest peaks are, in most cases, more reliable than those from the rest of the peaks. The other method is to take all the TDOA measurements into account, but a parameter is used in the innovation updating process to model the effect of false alarms. After the EKF step, the particles are drawn at the more relevant areas than using a prior importance function in general SIR-PF which only takes the past state estimates and measurements into account. The advantage of using an EKPF to track multiple nonconcurrent acoustic sources will be assessed through the simulated room environment experiments as well as real audio lab experiments.

In the following sections, the details of TDOA based tracking framework are firstly presented, including the two-step tracking system. The Langevin dynamical model and the reverberation measurement model are discussed in Section 5.2. The formulation of EKF and PF tracking approaches are introduced in Section 5.3 and 5.4 respectively. The EKPF for nonconcurrent multiple source tracking is subsequently developed in Section 5.5. Experiments and performance evaluation will be presented in Section 5.6 and the conclusions will be drawn in the final section.

5.2 Tracking framework

In Section 2.3, a broad range of tracking systems in terms of different measurement categories are introduced. Although TDOA measurement based tracking system is an indirect approach, it is still widely used because of its simplicity and being easily available in a number of scenarios. In this section, this tracking scheme is briefly introduced. The source dynamics and, in particular, the reverberant measurement model are also presented.

5.2.1 Acoustic source tracking system

In our tracking system, the two-step tracking scheme is employed, in which the TDOA measurements are extracted first and the locator is then applied. Fig. 5.1 shows this tracking scheme: PHAT-GCC method is used to extract the TDOA measurements, and the source position is estimated by the following tracker. It is worth mentioning that the DUET-GCC approach proposed

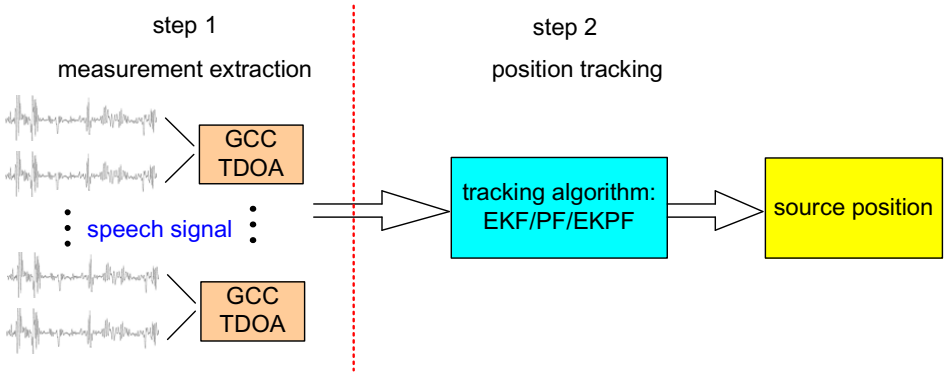


Figure 5.1: *Two-step (indirect) tracking scheme. The TDOA measurements are extracted from the received speech signals first, and the tracking algorithms are then applied to estimate the source position.*

in Section 4.4 on page 98 will not be considered here in extracting the TDOA measurements since at each time step, there is only one source active. PHAT-GCC method is found efficient and robust in the single source scenario.

Suppose that the tracking system consists of L spatially distributed (omni-directional) microphone pairs. The complete TDOA measurement vector for the EKF is constructed by the highest peak of PHAT-GCC function from each microphone pair, written as

$$\mathbf{z}_k = (\hat{\tau}_k^1, \ldots, \hat{\tau}_k^L). \tag{5.1}$$

Given a single source $\mathbf{x}_k$ with unknown position (x_k, y_k), it follows a nonlinear relationship with the TDOA measurement

$$\tau_k^\ell(\mathbf{x}_k) = \frac{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}{c}, \tag{5.2}$$

with c representing the sound velocity. $\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|$ is the difference of the distance between the acoustic source and the two microphones. Fig. 5.2 illustrates the position relationship between the source and the microphone receivers. The task of the tracking algorithm is to estimate the source position $\mathbf{x}_k$ at each time step according to a set of TDOA estimates $\mathbf{z}_k$, under the practical condition that it is assumed $\hat{\tau}_k^\ell \neq \tau_k^\ell(\mathbf{x}_k)$.

If the channel noise at each microphone pair is assumed to be an independent additive white Gaussian noise (WGN) with zero mean and equal variance, the maximum likelihood (ML)

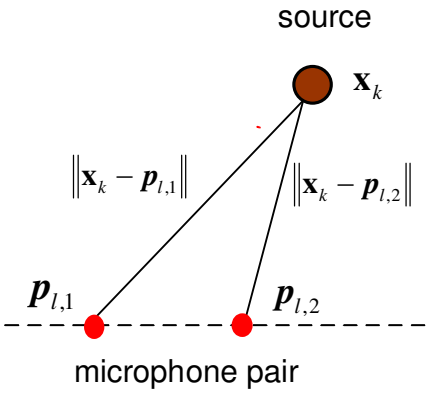


Figure 5.2: *Illustration of the geometry relationship between the source position and the microphone positions. The TDOA is the time difference of the acoustic arriving at these two microphones.*

criterion [27] for the location can be found by minimizing the least square error, given as

$$\hat{\mathbf{x}}_k = \underset{\mathbf{x}_k}{\operatorname{argmin}} \sum_{\ell=1}^{L} \left(\hat{\tau}_k^{\ell} - \tau_k^{\ell}(\mathbf{x}_k) \right)^2 . \tag{5.3}$$

The evaluation of $\mathbf{x}_k$ at each time step involves the optimisation of a non-linear function and necessitates the utilisation of the linear intersection methods, since no exact closed-form solution exists to equation (5.3).

5.2.2 State dynamic model

For the source tracking problem, especially when tracking a source moving with a complicated trajectory, source dynamical models play an important role in improving the tracking performance. This source motion can be modelled by the combination of the position, velocity, acceleration or even direction components [74]. Fortunately, the movement of the speakers in the room environment can always be assumed to be slow-paced (as mentioned in Section 5.1, usually less than 1m/s), and the Langevin motion model [6, 12] is found sufficient to model the source motion. The original source position vector is extended by appending a velocity component, given as

$$\mathbf{x}_k = (x_k, y_k, \dot{x_k}, \dot{y_k}), \tag{5.4}$$

with $(\dot{x}_k, \dot{y}_k)$ denoting the source moving velocity along the corresponding coordinate. The Langevin motion model can be written as

$$\mathbf{x}_k = \mathbf{A}_k \mathbf{x}_{k-1} + \mathbf{Q}_k \mathbf{v}_k, \tag{5.5}$$

where $\mathbf{v}_k$ is a normally distributed random vector, and the matrices $\mathbf{A}_k$ and $\mathbf{Q}_k$ are given by

$$\mathbf{A}_k = \begin{bmatrix} 1 & 0 & a\Delta T & 0 \\ 0 & 1 & 0 & a\Delta T \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix}, \tag{5.6}$$

and

$$\mathbf{Q}_k = \begin{bmatrix} b^2\Delta T & 0 & 0 & 0 \\ 0 & b^2\Delta T & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & b \end{bmatrix}, \tag{5.7}$$

with $\Delta T = T/f_s$ representing the time interval (in second) between time step k and $k-1$, and a and b are the position and velocity variance constants calculated according

$$a = \exp(-\beta \Delta T); \tag{5.8}$$

$$b = v\sqrt{1-a^2}, \tag{5.9}$$

in which v and β are the velocity parameter and the rate constant respectively. Table 5.2 gives several choices of variance constants a and b under different parameter pairs (v, β). In the following tracking approaches, equation (5.5) will be used to model the source motion dynamics. The model parameters used in [1, 6, 12] are $v = 1\text{ms}^{-1}$ and $\beta = 10\text{s}^{-1}$ respectively. In this thesis, these parameter values will also be used.

(v, β)	$(0.5, 5)$	$(0.8, 7)$	$(1, 10)$	$(1, 12)$	$(1.5, 20)$
(a, b)	$(0.73, 0.34)$	$(0.64, 0.62)$	$(0.53, 0.85)$	$(0.53, 1.23)$	$(0.28, 1.44)$

Table 5.2: *The variance constants a and b under different parameter pair (v, β).*

5.2.3 Reverberant measurement model

Due to the reverberation and background noise, the ghost peaks which correspond to error TDOAs may present in the GCC function. A number of peaks are thus collected to include the TDOAs generated by the real source as inclusive as possible. Assume that n_k^ℓ number of TDOA estimates can be obtained from the ℓth microphone pair at time step k, written in a vector as

$$\mathbf{z}_k^\ell = \{\hat{\tau}_{1,k}^\ell, \ldots, \hat{\tau}_{n_k^\ell,k}^\ell\}. \tag{5.10}$$

For the tracking system consisting of L microphone pairs, the complete measurement set is then

$$\mathcal{Z}_k = \{\mathbf{z}_k^1, \ldots, \mathbf{z}_k^L\}. \tag{5.11}$$

The main difference between this reverberant measurement model (5.11) and the traditional measurement model (5.1) is that the former allows the presence of false TDOA measurements, and the later only picks the TDOA from the largest peak at each microphone pair.

How many peaks are collected depends on the noise and reverberation level. Usually, the heavier the noise and reverberation, the more peaks are needed to be included to fully represent the source generated TDOAs. However, the number of false alarms will increase accordingly. In [6], the three largest peaks are picked as the TDOA measurements. However, in the low reverberant environment, it is unnecessary to collect so many peaks and if the reverberation is strong, three peaks may be not enough to include the real TDOA. In particular, when multiple sources exist, how many peaks are appropriate to present all the TDOAs generated by the sources is never known. According to the investigation in Chapter 4, a threshold with value of 0.7 on the GCC function is appropriate to generate a good probability of detection with limited false alarms. All the peaks exceed this threshold are collected to extract the TDOA measurements. Further, if there is no such a peak above this threshold, the largest peak is picked. Fig. 5.3 shows the TDOAs collected from microphone pair 1 under the wall reflection coefficients $\rho = 0.4$ and $\rho = 0.8$ respectively. The TDOAs are seriously deterioted when the reflection coefficients increase to 0.8. The probability of detection can be enhanced by picking more peaks, and so as the tracking performance. However, for the reverberation with wall reflection coefficients $\rho = 0.4$, it is unnecessary to pick several peaks to represent the TDOAs.

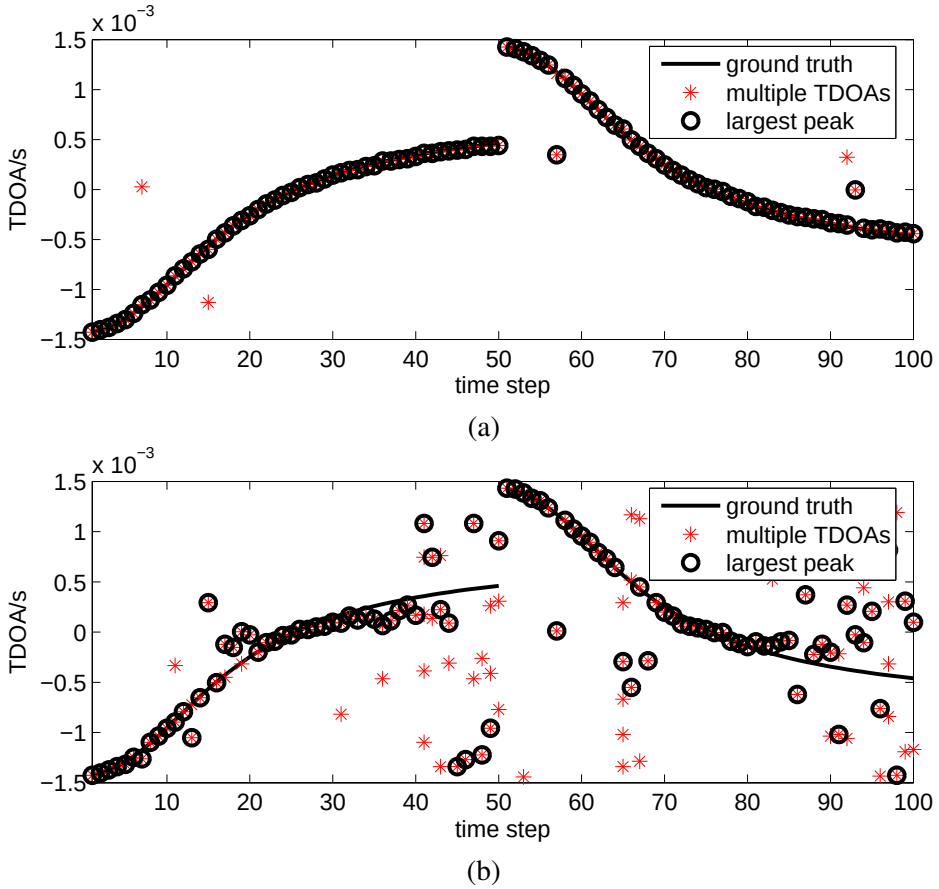


Figure 5.3: *Typical TDOA estimates from microphone pair 1 under the reverberant environment (a) $\rho = 0.4$, $T_{60} = 0.124s$; (b) $\rho = 0.8$, $T_{60} = 0.289s$. All the peaks above a threshold of 0.7 are picked to obtain the TDOAs.*

5.3 Extended Kalman filtering

Since the EKF will construct a basis of the derivation of the following EKPF tracking approach, it is fully presented in this section in terms of TDOA based acoustic source tracking. The derivation of the EKF in this section can also be found in [8].

5.3.1 Local linearisation

Since the measurement function (5.2) is nolinear, it has to be linearised to implement an EKF. The first-order Taylor expansion on $\tau_k^\ell(\mathbf{x}_k)$ from (5.2) can be stated as [8]:

$$\tau_k^\ell(\mathbf{x}_k) = \tau_k^\ell(\mathbf{x}_{k-1}) + \mathbf{c}_k^\ell \left[\mathbf{x}_k - \mathbf{x}_{k-1}\right]^T + \bar{n}_k, \tag{5.12}$$

where superscript T denotes transpose, $\bar{n}_k = O_{\mathbf{x}}(\mathbf{x}_k)$ is the higher order error of the time delay expansion, and $\mathbf{c}_k^\ell$ is the coefficient vector of Taylor expansion, given as

$$\mathbf{c}_k^\ell = \frac{1}{c}\left[\frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,1}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\|} - \frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,2}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}\right]\Bigg|_{\mathbf{x}_k = \mathbf{x}_{k-1}}. \tag{5.13}$$

Substituting the linearisation (5.12) into the ML criterion (5.3), one can get

$$\begin{aligned}\hat{\mathbf{x}}_k &\approx \underset{\mathbf{x}_k}{\operatorname{argmin}} \sum_{\ell=1}^{L} \left(\hat{\tau}_k^\ell - \tau_k^\ell(\mathbf{x}_{k-1}) - \mathbf{c}_k^\ell\left[\mathbf{x}_k - \mathbf{x}_{k-1}\right]\right)^2 \\ &= \underset{\mathbf{x}_k}{\operatorname{argmin}} \sum_{\ell=1}^{L} \left(\bar{\tau}_k^\ell - \mathbf{c}_k^\ell \mathbf{x}_k\right)^2, \end{aligned} \tag{5.14}$$

where

$$\bar{\tau}_k^\ell = \hat{\tau}_k^\ell - \tau_k^\ell(\mathbf{x}_{k-1}) + \mathbf{c}_k^\ell \mathbf{x}_{k-1}, \tag{5.15}$$

where $\hat{\tau}_k^\ell$ is the TDOA measurement extracted from GCC function, and $\tau_k^\ell(\mathbf{x}_{k-1})$ is the TDOA calculated from equation (5.2). The nonlinear measurement can thus be approximated as

$$\bar{\tau}_k^\ell = \mathbf{c}_k^\ell \mathbf{x}_k + \bar{n}_k. \tag{5.16}$$

Here the new measurement $\bar{\tau}_k^\ell$ has a linear form with the state $\mathbf{x}_k$. This process is the same as the linearisation described in Fig. 3.2, on page 55. Taking all the L microphone pairs into consideration, we can write the linearisation in a vector form. Define

$$\bar{\mathbf{z}}_k = \begin{bmatrix} \bar{\tau}_k^1 \\ \bar{\tau}_k^2 \\ \vdots \\ \bar{\tau}_k^L \end{bmatrix}, \qquad \mathbf{z}_k = \begin{bmatrix} \hat{\tau}_k^1 \\ \hat{\tau}_k^2 \\ \vdots \\ \hat{\tau}_k^L \end{bmatrix}, \tag{5.17}$$

and

$$\mathbf{C}_k = \begin{bmatrix} \mathbf{c}_k^1 \\ \mathbf{c}_k^2 \\ \vdots \\ \mathbf{c}_k^L \end{bmatrix}, \qquad \boldsymbol{\tau}_k(\mathbf{x}_k) = \begin{bmatrix} \tau_k^1(\mathbf{x}_k) \\ \tau_k^2(\mathbf{x}_k) \\ \vdots \\ \tau_k^L(\mathbf{x}_k) \end{bmatrix}, \tag{5.18}$$

and follow the equation (5.16), one can get the matrix form of the linear expression

$$\bar{\mathbf{z}}_k \approx \mathbf{C}_k \mathbf{x}_k + \mathbf{w}_k, \tag{5.19}$$

with the new measurements calculated as

$$\bar{\mathbf{z}}_k = \mathbf{z}_k - \boldsymbol{\tau}_k(\mathbf{x}_{k-1}) + \mathbf{C}_k \mathbf{x}_{k-1}. \tag{5.20}$$

Here $\mathbf{w}_k$ is assumed to be zero-mean Gaussian process with a variance of $\mathbf{R}_k$ which includes the higher order expansion error and the TDOAs measurement noise.

5.3.2 Tracking based on EKF

The application of EKF in the target tracking problem can be found in the open literature [73, 114]. Following the EKF algorithm depicted in Section 3.1.4 on page 54, and regarding equation (5.5) as the process equation, the state prediction can be obtained as

$$\mathbf{x}_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}_{k-1} + \sqrt{\mathbf{Q}_k}\mathbf{v}_k, \tag{5.21a}$$

$$\mathbf{P}_{k|k-1} = \hat{\mathbf{P}}_{k-1} + \mathbf{Q}_k, \tag{5.21b}$$

where $\hat{\mathbf{x}}_{k-1}$ is the state estimation at previous time step $k-1$, and $\mathbf{x}_{k|k-1}$ and $\mathbf{P}_{k|k-1}$ are the predicted source state and variance matrix respectively. The corresponding TDOA can be evaluated according to the measurement function (5.2), given as

$$\boldsymbol{\tau}_k(\mathbf{x}_{k|k-1}) = \left[\tau_k^1(\mathbf{x}_{k|k-1}), \tau_k^2(\mathbf{x}_{k|k-1}), \ldots, \tau_k^L(\mathbf{x}_{k|k-1})\right]^T. \tag{5.22}$$

Following the standard Kalman filtering procedure, the EKF gain can be calculated as

$$\mathbf{S}_k = \mathbf{R}_k + \mathbf{C}_k \mathbf{P}_{k|k-1} \mathbf{C}_k^T, \tag{5.23a}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{C}_k^T \mathbf{S}_k^{-1}. \tag{5.23b}$$

Finally the source state and the variance matrix are updated as

$$\hat{\mathbf{x}}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k(\mathbf{z}_k - \boldsymbol{\tau}_k(\mathbf{x}_{k|k-1})), \tag{5.24a}$$

$$\hat{\mathbf{P}}_k = \mathbf{P}_{k|k-1} - \mathbf{K}_k \mathbf{C}_k \mathbf{P}_{k|k-1}. \tag{5.24b}$$

Since the initial distribution is assumed to be Gaussian, the filtered distribution of the source state is also Gaussian and following the distribution as: $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_{1:k}) = \mathcal{N}(\mathbf{x}_k; \hat{\mathbf{x}}_k, \hat{\mathbf{P}}_k)$. This posterior distribution will be used as the importance function in the particle filter for the derivation of EKPF in Section 5.5.

The complete tracking algorithm is summarised in Algorithm 6. The EKF works well when the measurement noise is assumed to be Gaussian and the nonlinearity of the measurement function is relatively moderate [8, 9]. However, it only regards a single TDOA from each microphone pair (corresponding to the largest peak in the PHAT-GCC function) as its measurement. As mentioned in Section 5.2.3, due to the background noise and reverberation, the GCC function may present a number of dominant peaks, and all of these peaks are possibly corresponding the correct TDOA estimation. This phenomenon can also be found in Fig. 5.3(b). At several time steps (e.g., time step 19, 21 and 65), the correct TDOA estimates are extracted from some other peaks rather than from the largest peak in GCC function. The tracking performance is thus not satisfied by using this single-measurement tracking approach. In the next section, the particle filtering approach, which is able to incorporate the reverberant measurement model will be fully introduced.

Algorithm 6: EKF for acoustic source tracking.

Input: TDOA measurements $\mathbf{z}_{1:K}$.
Output: Sources position estimates $\mathbf{x}_k$.

Initialisation: $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\hat{\mathbf{P}}_0 \leftarrow \mathbf{P}_0$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
- predict the state $\mathbf{x}_{k|k-1}$ and the corresponding variance matrix $\mathbf{P}_{k|k-1}$ according to (5.21);
- calculate the predict measurements $\boldsymbol{\tau}(\mathbf{x}_{k|k-1})$ according to (5.22);
- calculate the filtering gain $\mathbf{K}_k$ according to (5.23b);
- update the state $\hat{\mathbf{x}}_k$ and variance matrix $\hat{\mathbf{P}}_k$ according to (5.24);
- output the estimates $\hat{\mathbf{x}}_k$.

end

5.4 Particle filtering

This section introduces the single acoustic source tracking applying general SIR-PF [6, 12] in the reverberant environment. Since the reverberant measurement model addressed in Section

5.2.3 is incorporated, the likelihood for the multiple TDOAs is formulated first. The importance function is then derived and complete tracking algorithm is presented.

5.4.1 Likelihood formulation

Consider the ground truth TDOA τ_k^ℓ and TDOA measurements $\mathbf{z}_k^\ell$ from the ℓth microphone pair. If the TDOA measurements can be assumed to be independent, the likelihood function can simply be written as the production of the likelihood from each independent TDOA, given as

$$p(\mathbf{z}_k^\ell|\tau_k^\ell) = \prod_{i=1}^{n_k^\ell} p(\hat{\tau}_{i,k}^\ell|\tau_k^\ell). \tag{5.25}$$

Since for each TDOA measurement vector $\mathbf{z}_k^\ell$ collected from a microphone pair, at most one TDOA is directly generated by the source (other peaks are generated by clutter), a variable $\lambda_{i,k}$, $i = 1, \ldots, n_k^\ell$ can be defined to indicate the association between each measurement and the source,

$$\lambda_{i,k} = \begin{cases} 1, & \text{the measurement is a detection;} \\ 0, & \text{the measurement is a false alarm.} \end{cases} \tag{5.26}$$

Based on this association on each independent TDOA measurement, two categories of hypotheses can be summarised for all the measurements obtained from a microphone pair, stated as

$$\begin{aligned} \mathcal{H}_{0,k} &\triangleq \{\lambda_{i,k} = 0; i = 1, \ldots, n_k^\ell\}, \\ \mathcal{H}_{i,k} &\triangleq \{\lambda_{i,k} = 1, \lambda_{j,k} = 0; i, j = 1, \ldots, n_k^\ell, j \neq i\}, \end{aligned} \tag{5.27}$$

where $\mathcal{H}_{0,k}$ denotes that none of the measurements is generated by the source, and $\mathcal{H}_{i,k}$ represents that the ith TDOA measurement $\tau_{i,k}^\ell$ is generated by the source, and all other TDOAs are generated by the clutter.

If the measurement is generated by the clutter, i.e., $\lambda_{i,k} = 0$, the likelihood is assumed to be a uniform distribution within the admissible TDOA range, given as

$$p(\hat{\tau}_{i,k}^\ell|\tau_k^\ell, \lambda_{i,k} = 0) = \mathcal{U}_\tau(\hat{\tau}_{i,k}^\ell)\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell) = \frac{1}{2\tau_{\text{max}}}\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell), \tag{5.28}$$

where $\tau = [-\tau_{\text{max}}\ \tau_{\text{max}}]$ denotes the possible TDOA range, and $\tau_{\text{max}} = d/c$ is the maximum TDOA. $\mathbb{I}_\tau(\cdot)$ is an indicator function that confines the TDOA measurements within the admis-

sible TDOA range τ. According to the equation (5.25), the likelihood for the hypotheses $\mathcal{H}_{0,k}$ can be expressed as

$$\begin{aligned} p(\mathbf{z}_k^\ell|\tau_k^\ell, \mathcal{H}_{0,k}) &= \prod_{i=1}^{n_k^\ell} p(\hat{\tau}_{i,k}^\ell|\tau_k^\ell, \lambda_{i,k}=0)\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell) \\ &= \frac{1}{(2\tau_{\max})^{n_k^\ell}}\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell). \end{aligned} \tag{5.29}$$

If the measurement is generated by a real source, the likelihood is assumed to be the true TDOA corrupted by an additive Gaussian noise with variance σ_τ^2 [6, 12, 115]. The likelihood can then be written as

$$p(\hat{\tau}_{i,k}^\ell|\tau_k^\ell, \lambda_{i,k}=1) = \mathcal{N}(\hat{\tau}_{i,k}^\ell; \tau_k^\ell, \sigma_\tau^2)\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell), \tag{5.30}$$

The general expression for the hypotheses $\mathcal{H}_{i,k}$ is thus

$$\begin{aligned} p(\mathbf{z}_k^\ell|\tau_k^\ell, \mathcal{H}_{i,k}) &= p(\hat{\tau}_{i,k}^\ell|\tau_k^\ell, \lambda_{i,k}=1)\prod_{\substack{j=1\\ j\neq i}}^{n_k^\ell} p(\hat{\tau}_{j,k}^\ell|\tau_k^\ell, \lambda_{j,k}=0) \\ &= \frac{1}{(2\tau_{\max})^{n_k^\ell-1}}\mathcal{N}(\hat{\tau}_{i,k}^\ell; \tau_k^\ell, \sigma_\tau^2)\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell). \end{aligned} \tag{5.31}$$

From now on, the likelihood for the two hypothesis categories in (5.27) have been formulated. The problem now is that for a given measurement vector $\mathbf{z}_k^\ell$, we don't know which one is generated by the real source. The correct hypothesis $\mathcal{H}_{i,k}$ is thus unknown *a priori*. In [6, 12], all the collected TODA estimates are deemed as equally important. Assume that the prior probability for $\mathcal{H}_{0,k}$ is q_0, i.e., $p(\mathcal{H}_{0,k}|\tau_k^\ell) = q_0$. The prior probability $p(\mathcal{H}_{i,k}|\tau_k^\ell)$ is thus equally weighted, given as

$$p(\mathcal{H}_{i,k}|\tau_k^\ell) = \frac{1-q_0}{n_k^\ell}; \qquad \forall \quad i = 1, \ldots, n_k^\ell. \tag{5.32}$$

Further, the complete likelihood over all the hypotheses from the ℓth microphone pair can be obtained by summing over all the hypotheses, that is

$$\begin{aligned} p(\mathbf{z}_k^\ell|\tau_k^\ell) &= \sum_{i=0}^{n_k^\ell} p(\mathcal{H}_{i,k}|\tau_k^\ell)p(\mathbf{z}_k^\ell|\tau_k^\ell, \mathcal{H}_{i,k}) \\ &= \frac{1}{(2\tau_{\max})^{n_k^\ell-1}}\left(\frac{q_0}{2\tau_{\max}} + \frac{1-q_0}{n_k^\ell}\sum_{i=1}^{n_k^\ell}\mathcal{N}(\hat{\tau}_{i,k}^\ell; \tau_k^\ell, \sigma_\tau^2)\mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell)\right). \end{aligned} \tag{5.33}$$

Since the measurements collected from all the microphone pairs are assumed to be independent, the extension of the likelihood from a microphone pair to over all L microphone pairs is straightforward

$$p(\mathcal{Z}_k|\boldsymbol{\tau}_k) = \prod_{\ell=1}^{L} p(\mathbf{z}_k^\ell|\tau_k^\ell), \tag{5.34}$$

where $p(\mathbf{z}_k^\ell|\tau_k^\ell)$ is given in (5.33). This likelihood depicts all kinds of hypotheses generated by the measurement set. It performs well in a number of acoustic source tracking scenarios [6,12,115]. Different likelihood constructions are discussed in [116], in which this intersection-based likelihood combination approach is further shown more robust than other models. The original work of this likelihood derivation is presented in [12].

5.4.2 Tracking based on PF

After defining the source dynamic model and the likelihood, the SIR algorithm depicted in section 3.2.4 on page 62 is easily implemented for acoustic source tracking. The key step in deriving the SIR algorithm is to compute the importance weights. Given a state sample $\mathbf{x}_{k-1}^{(i)}$, according to the importance updating equation (3.50) on page 61, the iterative updating of importance weight for acoustic source tracking can be written as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k|\boldsymbol{\tau}_k^{(i)}) p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k})}, \tag{5.35}$$

where $p(\mathcal{Z}_k|\boldsymbol{\tau}_k^{(i)})$ is the likelihood computed for each sample $\mathbf{x}_k^{(i)}$, and $p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$ is the transition density which can be evaluated from the source dynamic model (5.5).

The only problem left for applying a SIR filter is thus how to design the proposal distribution $q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k})$. In general SIR-PF approaches [6, 12], the samples are simply drawn according to the transition density $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, which leads to

$$q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k}) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}), \tag{5.36}$$

and the importance weight is updated accordingly as

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(\mathcal{Z}_k|\boldsymbol{\tau}_k^{(i)}). \tag{5.37}$$

This importance weight updating approach is firstly developed in [6]. One advantage is that it is

simple and easy to use, and sufficient to track a slow-paced moving source. In particular, if the source is following a simple trajectory, e.g., a diagonal line trajectory in Fig. 2.9 on page 44, the Langevin dynamical model is accurate enough to model the source motion. The samples are thus drawn efficiently and the general SIR-PF is found very robust in the adverse environment. However, the disadvantage is that this approach does not employ the information from the current measurements $\mathbf{z}_k$ to sample the particles, i.e., the proposal distribution in equation (5.36). Our later experiments will show that there will be a tracking-lag if the dynamical model mismatching happens.

The SIR tracking algorithm with prior importance function is summarised in Algorithm 7. The algorithm is the same as a general SIR particle filtering except the way that the likelihood is formed. In practice, the prior probability of q_0 (which represents association probability of all the measurements to the clutter) is obtained by empirical study. Higher value of q_0 denotes a lower probability of detection, and all the measurements are more likely to be associated with clutter. The advantage of the SIR particle filtering over the series of Kalman filtering in acoustic source tracking is that the SIR particle filtering can be directly apply to nonlinear measurement function. Further more, it allows a combination of multiple potential TDOA measurements, and thus explores a larger detection probability (of course, a larger false alarm rate as well), but without being affected by the large false alarm significantly.

5.5 Better Proposal Distribution: EKPF

Using the transition pdf as the proposal distribution has its advantage of easy implementation. However, it does not take the current measurements into account. That means the particles are drifting from the previous position estimates rather than the area with high likelihood. In this section, we attempt to derive an 'optimal' proposal distribution by employing an extended Kalman filtering. After the EKF step, all the particles are relocated according the information from the previous estimates as well as the current measurements, and thus at the high likelihood area. This extended Kalman particle filtering (EKPF) approach is firstly developed in the field of filtering theory in [113], and later widely employed in the target tracking problem [114]. The basic idea is that coarsely estimate the posterior distribution by employing an EKF first, and the samples are then drawn from this posterior distribution and refined by the particle filter.

This approach is novel in that it introduces the EKPF in acoustic source tracking. Compared to

Algorithm 7: SIR particle filtering for source tracking.

Input: Current TDOA measurements $\mathcal{Z}_k$.
Output: Sources position estimates $\mathbf{x}_k$.

Initialisation: draw particles $\hat{\mathbf{x}}_0^{(i)} \sim \mathcal{N}(\hat{\mathbf{x}}_0^{(i)}; \mathbf{x}_0, \mathbf{P}_0)$, and set the initial weight $\tilde{w}_0^{(i)} = 1/N$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
  Over all the particles:
  for $i \leftarrow 1$ **to** N **do**
    sampling $\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}) = \mathcal{N}(\hat{\mathbf{x}}_0^{(i)}; \mathbf{A}_k\mathbf{x}_{k-1}^{(i)}, \mathbf{Q}_k)$;
    transform the position part of the state samples into the TDOA measurements according to equation (5.2);
    computing the likelihood according to equation (5.34);
    computing the importance weight according to the equation (5.37):

$$w_k^{(i)} \propto \tilde{w}_{k-1}^{(i)} p(\mathcal{Z}_k|\boldsymbol{\tau}_k^{(i)});$$

  end
  Over all the particles:
  for $i \leftarrow 1$ **to** N **do**
    normalise the importance weight:

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{N} w_k^{(i)}};$$

  end
  replicate/discard the particles according to the high/low weights.
  output the estimates $\hat{\mathbf{x}}_k$.
end

the EKF approach [8], the EKPF developed here is able to incorporate the reverberant measurement model described in Section 5.2.3. Compared to the SIR-PF approach [6], the proposal distribution derived from the EKF is approximately optimal, and particles are drawn more efficiently.

5.5.1 Hypothesis prior based on PHAT-GCC amplitude

Other than the TDOA measurement itself, the corresponding PHAT-GCC function amplitude also carries some information for identifying the detections and the false alarms. Generally, the higher the amplitude, the more likely it is generated by the source. This phenomenon can also be found in Fig. 4.7, on page 86, where the RMS amplitude generated by detections is overwhelmingly larger than that generated by the false alarms.

Given the TDOA measurement vector $\mathbf{z}_k^\ell$ collected at the ℓth microphone pair, the corresponding TDOA amplitude from PHAT-GCC function can be written as

$$\{\hat{a}_{1,k}^\ell, \ldots, \hat{a}_{n_k^\ell,k}^\ell\}, \tag{5.38}$$

where $\hat{a}_{i,k}^\ell$, for $i = 1, \ldots, n_k^\ell$ is the amplitude of the TDOA measurement $\hat{\tau}_{i,k}^\ell$ obtained from the GCC function directly. The hypothesis prior model depicted by equation (5.32) assumes that all the TDOA measurement are equally important for the state estimation, and the prior for all the hypotheses $\mathcal{H}_{i,k}$, for $i = 1, \ldots, n_k^\ell$ are the same.

The assumption of the equal prior for all the hypothesis is true when the source is existed in an extremely noisy and reverberant environment. However, in the moderate reverberant or low reverberant environments, most of the TDOA detections come from those higher peaks. It is thus desired to incorporate the TDOA amplitude information into the hypothesis prior to make the final likelihood model appropriate in the different environments. Let the prior of the hypothesis $\mathcal{H}_{0,k}$ be q_0, and $\pi_{i,k}^\ell = \frac{\hat{a}_{i,k}^\ell}{\sum_i \hat{a}_{i,k}^\ell}$. The prior q_i of the hypothesis $\mathcal{H}_{i,k}$ can be calculated as

$$p(\mathcal{H}_{i,k}|\tau_k^\ell) = (1 - q_0)\pi_{i,k}^\ell; \qquad \forall \quad i = 1, \ldots, n_k^\ell. \tag{5.39}$$

This prior choice is to make the summation of all the priors equal to one, states as

$$\sum_{i=0}^{n_k^\ell} p(\mathcal{H}_{i,k}|\tau_k^\ell) = 1. \tag{5.40}$$

The likelihood model for the ℓth microphone pair is thus

$$\begin{aligned} p(\mathbf{z}_k^\ell|\tau_k^\ell) &= \sum_{i=0}^{n_k^\ell} p(\mathcal{H}_{i,k}|\tau_k^\ell) p(\mathbf{z}_k^\ell|\tau_k^\ell, \mathcal{H}_{i,k}) \\ &= \frac{1}{(2\tau_{\max})^{n_k^\ell - 1}} \left(\frac{q_0}{2\tau_{\max}} + (1 - q_0) \frac{\hat{a}_{i,k}^\ell}{\sum_i \hat{a}_{i,k}^\ell} \mathcal{N}(\hat{\tau}_{i,k}^\ell; \tau_k^\ell, \sigma_\tau^2) \mathbb{I}_\tau(\hat{\tau}_{i,k}^\ell) \right). \end{aligned} \tag{5.41}$$

The complete likelihood function for all the microphone pairs can be easily obtained according to equation (5.34). Rather than the SIR-PF approach in [6], where mutliple TDOAs are equally treated, this novel hypothesis prior emphasises the TDOAs from the higher PHAT-GCC peaks. In the noisy and reverberant environments, especially where a low threshold is chosen, taking the amplitude information into account is able to suppress the effect of the false alarms.

5.5.2 Proposal distribution

Figure 5.4 shows the different schemes of proposal distribution between our importance function and the transition prior importance function. If there is a sharp change of the source position (which is especially the case of the existence of multiple nonconcurrent speakers), the current source position varies hugely from its previous position estimates. The EKPF employs an EKF to estimate the state first, and the samples are then drawn around this posterior state estimate. The EKF proposal distribution thus leads to a more efficient sampling, in contrast to the general SIR-PF which draws the samples around the previous state estimates.

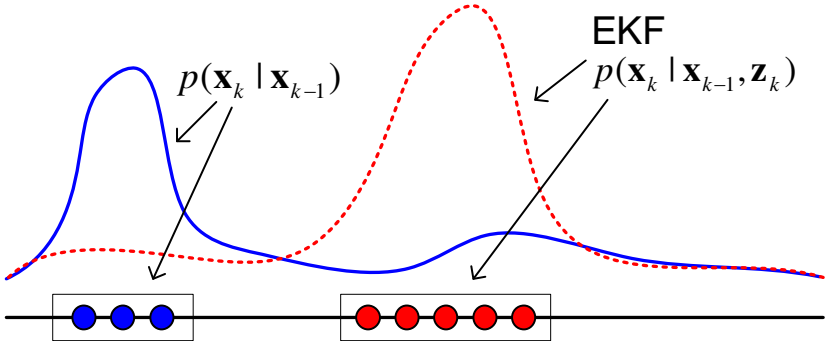


Figure 5.4: *Sampling from a prior distribution* vs. *sampling from an EKF posterior distribution.*

To enhance the probability of detection in the adverse environment, it is expected that the EKPF is able to take the multiple TDOAs from each microphone pair into account and incorporate the reverberant measurement model depicted in section 5.2.3, on page 122. Two approaches are developed to formulate the EKF step: 1) traditional EKF which only employs the TDOAs from the highest peaks of PHAT-GCC functions; 2) all the TDOA measurements are used to update the states, and a new innovation process is formulated using a parameter to model the effect of the false alarms.

Traditional EKF approach. The EKF formulation here is the same as the traditional one which uses the highest peak TDOAs from each microphone pair as the measurements. It is plausible to do this since in most cases, the highest peaks are very likely generated by the real source. The only difference is that a parameter is used to model the effect from the false alarms in the innovation process (equation (5.24a) on page 125), given as

$$\mathbf{y}_k = (1 - q_1)\left(\mathbf{z}_k^{\max} - \boldsymbol{\tau}_k(\mathbf{x}_{k|k-1})\right), \tag{5.42}$$

where $\mathbf{z}_k^{\max}$ are the measurements collected from each microphone pair with highest GCC peaks, and q_1 is a constant controlling the rate of the innovation from the measurements. The false alarms are modelled by carefully choosing the constant q_1, which is usually determined by the experimental study. Normally a smaller value of q_1 denotes that a reliable proposal distribution can be obtained by the EKF, and vice versa.

Multiple TDOA EKF approach. In this approach, all the TDOA measurements are used to update the samples. The innovation process of EKF is

$$\mathrm{y}_k^\ell = (1-q_1)\sum_{i=1}^{n_k^\ell} \pi_{i,k}^\ell \left(\mathbf{z}_k^\ell - \boldsymbol{\tau}_k^\ell(\mathbf{x}_{k|k-1})\right), \tag{5.43}$$

and the whole innovation vector is

$$\mathbf{y}_k = \left[\mathrm{y}_k^1, \ldots, \mathrm{y}_k^L\right]^T. \tag{5.44}$$

This innovation process is different from that in traditional EKF approach since all the TDOAs collected from the microphone pair are employed, and those TDOAs with higher peak amplitudes are regarded as more important measurements to the final state estimation.

It is worth pointing out that the EKPF using multiple TDOA measurements but without amplitude information is not considered here. Since only the GCC peaks above a certain threshold ($R_{\mathrm{TH}} = 0.7$ here) are used, the amplitudes of all these peaks are relatively large. The TDOAs are thus almost equally treated even though the amplitude priors are employed.

According to the update equation (5.24a), the state estimates can be written as

$$\bar{\mathbf{x}}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k \mathbf{y}_k. \tag{5.45}$$

The updated variance $\bar{\mathbf{P}}_k$ remains the same expression as it in equation (5.24b). Since each particle is redrawn according to this EKF step, the proposal distribution becomes

$$\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)} | \mathbf{x}_{0:k-1}^{(i)}, \mathcal{Z}_{1:k}) = \mathcal{N}(\mathbf{x}_k^{(i)}; \bar{\mathbf{x}}_k^{(i)}, \bar{\mathbf{P}}_k^{(i)}), \tag{5.46}$$

where $\bar{\mathbf{x}}_k^{(i)}$ and $\bar{\mathbf{P}}_k^{(i)}$ are the mean and covariance of Gaussian distribution for each particle respectively derived by the EKF. After the EKF step, the particles are relocated around the posterior distribution roughly. The final position estimates are obtained by further implementing a

particle filtering on these samples.

5.5.3 EKPF tracking algorithm

So far how to use the EKF to generate the proposal distribution has been discussed. The likelihood is constructed in the same way as described in Section 5.4.1. The prior distribution is the same as in general SIR-PF approach, in which it is determined by the transition probability

$$p(\mathbf{x}_k^{(i)}|\bar{\mathbf{x}}_{k-1}^{(i)}, \mathcal{Z}_{k-1}) = \mathcal{N}(\mathbf{x}_k^{(i)}; \mathbf{A}_k\bar{\mathbf{x}}_{k-1}^{(i)}, \mathbf{Q}_k). \tag{5.47}$$

The only difference here is that the states are sampled by the EKF. The weights are updated as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k|\boldsymbol{\tau}_k^{(i)})p(\mathbf{x}_k^{(i)}|\bar{\mathbf{x}}_{k-1}^{(i)})}{p(\mathbf{x}_k^{(i)}|\bar{\mathbf{x}}_{0:k-1}^{(i)}, \mathcal{Z}_{1:k})}, \tag{5.48}$$

which has the same form as in general SIR filter except that the particles are relocated by the EKF.

The EKPF tracking algorithm is summarised in Algorithm 8. The tracking framework is the same as that of SIR-PF tracking algorithm Algorithm 7 except two differences. First, EKPF employs an EKF step to draw the samples. Second, the calculation of the importance weight is different. The tracking algorithm actually adds a preprocessing step which uses an EKF to coarsely filter the predicted particles. The particles are thus redrawn at a high likelihood area rather than drifting from the motion dynamical equation. Traditional SIR-PF fails to do so since the particles are drawn only using the information from the motion dynamical equation, and a tracking-lag will be presented in catching up with the position of a new source. Of course, a sophisticated motion dynamical model may help to relief the tracking-lag brought by the model mismatch in general SIR-PF approach. However, such investigation is another perspective of acoustic source tracking and we will leave it as our future work.

5.6 Experiments

In this section, a series of experiments are presented to examine the tracking performance of all the tracking approaches introduced in the above sections. These approaches include:

Algorithm 8: EKPF for acoustic source tracking.

Input: Current TDOA measurements $\mathcal{Z}_k$.
Output: Sources position estimates $\hat{\mathbf{x}}_k$.

Initialisation: draw particles $\hat{\mathbf{x}}_0^{(i)} \sim \mathcal{N}(\hat{\mathbf{x}}_0^{(i)}; \mathbf{x}_0, \mathbf{P}_0)$, and set the initial weight $\tilde{w}_0^{(i)} = 1/N$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
  Over all the particles:
  for $i \leftarrow 1$ **to** N **do**
    implement EKF to obtain the new samples $\bar{\mathbf{x}}_k^{(i)}$;
    computing the likelihood according to equation (5.41);
    computing the importance weight according to the equation (5.48);
  end
  Over all the particles:
  for $i \leftarrow 1$ **to** N **do**
    normalise the importance weight:

$$\tilde{w}_k^{(i)} = \frac{w_k^{(i)}}{\sum_{i=1}^{N} w_k^{(i)}};$$

  end
  replicate/discard the particles according to the high/low weights.
  output the estimates $\hat{\mathbf{x}}_k$.
end

- EKF approach;
- general SIR-PF approach;
- EKPF with a traditional EKF step and incorporating the amplitude information in the likelihood model (EKPF);
- EKPF employs multiple TDOA EKF approach and incorporating the amplitude information in the likelihood model (multiTDOA-EKPF).

The experimental setup is described in Section 2.6.2 on page 43 and Section 2.6.3 on page 45. As mentioned in Section 5.5.2, since the amplitudes of GCC peaks are relatively large here, the prior in equation (5.39) of EKPF using multiple TDOAs but without amplitude information will be very similar as that in multiTDOA-EKPF. The tracking results of multiTDOA-EKPF without using the amplitude information is thus not presented here.

5.6.1 Root Mean Square Error

The position estimates can be obtained from each single run of the tracking algorithm directly. To fully assess and compare the tracking performance of these algorithms, different parameters have to be defined based on the position outputs of the algorithms over many Monte Carlo implementations. The root mean square error (RMSE) which directly illustrates the estimation divergence from the ground truth is employed to evaluate the tracking performance.

Suppose that J Monte Carlo simulations are implemented, and let $\hat{\mathbf{x}}_{j,k}$, $j = 1, \ldots, J$ and $\mathbf{x}_k$ represent the estimated source state at jth implementation and the ground truth, respectively. The Euclidean distance error between $\hat{\mathbf{x}}_{j,k}$ and $\mathbf{x}_k$ is defined as

$$d(\hat{\mathbf{x}}_{j,k}, \mathbf{x}_k) = \left\| \mathbf{B}\hat{\mathbf{x}}_{j,k} - \mathbf{B}\mathbf{x}_k \right\|, \tag{5.49}$$

where $\mathbf{B} = [\mathbf{I}\ \mathbf{0}]$ is a position extraction matrix such that $\mathbf{B}\mathbf{x}_k$ outputs the position part (x_k, y_k) of the state vector $\mathbf{x}_k$. The RMSE is defined as the root mean square of Euclidean distance error $d(\hat{\mathbf{x}}_{j,k}, \mathbf{x}_k)$ over the total number of Monte Carlo experiments J, given as

$$\varepsilon_k = \sqrt{\frac{1}{J}\sum_{j=1}^{J} d^2(\hat{\mathbf{x}}_{j,k}, \mathbf{x}_k)}. \tag{5.50}$$

Further averaging this error over all the time step leads to the overall RMSE, given as

$$\bar{\varepsilon} = \sqrt{\frac{1}{JK}\sum_{k=1}^{K}\sum_{j=1}^{J} d^2(\hat{\mathbf{x}}_{j,k}, \mathbf{x}_k)}. \tag{5.51}$$

The RMSE ε_k depicts how much the source location estimate deviates from the true source position at each time step. The capability of each algorithm in catching up with the position of new source can thus be illustrated by this parameter. The overall RMSE $\bar{\varepsilon}$ addresses the average error of each experiment. The lower the RMSE of the tracking algorithm presents, the more accurate the tracking algorithm is.

5.6.2 Experiment parameter setup

In this section, different noisy and reverberant environments are simulated to evaluate the performance of the tracking approaches. For all the simulated experiments, the room environ-

ment depicted in Fig. 2.9 on page 44 is used. The dimension of the simulated office room is $5\times4\times3\text{m}^3$. Four microphone pairs each with a 50cm separation are organized around the center of the walls. The audio signals are split into 50 frames with a frame length of 128ms, at a sampling frequency of 8000Hz. All the reverberations in the room are simulated using the imaging method [22], and different SNRs are simulated by adding the Gaussian white noise into the received speech signals. In the experiments, two speakers appear alternatively: one is active from frame index 1 to 50, and the other from the frame index 51 to 100. This leads to a tracking length of 12.8s. The sources are moving at a velocity of 0.5m/s roughly, which is comparable with the source velocities in [1,6]. The motion trajectories are diagonal line trajectories as shown in Fig. 2.9 on page 44.

For the source dynamics, the parameters v and β are set to 1ms^{-1} and 10s^{-1} respectively. This parameter setup is used in [1,6,12] and found accurate enough to model the source movement in our experimental studies. Since it is assumed that there is no prior information about the initial source position, the source position are initialised at the center of the room with the velocity on both orientations of 0.4m/s, i.e., $\mathbf{x}_0 = (2.5\ \ 2.0\ \ 0.4\ \ 0.4)^T$. The corresponding initial variance is set as

$$\mathbf{P}_0 = \text{diag}([1\ \ 1\ \ 0.1\ \ 0.1]), \tag{5.52}$$

with $\text{diag}(\cdot)$ denoting a diagonal matrix. These parameters are summarised in Table 5.3. Changing the parameters for the source dynamics and initialisation will lead to different converging velocity of the algorithm.

Other parameters for the tracking algorithms can be found in Table 5.4. The variance of the measurement noise $\mathbf{R}$ for the EKF is set the same as that for the EKPF, which is 1.25e^{-4}, one sample diverging from the measured TDOA. After the EKF step, the samples are relocated around the posterior distribution. The variance σ^2 in EKPF is thus set smaller than it in the general SIR-PF. q_0 and q_1 are the parameters which depict the affection of the reverberation in

parameter	value
v	$1\ \text{ms}^{-1}$
β	$10\ \text{s}^{-1}$
$\mathbf{x}_0$	$(2.5\ \ 2.0\ \ 0.4\ \ 0.4)^T$
$\mathbf{P}_0$	diag$[1\ \ 1\ \ 0.1\ \ 0.1]$

Table 5.3: *Parameters for Langevin motion model and initialisation.*

the PF and EKF step respectively and are chosen by the experiment studying.

It is worth pointing out that except the state initialisation parameter $\mathbf{x}_0$, the choices of all other parameters in Table 5.3 and Table 5.4 are based on extensive experimental studies. These parameters are found accurate enough for the experiments in the next sections and slightly different setup will not lead a significant difference on the tracking results.

parameter / algorithm	R	σ	q_0	q_1	N
EKF	0.5e^{-4}	-	-	-	-
SIR-PF	-	1.25e^{-4}	0.2	-	500
EKPF	1.25e^{-4}	0.5e^{-4}	0.2	0.1	100
multiTDOA-EKPF	1.25e^{-4}	0.5e^{-4}	0.2	0.1	100

Table 5.4: *Parameter setup for the tracking algorithms.*

5.6.3 Experiment results under the simulated room environments

5.6.3.1 Single experiment

First, the tracking results from a single experiment under the reverberation time $T_{60} = 0.163$s (with the wall reflection coefficients of 0.6) are presented. As described in Section 5.2.3, the threshold on the GCC function R_{TH} is set to be 0.7. Since the EKF method can not incorporate the reverberant measurement model, it only takes the TDOAs corresponding the largest peak as its measurements. The multiTDOA-EKPF and SIR-PF regard all the picked TDOAs (which are located in the admissible TDOA range $[-\tau_{\text{max}}\ \tau_{\text{max}}]$) as the measurements. The probability of detection and the false alarm rate for EKF and PF/multiTDOA-EKPF are given in Table 5.5. The TDOA measurements from microphone pair 1 and microphone pair 3 are displayed in Fig. 5.5. As multiple TDOAs are picked, the probability of detection can be enhanced. However, the false alarm rate increases as well.

Fig. 5.6 represents the tracking results from a single trial. It shows that the EKF based PF approaches lock on the new source more quickly than the SIR-PF does. Although EKF is able to find the new source quickly, it is not as robust as the EKPF and SIR-PF in dealing with the inaccurate TDOA measurements. Due to incorporating the reverberation measurement model, the multiTDOA-EKPF employs all the TDOA measurements and presents the best tracking result.

	EKF	PF/EKPF
probability of detection P_D	0.870	0.893
false alarm rate P_F	0.110	0.242

Table 5.5: *Probabilities of detection and false alarm for EKF and PF/multiTDOA-EKPF respectively ($T_{60} = 0.163s$).*

To fully analyse the tracking performance, the RMSE for 100 Monte Carlo runs is presented in Fig. 5.7. It shows that both the EKF based PF approaches are capable of finding the position of the new source quickly. However, at some time steps, some microphone pairs may only report false alarms rather than together with accurate TDOA measurements, e.g., microphone pair one at time step 16. This leads to a heavy false alarm and poor probability of detection at these time steps. The EKF filtering based on these false measurements will present unstable results, and make the following PF no longer able to draw the samples efficiently. This phenomenon can be seen from those peaks in the RMSE made by EKF and EKPF. Since the SIR-PF draws the samples around the previous state estimates, it is not sensitive to the sharp change of the measurements and presents the best performance to smooth the inaccurate mea-

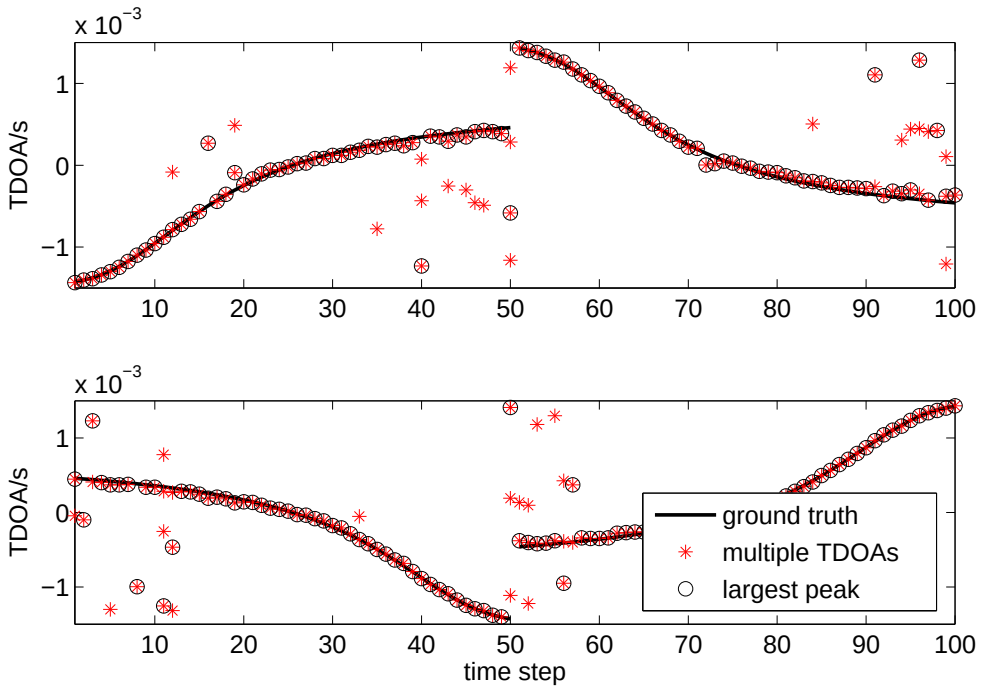


Figure 5.5: *TDOA measurements of microphone pair 1 and microphone pair 3 under the reverberant environment ($T_{60} = 0.163s$).*

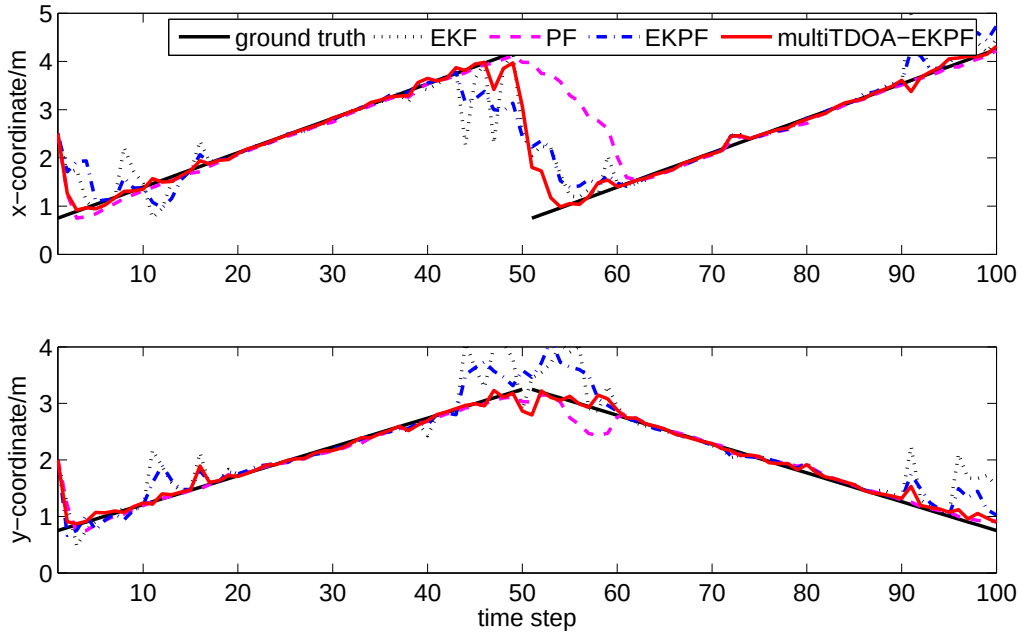


Figure 5.6: *Tracking results from a single trial under the reverberant environment ($T_{60} = 0.163s$)*

surements. However, the drawback is that the PF cannot lock on the new source quickly. The multiTDOA-EKPF approach, which incorporates both the reverberant measurement model and the EKF, is able to cope with false alarms due to reverberation/noise well, and also catch up with the position of the new source quickly.

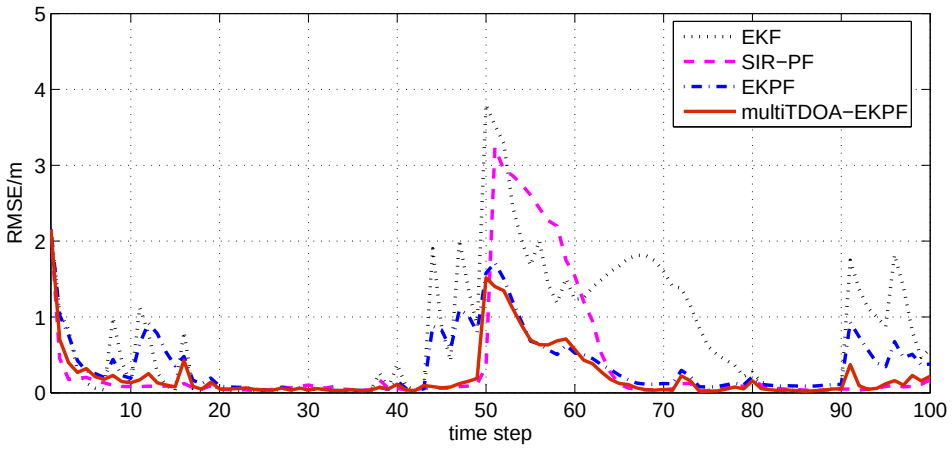


Figure 5.7: *RMSE over 100 Monte Carlo runs under the reverberant environment ($T_{60} = 0.163s$)*

5.6.3.2 Different reverberant and noisy environments

The algorithms are also implemented in different noisy and reverberant environments. The wall reflection efficient is set different values to simulate different reverberant environments, and WGN with different energy levels is added to the received signal to generate different SNRs. Table 5.6 and Table 5.7 present the exact probabilities of detection and false alarm under the different simulated reverberant and noisy environments respectively. For different reflection coefficients, the corresponding reverberation time T_{60} can be found in Table 2.1, on page 45.

	ρ	0.0	0.4	0.6	0.8	0.9
EKF	P_D	0.978	0.963	0.870	0.475	0.288
	P_F	0.022	0.037	0.110	0.369	0.433
PF/EKPF/	P_D	0.980	0.968	0.893	0.520	0.353
mutiTDOA-EKPF	P_F	0.027	0.058	0.242	0.634	0.786

Table 5.6: *Probabilities of detection and false alarm for EKF and PF/multiTDOA-EKPF under different reverberant environments.*

For all different reverberant environments, the SNR is set to be 30 dB. The average RMSE results under different reverberant environments over 100 Monte Carlo runs are presented in Fig. 5.8(a). It shows that the proposed EKPF approaches are able to track the nonconcurrent multiple sources very well in the moderate reverberant environment, and their performance are better than the general SIR-PF and EKF. Particularly, the multiTDOA-EKPF presents the best tracking results since it employs the reverberant measurement model as well as the optimum importance function. Fig. 5.8(b) shows the average RMSE in the different noisy environments. The performance of the proposed EKPF approaches are also better than that of SIR-PF in all experiments. EKF presents the best performance in the anechoic environment and high SNR environments. This is because in such cases the high probability of detection presents, the nonlinearity of the measurement model is able to be approximated by EKF quite well, and the

	SNR (dB)	0	5	10	15	20
EKF	P_D	0.588	0.848	0.948	0.960	0.975
	P_F	0.250	0.119	0.047	0.039	0.024
PF/EKPF/	P_D	0.680	0.885	0.950	0.965	0.983
mutiTDOA-EKPF	P_F	0.588	0.309	0.156	0.070	0.048

Table 5.7: *Probabilities of detection and false alarm for EKF and PF/multiTDOA-EKPF under different noisy environments.*

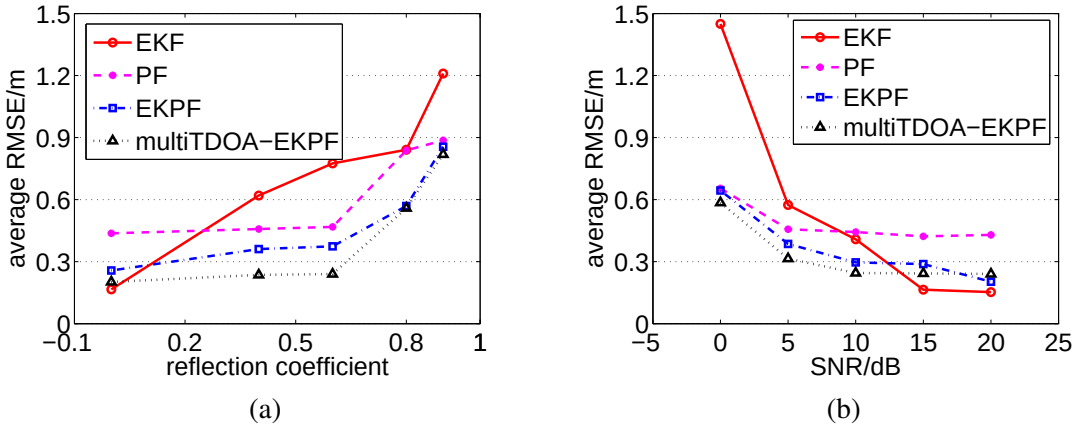


Figure 5.8: *Average RMSE under the different (a) reverberant environments; (b) noisy environments.*

EKF is able to catch up with the position of new source quickly. Further, all the methods are significantly deteriorated in the heavy reverberant and noisy environments.

5.6.4 Real room recording experiment

The complete real room recording environment is shown in Section 2.6.3, on page 45. The dimension of the lab is $8.1 \times 5.3 \times 3\text{m}^3$. Four microphone arrays each with five microphones are employed to receive the speech signals. 16 microphone pairs are thus available. The separation of each two adjacent microphones is 0.45m. Two sources move in the room with diagonal line trajectories: one is active from $(2.5,\ 1.5)$m to (6.0, 3.5)m, and the other follows from $(2.5,\ 3.5)$m to (6.0, 1.5)m. Detailed experiment setup can be found in Fig. 2.11 on page 46.

The whole signal is generated by cascading the received signal of source 2 to that of source 1. Examples of the received signals are presented in Fig. 4.26 on page 113. The TDOA estimation based on PHAT-GCC method is shown in Fig. 5.9. Since the ground truth is unknown, it is impossible to give exact probabilities of detection and false alarm, and thus only the TDOA measurement plots from some microphone pairs are presented here.

Figure 5.10 presents the tracking results of all the trackers listed at the beginning of this section for a single trial. Both the EKPF algorithms developed in this chapter track the sources accurately and lock on the position of the new source quickly. EKF fails to do so since large false alarm is presented for the TDOA measurements. Although the PF is able to track the

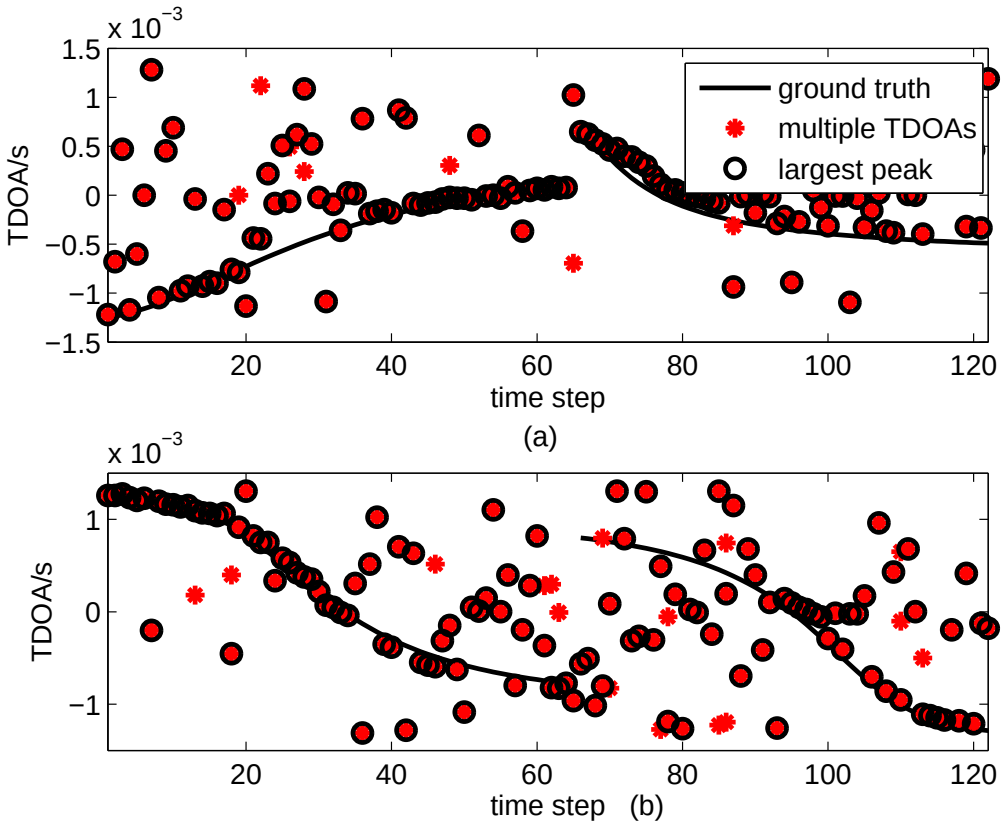


Figure 5.9: *TDOA measurement extracted from (a) microphone pair 4; and (b) microphone pair 14 in the real audio lab environment. Source 1 is active from time step 1 to 65, and then source 2 follows from time step 66 to 123.*

sources, it cannot catch up with the position of the new source quickly. Compared to the results from the single experiment of the simulated room environment in Section 5.6.3.1, the position estimation is worse. This is because the real room environment is more challenging. Also, the error from experiment system such as microphone positions would increase the tracking errors.

Figure 5.11 gives the tracking results over 100 Monte Carlo implementations. This statistical results further illustrate that the proposed EKPF algorithms are more accurate than the general SIR-PF and EKF in tracking the nonconcurrent multiple sources. The RMSE also presents a transition behavior: at those time steps where source switches, the RMSE increases sharply. The algorithms then converges to the position of the new source. However, the proposed EKPF approaches are able to find the position of the new source quickly, while the general SIR-PF generally needs much more time steps to lock on the position of the new source. The average RMSE is given in Table 5.8. Again, the errors from the experiment system, particular from the

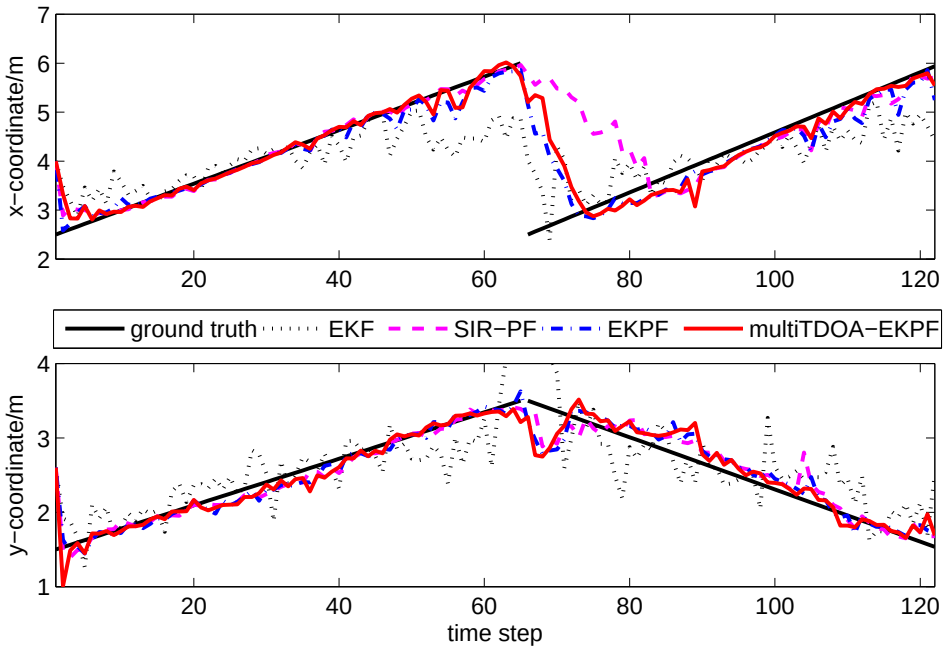


Figure 5.10: *Tracking results from a single trial in the real audio lab environment.*

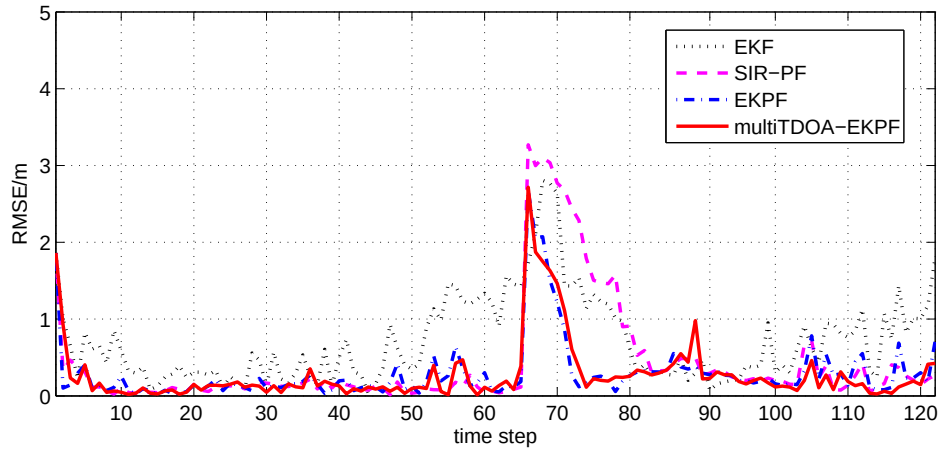


Figure 5.11: *RMSE over 100 Monte Carlo runs in the real audio lab environment.*

algorithm	EKF	SIR-PF	EKPF	multiTDOA-EKPF
RMSE	0.843	0.544	0.398	0.396

Table 5.8: *Average RMSE in the real audio lab environment.*

estimation of the ground truth of the source positions, increases the final RMSE.

5.7 Chapter summary

In this chapter, the problem of nonconcurrent multiple acoustic source tracking has been discussed. Two EKPF approaches are developed to track the nonconcurrent multiple sources: one employs a single TDOA (corresponding the maximum peak) to update the EKF step; the other uses multiple TDOAs (reverberant measurement model) to update the EKF step. Compared to the general SIR-PF, the EKPF approaches are able to sample the particles according to the optimal importance function. It is thus able to take the current measurements into account and lock on the position of the new source quickly. A single experiment in the reverberant environment is organised to illustrate the advantages of the EKPF approaches over the general SIR-PF and EKF. It shows that the tracking-lag problem of SIR-PF can be alleviated well by using the EKPF approaches.

The experiment results under different simulated noisy and reverberant environments further illustrate the tracking performance of EKPF approaches are better than general SIR-PF and EKF, especially in the moderate adverse environments. However, in the heavy reverberant and noisy environments, all these methods fail to report the source positions accurately mainly because the TDOA measurements are collapsed. More robust TDOA measurement extraction approaches should be developed under such cases.

The tracking approaches are also implemented in the real audio lab environment. The experiment presents very similar results: the proposed EKPF approaches are able to track the sources as well as to catch up with the position switch between the sources with a satisfactory accuracy.

Although the multiTDOA-EKPF algorithm introduces the amplitude information of each TDOA measurement as its prior, the amplitude information used here is simply based on its proportion among all the amplitudes of TDOAs, and how this amplitude information will affect the tracking results is not studied. Fig. 4.7 on page 86 shows that the RMS amplitudes from source and clutter are different. It is a great interest to thoroughly investigate the amplitudes from the sources and clutter, and then build probabilistic models for these amplitudes accordingly. The amplitude information can thus be incorporated in a full probabilistic sense.

Chapter 6
Unknown number of multiple acoustic source tracking

In Chapter 3 and Chapter 5, a Bayesian tracking system was derived for the single source tracking problem. Although the tracking system presented in Chapter 5 allows a presence of multiple sources, the sources are assumed to be nonconcurrently active. The tracking algorithm is, in fact, dealing with a single source at each time step. However, in a number of speech applications such as surveillance, multimedia conferencing, and selective speech enhancement, it requires the system to localise multiple simultaneously active sources. In such scenarios, the tracking algorithm should be able to track the positions of individual sources as well as to determine the number of sources. In contrast to the single source tracking problem which has been extensively studied, the problem of tracking multiple acoustic sources has received much less attention.

A multiple hypothesis tracking (MHT) based particle filtering approach is introduced and modified in this chapter to track the unknown and time-varying number of speakers. The tracker is built within a RFS multisource Bayesian filtering framework and incorporates a data association technique. The source state is constructed by the position set and an additional association variable. To reduce the estimation variance and sampling efficiently, a Rao-Blackwellisation technique is employed, by which the position states are marginalised by using an extended Kalman filtering, and only the data association variable is handled by the particle filtering. Rather than the traditional data association approaches where the heuristic techniques are used to prune or determine the hypothesis, the proposed particle filtering implementation theoretically allows random hypothesis-pruning.

6.1 Introduction

Various multisensor multitarget tracking techniques are introduced for the multiple speakers tracking problem in the past few years, such as interacting multiple model (IMM) [41, 70]

method and the random finite set multiple Bayesian filtering approach [1, 15, 16, 72]. The approaches in [41, 70] do not allow a time-varying number of sources, and only present the tracking results from single trials. In [1], Ma et al present an efficient algorithm to track multiple and time-varying number of sources and have fully analysed the experiment results. However, it focuses on introducing an updated tracking technique and lack of careful consideration of the room acoustic itself. Further, all these approaches are neither tested in a broad range of different noisy and reverberant environments nor in the real room experiment. In this chapter, a TDOA based multi-speaker tracking system is designed based on the following three requirements: 1) differentiation of the measurements from real sources or the clutter in the room environment; 2) enumerating the time-varying number of sources; and 3) filtering the corresponding source position estimates.

The tracking approach is based on random finite set framework with a Rao-Blackwellised particle filtering implementation. The RBPF is first proposed for a fixed number of multitargets in [117] and then extended to the unknown number of targets tracking scenario in [118]. In essence, the source position states which are assumed to follow a linear Gaussian model which can be integrated out in a closed form, and the PF is employed to handle the data association problem. As illustrated in Section 3.3 on page 63, such marginalisation of the source states can be regarded as using an infinite number of samples to replace the finite particle set in PF. The estimation variance can thus be reduced, and fewer particles are needed for the same accuracy. The PF implementation allows a random hypothesis-pruning by evaluating the importance of each association. The author in [119] fully illustrates this approach and implement it within a random finite set framework, in which the source dynamics are described by birth, survival, or death models, and a latent variable is incorporated to identify the data associations. Although some simulations are presented to illustrate the advantages of the tracking algorithm, none of them is implemented in the acoustic source tracking problem. Also, no statistical error measures are given for such tracking algorithms. It is thus difficult to compare the tracking performance with other techniques.

The following modifications are made for the tracking approach in [119] to construct the data association based RBPF method for a time-varying number of speakers.

1. The TDOA measurement function is linearised to form an EKF step. This step is the same as described in Section 5.3 on page 123, where the EKF is formulated to obtain the optimal importance distribution in the single source tracking case.

2. Since only one source is allowed to be born at each time step, the birth model is simply a prior probability based on experimental study, while the Poisson birth model is used in [119].

3. The source death is determined by modelling the expected track length (lifetime) using a Gamma distribution. Generally, the longer the source is not associated, the higher is the death probability. This model is the same as the death process in [118]. A death model which is symmetrical to the birth model is employed in [119].

4. Different error measures are introduced in this chapter to fully analyse the tracking performance.

Other than the state distribution, the likelihood used to evaluate the importance of different association hypotheses is also obtained by the EKF. By modeling the expected track length of the source, the modified tracking algorithm is able to keep the source in the scene even it is not associated with any measurement at a short interval. This matches the TDOA based tracking scenario in that multisource TDOAs can rarely be obtained at all the microphone pairs when sources are simultaneously active in the reverberant environment. Usually, the reliable TDOAs can only be obtained from the microphone pairs which are close enough to the sources, where both the high SNR and SRR can be achieved. It is worth pointing out that the number of simultaneously active speakers is assumed small so that the computation is affordable by employing this MHT based approach.

The main differences between the tracking system developed in this chapter and the existing RFS Bayesian filtering for acoustic source tracking in [1] are that the data association and Rao-Blackwellisation step are used here. The tracking approach in [1] is depicted in the Fig. 6.1(a), in which all the measurements at a time step are used to update the states, and no data association technique is operated. The states are estimated by using the particle filtering. Fig. 6.1(b) illustrates the processing model of our tracking system. For each measurement, its relationship with the source or clutter is evaluated. At each time step, the measurements are thus processed one after another, and only the measurement which is associated to the source are used to update the state. Further, the states and the likelihood of a source in our approach are obtained by using the EKF, while in [1], the likelihood is the Gaussian distribution described in equation (5.30) on page 128 and the states are estimated by the particle filtering. The differences between our approach and RFS bayesian filtering are summarised in Table 6.1.

	RFS particle filtering	RFS RBPF
measurement model	reverberant measurement model	singleton measurement
likelihood (from source)	Gaussian distribution	EKF
state estimation	PF filtering	Rao-Blackwellisation PF

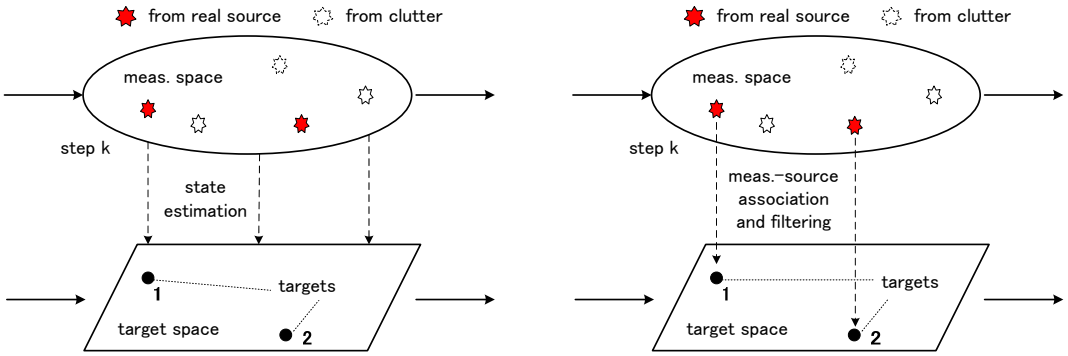
Table 6.1: *Differences between RFS particle filtering and RFS Rao-Blackwellised particle filtering (our approach).*

When tracking multiple simultaneously active sources, extracting the TDOAs for all the sources are almost impossible by using the traditional PHAT-GCC method. Experiments in Section 4.5 show that even in a moderate reverberant environment, the probability of detection is very low, and the probability of false alarm can be very high if a small threshold is picked. As illustrated in Section 4.4, the DUET-GCC method is able to estimate the TDOAs of multiple sources since the sources are separated in the time-frequency domain first, and the PHAT-GCC method is then applied to the spectrogram of each source individually to extract the TDOA measurements. Experiments in Section 4.4 also showed that DUET-GCC performs better than traditional PHAT-GCC method under such a scenario. In this chapter, both the TDOA measurements extracted from DUET-GCC and traditional PHAT-GCC will be used to track the sources, and the final tracking performance will be fully analysed. Similar as in Chapter 5, a real audio lab experiment is also organised to evaluate the tracking performance.

The rest sections are organised as follows. The RFS measurement and multiple source dynamical models are presented in Section 6.2. The Rao-Blackwellisation formulation is illustrated in Section 6.3. The source dynamics: the birth, death and survival processes are given in Section 6.4. The complete tracking algorithm and performance evaluation measures are summarised in Section 6.5. The proposed tracking approach is implemented in different simulated noisy and reverberant environments as well as in the real room environment in Section 6.6. Finally, some conclusions are drawn in the last section. Note that part of the work in this chapter is already published in [120].

6.2 RFS acoustic source tracking models

The RFS has been shown an efficient framework for multiple source tracking since it naturally depicts the randomness of the source number as well as source positions. In this section, the RFS multiple acoustic source tracking framework is firstly introduced. The basic concepts of



(a) (b)

Figure 6.1: *(a) All the measurements are used to update the source states without a data association technique in RFS particle filtering approach [1]; (b) Measurement-source associations are operated and the states are estimated from the corresponding measurements in RFS RBPF tracking system.*

data association are also presented.

6.2.1 Measurement model

The measurement set $\mathcal{Z}_k$ collected at time step k is the same as presented in Section 5.2.3, on page 122. It is formulated by the source-generated measurements and the false alarms. To simplify the expression, all the measurements are included in a set and the index of microphone pair ℓ can be ignored. The measurement state can be stated as

$$\begin{aligned} \underbrace{\mathcal{Z}_k}_{\text{measurement set}} &= \bigcup_{\ell=1}^{L} \{\hat{\tau}_{1,k}^{\ell}, \ldots, \hat{\tau}_{n_k^{\ell},k}^{\ell}\} \\ &= \underbrace{\{\tilde{\tau}_{1,k}, \ldots, \tilde{\tau}_{\tilde{n}_k,k}\}}_{\text{source generated}} \cup \underbrace{\{\bar{\tau}_{1,k}, \ldots, \bar{\tau}_{\bar{n}_k,k}\}}_{\text{false alarms}}, \end{aligned} \tag{6.1}$$

where $\tilde{n}_k$ and $\bar{n}_k$ represent the number of source generated measurements and the number of false alarms respectively. The cardinality of the measurement set is thus $N_k = \sum_{\ell} n_k^{\ell} = \tilde{n}_k + \bar{n}_k$. Since in our tracking system, the data association is considered, each measurement is associated either with a source or with a clutter. At each processing, the measurement set is thus a singleton, given as

$$z_k = \{\hat{\tau}_{n,k}\}, \tag{6.2}$$

for all $n = 1, \ldots, N_k$. The measurement set from all the microphone pairs are thus processed one after another sequentially (with an arbitrary order). The source dynamic model will only be operated when the time step k changes. This means that the association and state filtering will be manipulated for each measurement for N_k times. Particularly, the expression $z_{1:k}$ will refer to all the measurements from start to current processing, and $z_{1:k-1}$ corresponds to all the measurements before current processing, i.e., all the measurements from time step 1 to time step $k-1$. The Bayesian network of the measurement model is illustrated in Fig. 6.2(a). The likelihood density of the measurements conditions on the source presence $\mathcal{X}$ as well as different data associations.

The same as the single source tracking scenario described in Section 5.2.3 on page 122, several local peaks of the GCC function are often picked to fully represent the potential TDOAs as well as to handle the reverberation in the existing multiple acoustic source tracking approaches [1, 15, 16, 72]. In our tracking system, only the peaks with relatively large values (threshold $R_{TH} > 0.9$), as described in Table 4.5 on page 109) and the highest peak are picked from each microphone pair; since these peaks are more likely generated by the source with higher SRR rather than by the clutter due to the reverberations. Reliable TDOA measurements can thus be obtained with limited false alarms. The cost of this measurement extraction is that the probability of detection will be low, especially when multiple sources are simultaneously active. For the DUET-GCC method, since the TDOA from clutter can rarely be clustered, the false alarm rate is generally smaller. In Section 6.4, a source survival/death process will be fully derived to deal with the problem of detection missing.

6.2.2 Multiple source dynamical model

The source dynamic is relatively easy to model for the single source case, e.g., by a random walk model [121] or the Langevin model [6], since one and only one source is assumed to be active and only the modeling of source trajectory is needed. For the multiple source tracking, more complicated dynamic models should be incorporated due to the uncertainty of the source appearance/disappearance. Three categories of the source behaviors are considered:

- source survival;
- new born source;
- source dies.

Any kind of source dynamics can thus be modelled by formulating a combination of these three behaviors. [1]

Source survival: no source birth or death

Suppose that there are no source births or deaths in the scene. Since the sources are statistically independent, each source motion can be assumed to follow the Langevin model introduced by equation (5.5) on page 121 in Section 5.2.2. For the tracking problem, the motion model is an important factor that affects the tracking performance. Since a simple trajectory and slow-paced movements are assumed in our experiments, the standard Langevin model is simply employed. Let $\mathcal{X}_{k-1} = \{\mathbf{x}_{1,k-1}, \ldots, \mathbf{x}_{M_{k-1},k-1}\}$ represent the state vector at time step $k-1$. The predicted source state set can be written as

$$\underbrace{\mathcal{X}_{k|k-1}}_{\text{predicted state set}} = \underbrace{\{\mathbf{x}_{1,k|k-1}, \ldots, \mathbf{x}_{M_{k-1},k|k-1}\}}_{\text{predicted state}}, \tag{6.3}$$

in which each element $\mathbf{x}_{m,k|k-1}$ is evolved following the Langevin model, given as

$$\mathbf{x}_{m,k|k-1} = \mathbf{A}_k \mathbf{x}_{m,k-1} + \mathbf{Q}_k \mathbf{v}_k. \tag{6.4}$$

Since there is no source birth or death, the number of the sources remains M_{k-1}, i.e., $|\mathcal{X}_{k|k-1}| = |\mathcal{X}_{k-1}|$. The sources are assumed to be independent of each other. The transition density can thus be written as

$$\begin{aligned} p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{X}_{k-1}) &= \prod_{m=1}^{M_{k-1}} p(\mathbf{x}_{m,k|k-1}|\mathbf{x}_{m,k-1}) \\ &= \prod_{m=1}^{M_{k-1}} \mathcal{N}(\mathbf{x}_{m,k|k-1}; \mathbf{A}_k \mathbf{x}_{m,k-1}, \mathbf{Q}_k), \end{aligned} \tag{6.5}$$

New source born

If new sources appear, the predicted states can be formulated as

$$\underbrace{\mathcal{X}_{k|k-1}}_{\text{predicted state set}} = \underbrace{\bar{\mathcal{X}}_{k|k-1}}_{\text{predicted survival state set}} \cup \underbrace{\mathcal{B}_k}_{\text{new state set}}, \tag{6.6}$$

[1] In the following chapters, the terms "source birth" and "source death" are used to denote the source appearance and the source disappearance respectively. Sources "die" if they disappear or become nonactive from the scene, and are "born" if they appear or become active in the field of view.

where $\bar{\mathcal{X}}_{k|k-1}$ is the predicted state set of $\mathcal{X}_{k-1}$, and $\mathcal{B}_k$ is the set of new born sources. $\cup$ denotes the union of the set which simply adds the new states to the original set. Suppose that new states are initialised at $\mathbf{x}_0$, and μ is the number of the new born sources, the expression (6.6) can be written as

$$\underbrace{\mathcal{X}_{k|k-1}}_{\text{predicted state set}} = \underbrace{\{\mathbf{x}_{1,k|k-1}, \ldots, \mathbf{x}_{M_{k-1},k|k-1}\}}_{\text{predicted survival states}} \cup \underbrace{\{\mathbf{x}_{1,0}, \ldots, \mathbf{x}_{\mu,0}\}}_{\text{new states}}. \tag{6.7}$$

The total number of the sources at time step k is thus $M_k = M_{k-1} + \mu$ ($\mu = 1$ in this chapter since only one new born source at each time step is allowed). The transition probability can be derived as

$$p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{X}_{k-1}) = \sum_{\mathcal{B}_k \subseteq \mathcal{X}_{k|k-1}} p_b(\mathcal{B}_k) p_{k|k-1}(\mathcal{X}_{k|k-1} \setminus \mathcal{B}_k|\mathcal{X}_{k-1}), \tag{6.8}$$

where $p_b(\cdot)$ represent the density for the new born source, and $\setminus$ represents set minus. Here $\mathcal{X}_{k|k-1} \setminus \mathcal{B}_k = \bar{\mathcal{X}}_{k|k-1}$. Given a birth prior P_b, the density for the new born sources can be written as

$$p_b(\mathcal{B}_k) = \begin{cases} 1 - P_\mathrm{b}, & \mathcal{B}_k = \emptyset; \\ P_\mathrm{b} p(\mathbf{x}_0), & \mathcal{B}_k = \{\mathbf{x}_0\}; \\ 0, & \text{otherwise.} \end{cases} \tag{6.9}$$

with $p(\mathbf{x}_0)$ denoting an initial state distribution.

Source death

If the sources disappear, then

$$\underbrace{\mathcal{X}_{k|k-1}}_{\text{predicted state set}} = \underbrace{\{\mathbf{x}_{1,k|k-1}, \ldots, \mathbf{x}_{M_{k-1},k|k-1}\}}_{\text{predicted states}} \setminus \underbrace{\mathcal{D}_k}_{\text{death sources}}, \tag{6.10}$$

where $\mathcal{D}_k$ represents the set of death sources, given as

$$\mathcal{D}_k = \{\mathbf{x}_{m_1,k|k-1}, \ldots, \mathbf{x}_{m_\nu,k|k-1}\}, \tag{6.11}$$

where ν is the number of the disappearing sources. For a dead source, just simply set the corresponding state as an empty set, that is

$$\mathbf{x}_{m,k|k-1} = \emptyset; \quad \forall : m = m_j,\ j = 1, \ldots, \nu. \tag{6.12}$$

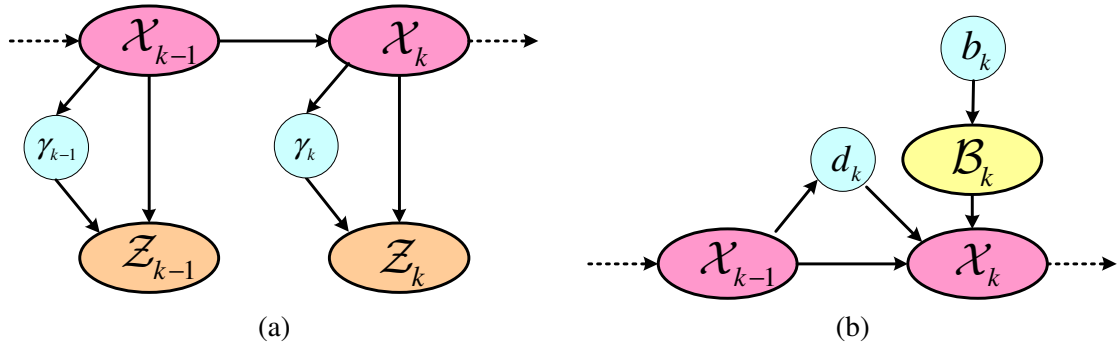


Figure 6.2: *Bayesian network representing of (a) the measurement model; (b) the source dynamic model. The measurement is associated with a source according to the association indicator $\boldsymbol{\gamma}$, and the source birth and death are determined by the birth and death indicator $\mathbf{b}$ and $\mathbf{d}$ respectively.*

Given a death probability P_{d}, the corresponding state transition density is thus

$$p_{k|k-1}(\mathcal{X}_{k|k-1}|\mathcal{X}_{k-1}) = \left\{ \prod_{\forall:\, \{\mathbf{x}_{m,k-1}\} \not\subseteq \mathcal{X}_{k|k-1}} P_{\mathrm{d}} \right\} \left\{ \prod_{\forall:\, \{\mathbf{x}_{m,k-1}\} \subseteq \mathcal{X}_{k|k-1}} (1 - P_{\mathrm{d}}) p(\mathbf{x}_{m,k|k-1}|\mathbf{x}_{m,k-1}) \right\}, \tag{6.13}$$

where the first product is taken for all the death sources, and the second is for all the survival sources. The complete Bayesian network of multiple source dynamical model is illustrated in Fig. 6.2(b). The detailed discussion about the birth prior P_{b}, the death probability P_{d} and the transition priors will be presented in Section 6.4.

6.2.3 Tracking via data association

The assignment process between the sources/clutter and the measurements is called data association. The main difficulty in multiple source tracking is to associate the measurements with the sources or the clutter correctly. To clarify our technique, the basic concepts about the data association problem is introduced in this section.

Assuming that there are M_k sources the kth time step, i.e. $|\mathcal{X}_k| = M_k$. For each singleton measurement set z_k, the association hypothesis $\boldsymbol{\gamma}_k$ is defined as

$$\boldsymbol{\gamma}_k : z_k \mapsto \{0, 1, \ldots, M_k, M_k + 1\}, \qquad \forall : z_k \subseteq \mathcal{Z}_k, \tag{6.14}$$

and

- $\gamma_k = 0$ means that the nth measurement z_k is associated to a false alarm;
- $\gamma_k = m$, for $m = 1, \ldots, M_k$ denotes that the nth measurement z_k is associated to source m;
- $\gamma_k = M_k + 1$ denotes that the nth measurement z_k is associated to a new born source $M_k + 1$;

Note that each measurement is able to have one association only. Fig. 6.3 illustrates the association hypothesis between the all the measurements collected at time step k and the sources. The task for the tracking algorithm is to associate the measurements with the sources correctly and filter the source states.

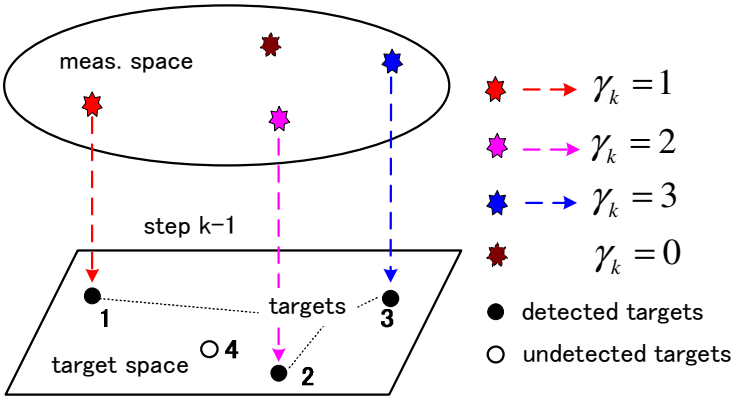


Figure 6.3: *Illustration of the association hypotheses. Each measurement is able to be associated with either a source or clutter. The task for the tracking algorithm is to find the correct association and filter the source states accordingly.*

6.3 Rao-Blackwellisation formulation

Due to the presence of multiple sources, the state to be estimated is no longer simply the position of a single source, but a set of positions as well as the number of sources. This leads a high dimensional source state, and accurate estimation is very difficult to achieve. Following the Rao-Blackwellisation theory introduced in Section 3.3 on page 63, a Rao-Blackwellisation formulation for the multiple source tracking problem is presented in this section. The source

position states are marginalised out by using an EKF, and only the association variable is estimated by the particle filtering. The derivation of the association priors, general likelihood and optimal importance function will also be presented.

6.3.1 Rao-Blackwellised formulation

Supposing that $\boldsymbol{b}_k$ and $\boldsymbol{d}_k$ are the variables indicating the birth and the death processes of the source respectively, the dynamics of the source can be fully represented by an extension of the predefined association variable $\boldsymbol{\gamma}_k$, given as

$$\boldsymbol{\theta}_k = (\boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k). \tag{6.15}$$

Here $\boldsymbol{\theta}_k$ can be regarded as an association variable which encapsulates the survive, birth, and death processes. $\boldsymbol{b}_k$ and $\boldsymbol{d}_k$ indicate the birth and death processes respectively, with value 1 denoting that the birth and death happen and 0 otherwise. The original state $\mathcal{X}_k$ to be estimated is thus extended to

$$(\mathcal{X}_k, \boldsymbol{\theta}_k) = (\{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{M_k,k}\}, \boldsymbol{\theta}_k). \tag{6.16}$$

Given a single measurement z_k to be processed at the time step k, our interest here is to estimate the joint posterior distribution $p(\mathcal{X}_k, \boldsymbol{\theta}_k | z_{1:k})$, which can be decomposed to the conditional source distribution $p(\mathcal{X}_k | \boldsymbol{\theta}_k, z_{1:k})$ and the association posterior density $p(\boldsymbol{\theta}_k | z_{1:k})$, given by

$$p(\mathcal{X}_k, \boldsymbol{\theta}_k | z_{1:k}) = \underbrace{p(\mathcal{X}_k | \boldsymbol{\theta}_k, z_{1:k})}_{\text{EKF approximation}} \underbrace{p(\boldsymbol{\theta}_k | z_{1:k})}_{\text{PF}}. \tag{6.17}$$

Conditional on $\boldsymbol{\theta}_k$, the position states $p(\mathcal{X}_k | \boldsymbol{\theta}_k, z_{1:k})$ can be estimated by applying an EKF as described in Section 5.3, and only the latent variable $\boldsymbol{\theta}_k$ is needed to be handled by a particle filtering. The association posterior density $p(\boldsymbol{\theta}_k | z_{1:k})$ is approximated by the Monte Carlo simulation, given as

$$p(\boldsymbol{\theta}_k | z_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k), \tag{6.18}$$

where $\delta(\cdot)$ is a dirac function only with values when $\boldsymbol{\theta}_k = \boldsymbol{\theta}_k^{(i)}$, and N denotes the number of particles. Following the RBPF expression (3.69) on page 68, the posterior distribution can be

obtained by

$$p(\mathcal{X}_k, \boldsymbol{\theta}_k | z_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k) p(\mathcal{X}_k^{(i)} | \boldsymbol{\theta}_k^{(i)}, z_{1:k}), \tag{6.19}$$

where $p(\mathcal{X}_k, \boldsymbol{\theta}_k | z_k)$ is a multi-modality Gaussian distribution.

The formulation of the importance weight is also the same as described in 3.3.2, on page 67. Suppose that the importance distribution of the association hypothesis variable $\boldsymbol{\theta}_k$ is

$$\boldsymbol{\theta}_k^{(i)} \sim q(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k}), \tag{6.20}$$

the weight of Rao-Blackwellised particle filter can be updated as

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k | \boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) p(\boldsymbol{\theta}_k^{(i)} | \boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})}{q(\boldsymbol{\theta}_k^{(i)} | \boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k})}, \tag{6.21}$$

where $p(z_k | \boldsymbol{\theta}_k^{(i)}, z_{1:k-1})$ is the hypothesis likelihood, and $p(\boldsymbol{\theta}_k^{(i)} | \boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})$ is the prior of the hypothesis. There is no $\mathcal{X}$ term in this expression since the source state is marginalized by an EKF.

The top-level procedure of RBPF can be found in Algorithm 9. By using a Rao-Blackwellised step, the source states are estimated by the EKF, and the hypotheses are sampled and replicated/discarded according to the high/low importance weight if necessary.

6.3.2 Association priors

To calculate the prior of the hypothesis variable $p(\boldsymbol{\theta}_k^{(i)} | \boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})$, the relation between the birth, survive, and death process should be clarified. As described in Fig. 6.2(b), a source is born with a prior birth probability, and is independent with any of the existing sources. Generally, the probability of source death is dependent only on its previous existence. The prior of the association indicator is dependent only on the number of sources based on the assumption at current time step k. The prior of the association variable can thus be written as

$$p(\boldsymbol{\theta}_k^{(i)} | \boldsymbol{\theta}_{1:k-1}^{(i)}, z_{1:k-1}) = p(\boldsymbol{\gamma}_k^{(i)} | \boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)}) p(\boldsymbol{d}_k^{(i)} | \boldsymbol{d}_{k-1}^{(i)}) p(\boldsymbol{b}_k^{(i)}), \tag{6.22}$$

where $p(\boldsymbol{b}_k^{(i)})$ and $p(\boldsymbol{d}_k^{(i)} | \boldsymbol{d}_{k-1}^{(i)})$ are the prior density of the new born source and a death process, which correspond to the probability of birth P_b in equation (6.9) and the probability of death P_d

Algorithm 9: Top-level procedure of the RBPF.

Input: Extracted TDOA measurements.
Output: Multiple sources positions.

Initialisation: $w_0^{(1:N)} \leftarrow 1/N$; $\mathcal{X}_0^{(1:N)} \leftarrow \emptyset$.
Over all the time step:
for $k \leftarrow 1$ **to** K **do**
  Predict the state $\hat{\mathcal{X}}_{k|k-1}^{1:N}$according to the Langevin motion model for all the particles.
  Over all the measurements at time step k:
  for $n \leftarrow 1$ **to** N_k **do**
    Over all the particles:
    for $i \leftarrow 1$ **to** N **do**
      - generate different hypothesis $\boldsymbol{\theta}_k^{(i)}$;
      - evaluate the importance function $w_k^{(i)}$ according to equation (6.21); see Algorithm 13 for details.
    end
    Weight normalization: $w_k^{(i)} = w_k^{(i)} / \sum_{i=1}^{N} w_k^{(i)}$. $\quad 1 \leq i \leq N$
  end
  Output the estimates.
  Resample ($\mathcal{X}_k$, w_k) if necessary.
end

in equation (6.13) respectively. $p(\boldsymbol{\gamma}_k^{(i)}|\boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)})$ is the prior probability of the association indicator, defined as

$$p(\boldsymbol{\gamma}_k^{(i)} = \gamma|\boldsymbol{b}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{\gamma}_{k-1}^{(i)}) = \begin{cases} p_f, & \gamma = 0; \\ \frac{1-p_f}{M_k}, & \gamma = m; \\ 0, & \text{otherwise.} \end{cases} \tag{6.23}$$

where p_f is the prior probability of false alarm, and the source number M_k is defined as

$$M_k = \begin{cases} |\mathcal{X}_k \cup \mathcal{B}_k| = M_{k-1} + 1, & \text{birth happens;} \\ |\mathcal{X}_k \setminus \mathcal{D}_k| = M_{k-1} - 1, & \text{death happens;} \\ |\mathcal{X}_k| = M_{k-1}, & \text{otherwise.} \end{cases} \tag{6.24}$$

where $|\cdot|$ stands for the cardinality. Since the probability of false alarm is p_f, the probability for all the sources is $1 - p_f$. To keep the summation of the association prior to be unity, a reasonable choice for setting the probability for each source is thus to distribute the probability of all the sources equally, i.e., $(1 - p_f)/M_k$.

6.3.3 Likelihood function

The general expression of the association likelihood density can be written as

$$p(z_k|\mathcal{X}_{k|k-1}) = p_f p(z_k|\gamma_k = 0) + \sum_{\gamma} \frac{1-p_f}{M_k} p(z_k|\gamma_k = m), \tag{6.25}$$

where $p(z_k|\gamma_k = 0)$ and $p(z_k|\gamma_k = m)$ are the likelihood for the false alarm and the source generated measurement respectively. Fig. 6.2(a) illustrates the Bayesian network of the detailed measurement model. The measurements are associated with either a source or a false alarm by the indicator $\boldsymbol{\gamma}_k$, and only the measurements associated the real sources are used to update the corresponding states.

Due to the reverberation and the interference among source signals themselves, some measurements can be false alarms. The same as described in Section 5.4.1 on page 127, a uniform distribution over the possible TDOA interval is given in the case that the measurement is a false alarm,

$$p(z_k|\gamma_k = 0) = \mathcal{U}_{[-\tau_{\max},\tau_{\max}]}(z_k) = \frac{1}{2\tau_{\max}}, \tag{6.26}$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay which can only happen when the microphone pair and the source lie exactly on a line.

If the measurement z_k is generated by the mth target, it follows the nonlinear relationship with the source position $\mathbf{x}_{m,k}$ as described in equation (5.2) on page 119, rewritten as

$$z_k = \frac{\|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,2}\|}{c}. \tag{6.27}$$

Since this measurement model is nonlinear, the extended Kalman filter is employed here to evaluate the likelihood, given by

$$\begin{aligned} p(z_k|\gamma_k = m) &= \text{EKF}(z_k; \mathbf{H}_{m,k}\mathbf{x}_{m,k}, \mathbf{R}_{m,k}) \\ &= \mathcal{N}(z_k; \tau_k^{\ell}(\mathbf{x}_{m,k}), \mathbf{S}_k), \end{aligned} \tag{6.28}$$

where $\tau_k^{\ell}(\mathbf{x}_{m,k})$ is the EKF measurement prediction, and $\mathbf{S}_k$ is given by equation (5.23a), on page 125. EKF$(\cdot)$ denotes the implementation of EKF, in which the derivation of the EKF model matrix $\mathbf{H}_{m,k}$ and $\mathbf{R}_{m,k}$ and its particle filtering implementation can be found in Section 5.3. Further, if the measurement is associated with a new born source, the same EKF implementation

will apply, given as

$$\begin{aligned} p(z_k|\gamma_k = M_{k-1}+1) &= \text{EKF}(z_k; \mathbf{H}_{0,k}\mathbf{x}_{0,k}, \mathbf{R}_{0,k}) \\ &= \mathcal{N}(z_k; \tau_k^{\ell}(\mathbf{x}_{0,k}), \mathbf{S}_k), \end{aligned} \tag{6.29}$$

where $\mathbf{x}_{0,k}$ is an initial state given for all the new born source, and $\mathbf{H}_{0,k}$ and $\mathbf{R}_{0,k}$ can be obtained by simply taking the deviation at $\mathbf{x}_{0,k}$. Apart from the likelihood, the EKF step here also provides the filtered position state distribution if the measurement is associated to a source.

6.3.4 Optimal importance function

As mentioned in Section 3.2.4, on page 62, the performance of particle filtering is highly depending on the design of the important function. The optimal importance function [85], $q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k}) = p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k})$, which has been proved able to minimize the variance of the importance weight $w_k^{(i)}$ conditional upon the previous states $\mathbf{x}_{1:k-1}^{(i)}$ and the measurements $z_{1:k}$, is employed here. Again, in the RBPF implementation, since the position states $\mathcal{X}_k$ is marginalized out by the EKF, the measurement is only conditional on the hypothesis variable $\boldsymbol{\theta}_k^{(i)}$. The optimal importance distribution can be stated as

$$\begin{aligned} \boldsymbol{\theta}_k^{(i)} &\sim q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k}) \\ &= \frac{p(z_k|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1})p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})}{p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})}, \end{aligned} \tag{6.30}$$

in which the hypothesis prior $p(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1})$ can be calculated according to equation (6.22). The calculations of the denominator term and the likelihood $p(z_k|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1})$ are as follows.

The denominator term in equation (6.30) is a proportional constant which can be calculated as

$$p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1}) = \int p(z_k, \boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1}) d\boldsymbol{\theta}_k. \tag{6.31}$$

The association variable $\boldsymbol{\theta}_k$ can be decomposed to separated terms and computed as

$$\begin{aligned} p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1}) &= \int\int\int p(z_k, \boldsymbol{\gamma}_k, \boldsymbol{d}_k, \boldsymbol{b}_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1}) d\boldsymbol{\gamma}_k d\boldsymbol{d}_k d\boldsymbol{b}_k \\ &= \sum_{\boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k} p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, \boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k, z_{1:k-1}) \\ &\quad p(\boldsymbol{\gamma}_k|\boldsymbol{d}_k, \boldsymbol{b}_k, \boldsymbol{\theta}_{k-1}^{(i)}) p(\boldsymbol{d}_k|\boldsymbol{d}_{k-1}^{(i)}) p(\boldsymbol{b}_k), \end{aligned} \tag{6.32}$$

where $p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, \boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k, z_{1:k-1})$ is actually the likelihood based on different hypotheses $(\boldsymbol{\gamma}_k, \boldsymbol{b}_k, \boldsymbol{d}_k)$. Given a hypothesis $(\gamma_k^{(i)}, b_k^{(i)}, d_k^{(i)})$, $p(z_k|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1})$ can be calculated as

$$\begin{aligned} p(z_k|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) &= \int p(z_k, \mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) \delta\mathcal{X} \\ &= \int p(z_k|\boldsymbol{\theta}_k^{(i)}, \mathcal{X}) p(\mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) \delta\mathcal{X}. \end{aligned} \tag{6.33}$$

Since the measurement only has a relationship with the state if it is associated with a source, above expression can be simplified as

$$p(z_k|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) = \int p(z_k|\gamma_k^{(i)}, \mathcal{X}) p(\mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) \delta\mathcal{X}. \tag{6.34}$$

If the measurement is associated with a clutter, i.e., $\gamma_k^{(i)} = 0$, $p(z_k|\gamma_k^{(i)}, \mathcal{X})$ follows the distribution described by equation (6.26). The above expression (6.33) becomes

$$p(z_k|\gamma_k^{(i)} = 0) \int p(\mathcal{X}|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) \delta\mathcal{X} = \frac{1}{2\tau_{\max}}. \tag{6.35}$$

In the case that the measurement is associated with a source, i.e., $\gamma_k^{(i)} = m \geq 0$, for $m = 1, \ldots, M_k$, the integral in (6.33) can be written as

$$\int p(z_k|\gamma_k^{(i)} = m, \mathbf{x}_{m,k}) p(\mathbf{x}_{m,k}|\boldsymbol{\theta}_k^{(i)}, z_{1:k-1}) d\mathbf{x}_{m,k} = p(z_k|\gamma_k^{(i)} = m). \tag{6.36}$$

It thus turns out the EKF likelihood computed by using equation (6.28) or equation (6.29). This means when formulating the denominator term, the likelihood is also calculated, and the calculation of the integration in equation (6.31) is simply the summation of the probabilities of all the hypotheses.

Substituting the optimal importance function (6.30) into the weight updating equation (6.21),

we can get the new expression of the weight updating,

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(z_k|\boldsymbol{\theta}_{k-1}^{(i)}, z_{1:k-1}). \tag{6.37}$$

The complete algorithm to compute this optimal importance function will be summarised in Section 6.5 where the tracking algorithm is formulated. The difficulties now become modelling the birth and death processes. In the next Section, detailed illustration of designing the birth and death processes will be presented.

6.4 Source dynamic models

The purpose of this section is to model different source dynamics. Since a time-varying number of the sources is allowed as mentioned in section 6.2.2, on page 152, the source dynamics can be depicted as birth, death, and survival (no birth and death happens) processes. The detailed models are given as follows.

6.4.1 Birth process

Assuming that a birth happens with a probability of P_b, and the birth happens independent with any existing sources, the birth process can be formulated as equation (6.9). For a new born source, an initial state $\mathbf{x}_{0,k}$ is given, where $\mathbf{x}_{0,k}$ is assumed to follow a Gaussian distribution with the state mean $\mathbf{m}_0$ and covariance matrix $\mathbf{P}_0$ respectively. After a birth process, the new state can be expressed as

$$\mathcal{X}_{k|k-1} = \bar{\mathcal{X}}_{k|k-1} \cup \{\mathbf{x}_{0,k}\}. \tag{6.38}$$

In practice, P_b is not known, and its value is usually determined by experimental study. Generally speaking, increasing the value of the birth probability P_b is expected to enhance the discovering of the new source. However, an overly large value may increase the risk of overestimation of the source number. The algorithm for the birth process is summarised in Algorithm 10.

Algorithm 10: Form the birth process.

// At time step k for ith particle. The number of sources: $|\mathcal{X}_{k-1}^{(i)}| = M$.
Input: Initialisation state and variance matrix $(\mathbf{x}_0, \mathbf{P}_0)$; measurement z_k.
Output: $(p_{\mathrm{b}}, \mathbf{x}_{M+1,k}, \mathrm{l}_{\mathrm{b}})$
if *birth happens* **then**
  - $p_{\mathrm{b}}(1) \leftarrow P_{\mathrm{b}}$; $\mathcal{B}_k \leftarrow \{\mathbf{x}_0\}$;
  - operate the EKF for the initialisation state: // l_b *represents the likelihood*
    $(\mathbf{x}_{M+1,k}, \mathrm{l}_{\mathrm{b}}(1)) \leftarrow \mathrm{EKF}(z_k; \mathbf{H}_{0,k}\mathbf{x}_0, \mathbf{P}_0)$.
else
  - $p_{\mathrm{b}}(0) \leftarrow 1 - P_{\mathrm{b}}$; $\mathrm{l}_{\mathrm{b}}(0) \leftarrow 1$.
end

6.4.2 Death process

After assigning the measurement to an existing source, the track of the source will be existed until a death happens. Suppose that the expected length of the source existence (track) is T_m, which denotes the time period from the last source-measurement association. The conditional probability $P(T_m)$ for the expected track length T_m will be modelled. The exponential density of the track length is introduced in [122] and has been found to match the experimental data closely for targets crossing through a fixed surveillance area. In this work, the track length of source is modeled by a gamma distribution [118]. The gamma probability density function is widely used in reliability models of lifetimes, and is more flexible than the exponential distribution in that it can be regarded as a summation of multiple exponential distributions and can be used to model the variables that seem to be highly skewed.

Suppose that T is the length of a frame, and the sampling frequency is f_s. The time interval between two consecutive frames is thus

$$t_0 = \frac{T}{f_s}. \tag{6.39}$$

where t_0 is real time in seconds. Further assume that k_m is the last frame that the source m is associated, and the corresponding time stamp is $t_m = k_m t_0$. The source is still existing at the previous time step t_{k-1} (the corresponding time stamp $t_{k-1} = (k-1)t_0$), which means that during the period $\Delta t_m = t_{k-1} - t_m$ the source is not associated but remains in the scene. Given the condition $T_m \geq t_{k-1} - t_m$, we are interested in the probability that the source is dead at current time step t_k, with $t_k = t_{k-1} + t_0$. Fig. 6.4 gives an illustration of the expected track length and the evaluation of the probability of death. The probability of the expected track

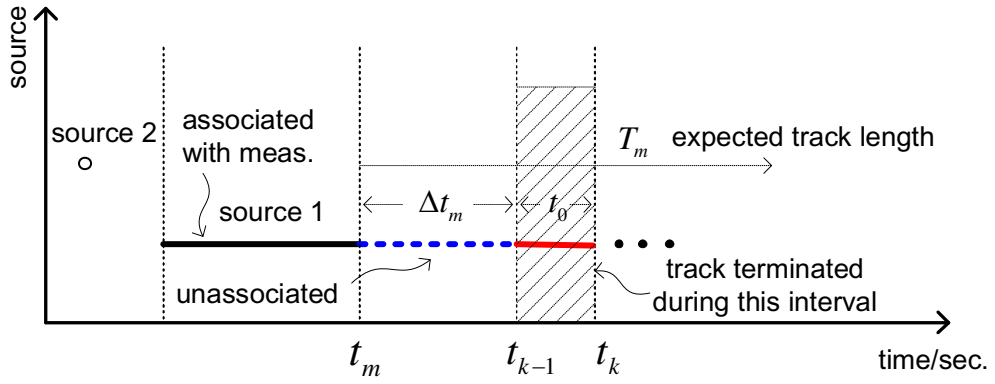


Figure 6.4: *Illustration of the expected track length T_m. The last time the source 1 is associated with a measurement is t_m, and during the past period $\Delta t_m = t_{k-1} - t_m$ the source is still active but not associated. The probability that the source 1 is dead at the current time t_k is the probability that the expected track length T_m terminates during the time interval $[t_{k-1}\ t_k]$.*

length of mth source follows a gamma distribution, given as

$$T_m \sim \mathcal{G}(T_m|\alpha,\beta) = {T_m}^{(\alpha-1)} \frac{\beta^\alpha e^{\frac{-T_m}{\beta}}}{\Gamma(\alpha)}, \tag{6.40}$$

where $\mathcal{G}(\cdot|\alpha,\beta)$ denotes the gamma distribution, and α and β are the shape parameter and scale parameter respectively. Fig. 6.5(a) plots the pdf of the expected track length (Gamma probability density functions) under several different parameter pairs (α,β). The probability that the source is dead at current time step t_k is thus

$$p(\boldsymbol{d}_k|\mathbf{x}_{m,k} = \emptyset, \mathbf{x}_{m,k-1}, t_m) = P(T_m \in [\Delta t_m, \Delta t_m + t_0]\,|T_m \geq \Delta t_m). \tag{6.41}$$

Note that the expression $p(\boldsymbol{d}_k|\mathbf{x}_{m,k} = \emptyset, \mathbf{x}_{m,k-1}, t_m)$ here is a full expansion of the death prior $p(\boldsymbol{d}_k^{(i)}|\boldsymbol{d}_{k-1}^{(i)})$ in equation (6.22).

According the definition of the conditional probability, above equation (6.41) can be written as

$$\begin{aligned} p(\boldsymbol{d}_k|\cdot) &= \frac{P\big(T_m \in [\Delta t_m, \Delta t_m + t_0] \cap T_m \geq \Delta t_m\big)}{P(T_m \geq \Delta t_m)} \\ &= \frac{P(T_m \in [\Delta t_m, \Delta t_m + t_0])}{P(T_m \geq \Delta t_m)} \\ &= \frac{\int_{\Delta t_m}^{\Delta t_m + t_0} \mathcal{G}(T_m|\alpha, \beta) dT_m}{\int_{\Delta t_m}^{\infty} \mathcal{G}(T_m|\alpha, \beta) dT_m}. \end{aligned} \tag{6.42}$$

When $\Delta t_m = 0$, it becomes a cumulative distribution function of the gamma pdf since

$$\int_0^{\infty} \mathcal{G}(T_m|\alpha, \beta) dT_m = 1. \tag{6.43}$$

The gamma cdf is defined as

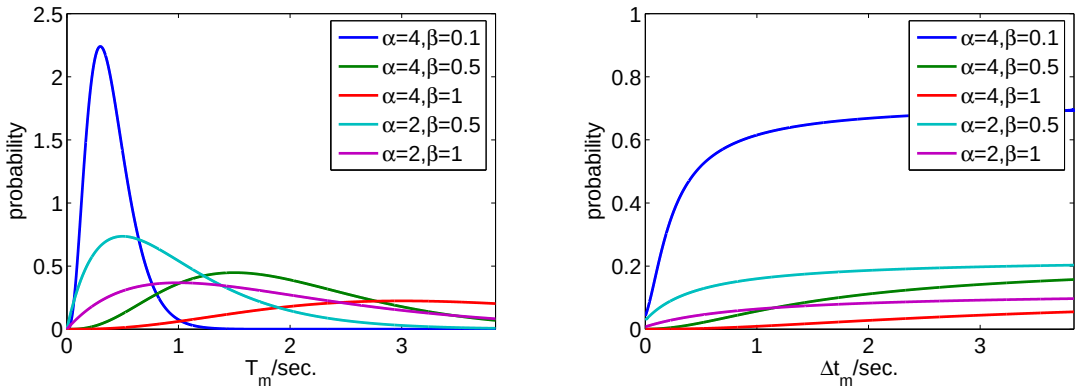
$$G(T_m|\alpha, \beta) = \int_0^{T_m} \mathcal{G}(x|\alpha, \beta) dx. \tag{6.44}$$

In the case of $\Delta t_m > 0$, the expression (6.42) can be written as

$$\begin{aligned} p(\boldsymbol{d}_k|\cdot) &= \frac{\int_0^{\Delta t_m + t_0} \mathcal{G}(T_m|\alpha, \beta) dT_m - \int_0^{\Delta t_m} \mathcal{G}(T_m|\alpha, \beta) dT_m}{1 - \int_0^{\Delta t_m} \mathcal{G}(T_m|\alpha, \beta) dT_m} \\ &= \frac{G(\Delta t_m + t_0|\alpha, \beta) - G(\Delta t_m|\alpha, \beta)}{1 - G(\Delta t_m|\alpha, \beta)}. \end{aligned} \tag{6.45}$$

Fig. 6.5(b) presents several choices of death probability under different Gamma parameters. The gamma parameter pair (α, β) controls how fast the source dies. Given a Gamma distribution, the death probability is actually determined by the period that it is not associated but still alive $\Delta t_m = t_{k-1} - t_m$. Normally, the larger is Δt_m, the higher possibility the source m dies. Once the death of the source is determined, we can simply set the corresponding state as an empty set, i.e., $\mathbf{x}_{m,k} = \emptyset$. The steps for computing the death probability are shown in Algorithm 11.

Source tracks may be placed only for the initial scan due to incorrect birth. This leads a zero track length, and is essentially indistinguishable from the clutter for online algorithms. However, it can easily be eliminated by post-processing; simply delete those estimated sources with very short (zero or one frame interval) existence period. For example, source 2 marked in Fig. 6.4, which may only exist at a one-point time.



(a) (b)

Figure 6.5: *(a) Gamma probability density function under the different Gamma parameters; (b) probability of death under the different Gamma parameters.*

6.4.3 Source survival

The states of the survival sources are constructed by the existing states after the death operation. If the source is not dead at current time step k, it is surviving with a probability of p_{s}. Suppose that the death probability for the mth source, $m = 1, \ldots, M_{k-1}$ is $p(\boldsymbol{d}_k|\mathbf{x}_{m,k} = \emptyset, \mathbf{x}_{m,k-1}, t_m))$ ($p_{\mathrm{d}}(m)$ for short in the algorithm), the probability of survival is thus

$$p_{\mathrm{s}}(m) = 1 - p(\boldsymbol{d}_k|\mathbf{x}_{m,k} = \emptyset, \mathbf{x}_{m,k-1}, t_m)). \tag{6.46}$$

Algorithm 11: Calculate the death probability.

// At time step k for ith particle. The number of sources: $|\mathcal{X}_{k-1}^{(i)}| = M$.
Input: Gamma parameters (α, β); time stamp of the source t_m.
Output: The probability of death p_{d}
for $m \leftarrow 1$ **to** M **do**
  - calculate the interval that the source is not associated $\Delta t_m = t_{k-1} - t_m$;
  if $\Delta t_m = 0$ **then**
    - calculate the death probability $p_{\mathrm{d}}(m)$ according to equation (6.44);
  else
    - calculate the death probability $p_{\mathrm{d}}(m)$ according to equation (6.45).
  end
end

According to the source death model (6.10), the total probability of the survival sources is thus

$$P_{\mathrm{s}}(m) = \prod_{\forall:\ \{\mathbf{x}_{m,k|k-1}\} \subseteq \mathcal{X}_{k|k-1}} \{1 - p(\boldsymbol{d}_k | \mathbf{x}_{m,k|k-1} = \emptyset, \mathbf{x}_{m,k-1}, t_m)\}. \tag{6.47}$$

The algorithm for calculating the probability of survival sources is presented in Algorithm 12.

6.5 Particle filtering implementation

The Rao-Blackwellisation step is formulated in Section 6.3. This section focuses on a particle filtering implementation of the proposed RFS data association tracking approach. First, the RFS particle filtering algorithm and the state extraction approach are presented. Different error measures to evaluate the tracking performance are then introduced.

6.5.1 RBPF implementation

Several assumptions are made to reduce the exhaustive associations in the variable $\boldsymbol{\theta}_k^{(i)}$ when implementing the algorithm:

- at most one source can be born at a time step k;
- at most one source can die at a time step k;
- the total number of sources is bounded at $N_{\max}$.
- the source can only be generated within the boundary of the room.

Algorithm 12: Calculate the probability of survival sources.

// At time step k for the ith particle. The number of sources: $|\mathcal{X}_{k-1}^{(i)}| = M$.
// Given the death probability p_d calculated according to Algorithm 11.
Input: The death probability p_{d}; measurement z_k; predicted states.
Output: Survival probability P_{s}; states and the likelihood of the survival states $(\mathcal{X}_k^{(i)}, \mathbf{l}_{\mathrm{s}})$.
for $m \leftarrow 1$ **to** M **do**
  - calculate the probability of the mth survival sources $p_{\mathrm{s}}(m)$ according to equation (6.46);
  - calculate the total probability of the survival sources $P_{\mathrm{s}}(m)$ according to equation (6.47);
  - operate the EKF for the existing states: *// $\mathbf{l}_s$ represents the likelihood*
    $(\mathbf{x}_{m,k}, \mathbf{l}_{\mathrm{s}}(m)) \leftarrow \mathrm{EKF}(z_k; \mathbf{H}_{m,k}\mathbf{x}_{m,k|k-1}, \mathbf{R}_{m,k|k-1})$.
end

The restriction of at most one source can be born or die at a time step k is to guarantee that the association and the combinations are always limited. In practice, the number of simultaneously active speakers is always assumed to be small, and thus the maximum number of the sources is bounded to $N_{\max}$ to reduce unnecessary associations. This is easily obtained by set the birth probability as 0 when the maximum number of sources achieves, i.e., $|\mathcal{X}_{k-1}| = N_{\max}$. These assumptions hence keep the complexity of the algorithm. The algorithm of the importance function calculation is summarised in Algorithm 13. The advantage of using a particle filtering here is that it allows a random hypothesis pruning rather than the typical heuristic hypothesis pruning applied in the traditional data association based tracking algorithms.

Algorithm 13: Calculate the optimal importance function.

// *At time step k for the ith particle. The number of sources*: $|\mathcal{X}_{k-1}^{(i)}| = M$.
Input: predicted states $\mathcal{X}_{k|k-1}^{(i)}$ and importance weight at previous time step $\tilde{w}_{k-1}^{(i)}$.
Output: updated states $\mathcal{X}_k^{(i)}$ and the weight $w_k^{(i)}$.
Initialisation: $p_{\mathrm{d}}(0) = 1$.
Association prior: $p_\gamma(0) = p_f$;
$p_\gamma(j) = \frac{1-p_f}{n}, \quad 0 < j \leq M; \quad n$ is M_k in equation (6.24).
- calculate p_{d} according to Algorithm 10.
- calculate $(p_{\mathrm{b}}, \mathbf{x}_{M+1,k}, \mathbf{l}_{\mathrm{b}})$ according to Algorithm 11.
- calculate survival probability p_{s}, states and the likelihood of the survival sources $(\mathcal{X}_k^{(i)}, \mathbf{l}_{\mathrm{s}})$ according to Algorithm 12.
- calculate the likelihood for clutter $\mathbf{L}(0,0)$ according to equation (6.26).
- set the likelihood as $\mathbf{L}(0,j) = \mathbf{l}_{\mathrm{s}}(j)$;$\mathbf{L}(1,j) = \mathbf{l}_{\mathrm{b}}(1); \quad 0 < j \leq M$
Formulate the probabilities for different hypotheses according to the numerator of equation (6.30) :
for $\ell \leftarrow 0$ **to** 1 **do**
  for $m \leftarrow 0$ **to** M **do**
    for $j \leftarrow 0$ **to** M **do**
      $\mathbf{P}(\ell, m, j) \leftarrow p_{\mathrm{b}}(\ell) p_{\mathrm{d}}(m) p_\gamma(j) p_{\mathrm{s}}(m) \mathbf{L}(\ell, j);$ $\quad j \neq m$ if $m \neq 0$
    end
  end
end
calculate the denominator term of equation (6.30) according to (6.32):
$\hat{\mathbf{P}} \leftarrow \sum_{\ell,m,j} \mathbf{P}(\ell, m, j)$.
normalise the hypothesis probability: $\mathbf{P}(\cdot,\cdot,\cdot) \leftarrow \mathbf{P}(\cdot,\cdot,\cdot)/\hat{\mathbf{P}}$.
draw the hypothesis: $\boldsymbol{\theta}_k^{(i)} = (\boldsymbol{\gamma}_k^{(i)}, \boldsymbol{d}_k^{(i)}, \boldsymbol{b}_k^{(i)}) \sim \mathbf{P}(\cdot,\cdot,\cdot)$, and the updated states $\mathcal{X}_k^{(i)}$ accordingly;
update the particle weight: $w_k^{(i)} = \tilde{w}_{k-1}^{(i)} \hat{\mathbf{P}}$.

Due to the multi-modality of the posterior distribution obtained by RBPF, extract the final state estimation is not as straightforward as in the particle filtering for the single source scenario. The histogram like visualisation probability hypothesis density (PHD) derived can be obtained as [119]

$$D(\hat{\mathbf{x}}_k) = \sum_{i=1}^{N} w_k^{(i)} \sum_{j=1}^{M_k^{(i)}} \mathcal{N}(\mathbf{x}_k; \mathbf{m}_j^{(i)}, \mathbf{P}_j^{(i)}), \tag{6.48}$$

where $M_k^{(i)}$ is the dimension of the ith particle $\mathcal{X}_k^{(i)}$, i.e., $M_k^{(i)} = |\mathcal{X}_k^{(i)}|$, and $\mathbf{m}_j^{(i)}$ and $\mathbf{P}_j^{(i)}$ are the mean vector and the variance matrix of jth source in the ith particle respectively. The PHD visualisation of RBPF posterior is straightforward, but since it requires the approximation of all the Gaussian densities, the computation is expensive.

Following a similar hypothesis extraction approach in MHT, Vihola [119] also proposes a "winner particle" extraction approach. Let

$$\hat{w}_k^{(i)} = \sum_{i'=1}^{N} w_k^{(i')} \delta_{\boldsymbol{\theta}_k^{(i)}}(\boldsymbol{\theta}_k^{(i')}). \tag{6.49}$$

The term $\hat{w}_k^{(i)}$ is thus the duplicate of all the same hypotheses. The best hypothesis can then be extracted as

$$\hat{I}_k = \arg\max_{1 \le i \le N} \hat{w}_k^{(i)}. \tag{6.50}$$

The source states can thus be estimated from this best hypothesis. The number of the sources is simply the cardinality of each state estimation.

6.5.2 Multiple state error measures

In addition to visualizing the states of multiple sources $\hat{\mathcal{X}}$ for a single trial, it is also necessary to find measures to evaluate the estimation performance over many Monte Carlo trials. The following evaluations are used: the percentage the estimator can estimate the right number of the sources, and given the correct estimation of the source number, how far the number and position estimates deviate from the ground truth. Suppose J Monte Carlo simulations are implemented, and let $\hat{\mathcal{X}}_{j,k}$, $j = 1, \cdots, J$ and $\mathcal{X}_k$ represent the state estimation at jth implementation and the ground truth respectively, and $\hat{M}_{j,k} = |\hat{\mathcal{X}}_{j,k}|$ is the source number estimation.

Probability of correct number estimation For the multiple sources estimation, it is obviously

interesting to know the probability that the source number estimate matches the actual number of the sources. The probability of the correct number estimation is thus defined as

$$P_k = \frac{1}{J}\sum_{j=1}^{J} \delta_{|\mathcal{X}_k|}\left(|\hat{\mathcal{X}}_{j,k}|\right) \times 100\%, \tag{6.51}$$

The probability of correct number estimation illustrates the percentage that the tracking algorithm reports the number of the sources correctly.

Cardinality error of the source number estimation The cardinality error of the source number estimation ϵ_k is defined as

$$\epsilon_k = \sqrt{\sum_{j=1}^{J} \frac{1}{J}\left|\hat{M}_{j,k} - M_k\right|^2}. \tag{6.52}$$

It gives the root mean square error (RMSE) over multiple implementations between the source number estimates and the ground truth.

Global mean deviation It is even more important to know the error between the estimated positions and the ground truth. The family of Wasserstein distances (WD) [123, 124], which accounts for the deviations in cardinality, is able to evaluate the distance between the two sets. Consider any state estimates $\hat{\mathcal{X}}_k$ and the ground truth $\mathcal{X}_k$, and cardinalities $\hat{M}_k = |\hat{\mathcal{X}}_k|$ and $M_k = |\mathcal{X}_k|$. The WD is defined as

$$d_p(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \min_{C} \sqrt[p]{\sum_{i=1}^{M_k}\sum_{j=1}^{\hat{M}_k} C_{i,j} d(\hat{\mathbf{x}}_{j,k}, \mathbf{x}_{i,k})^p}, \tag{6.53}$$

where the infimum is taken over all $\hat{M}_k \times M_k$ transportation matrices C. The transportation matrices C is defined as

$$\sum_{i=1}^{M_k} C_{i,j} = \frac{1}{\hat{M}_k}, \qquad \sum_{j=1}^{\hat{M}_k} C_{i,j} = \frac{1}{M_k}, \tag{6.54}$$

for all $i = 1, \ldots, M_k$ and $j = 1, \ldots, \hat{M}_k$. This implies

$$\sum_{i=1}^{M_k}\sum_{j=1}^{\hat{M}_k} C_{i,j} = 1. \tag{6.55}$$

In equation (6.53), the larger the value of p, the more the metric penalises the error on number estimation. Defining a localization error is still a problem on its own when the state cardinality estimate is incorrect, i.e., $|\hat{\mathcal{X}}_k| \neq |\mathcal{X}_k|$. Equation (6.53) partly solves this problem. However, the metric actually depends on how well the numbers of estimate points among the actual objects are balanced. For example, suppose there is only one source exist, but five estimates which are very close to the ground truth are obtained from the tracking algorithm. In such a case, the cardinality of the estimated state is far away from the ground truth. Since the estimates are perfectly balanced for the ground truth source, the metric does not detect the cardinality error and subsequently only presents a small error. The downside of this metric is summarised and a more optimal metric for the multisource estimations is presented in [125].

Since the probability of correct number of estimation is already defined, the deviations under the correct number estimation will be considered here. Let $|\hat{\mathcal{X}}_k| = |\mathcal{X}_k| = M_k$. The multiple speaker deviation can be formulated as [1]

$$d(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \min_{\sigma} \sqrt{\frac{1}{M_k} \sum_{i=1}^{M_k} \|\mathbf{B}\hat{\mathbf{x}}_{\sigma_i,k} - \mathbf{B}\mathbf{x}_{i,k}\|^2}, \tag{6.56}$$

where the minimum is taken over all permutations on the numbers σ, and $\mathbf{B} = [\mathbf{I}\,\mathbf{0}]$ is a position extraction matrix such that $\mathbf{B}\mathbf{x}_k$ outputs the position part (x_k, y_k) of the state vector $\mathbf{x}_k$.

Special case Let $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_k\}$ and $\mathcal{X}_k = \{\mathbf{x}_k\}$. The mean deviation $d(\hat{\mathcal{X}}_k, \mathcal{X}_k)$ defined in (6.56) reduces to the Euclidean distance between the estimates $\hat{\mathbf{x}}_k$ and the ground truth $\mathbf{x}_k$:

$$d(\hat{\mathcal{X}}_k, \mathcal{X}_k) = \sqrt{\|\mathbf{B}\hat{\mathbf{x}}_k - \mathbf{B}\mathbf{x}_k\|^2}. \tag{6.57}$$

With (6.56), a mean deviation under the correct estimation of the source number can be defined as

$$\xi_k = \mathbb{E}\left(d(\hat{\mathcal{X}}_{j,k}, \mathcal{X}_k) \middle| |\hat{\mathcal{X}}_{j,k}| = |\mathcal{X}_k| \right). \tag{6.58}$$

The measures (6.51) and (6.58) have been used to evaluate the performance tracking algorithm in [1]. In the following chapters, all the three measures defined in equations (6.51), (6.52), and (6.58) will be employed to evaluate the performance of the proposed tracking systems. In general, equation (6.51) and (6.52) are able to give the performance of the source number estimation. Equation (6.58) gives the position estimation errors conditional on the correct number estimation. All these measures are necessary in evaluating the performance of a tracking algo-

rithm since for multiple sources, both the accuracy of position estimation and source number estimation are interested. Further, all these measures represent the cardinality estimation error and position estimation error straightforwardly.

6.6 Experiments

In this section, the experiments based on the simulated reverberant environment as well as the real room environment are organised to evaluate the performance of the proposed algorithm. The whole experiment is almost the same as in the previous chapter except that two speakers have time-varying appearance and have an overlap during a period.

Since the sources are following the same trajectories in all of our experiments, the parameters for the EKF step are set the same as in Table 5.3, in Section 5.6.2, on page 137. The parameters ς and η in Langevin model have no significant influence on the tracking performance because the particle filtering itself allows a flexible state transition model. All the initial positions of the sources are assumed to be unknown and the central position of the room is given if a new source $\mathbf{x}_0$ is born. The measurement noise variance $\mathbf{R}$ is set to $5 \times (0.1)^9$, which is determined by experimental study. The initial covariance matrix $\mathbf{P}_0$ and the measurement noise variance $\mathbf{R}$ control the convergence velocity of the position tracking; a larger value in the matrix makes the EKF converge to the real position faster but may lead to a larger variance.

It is difficult to compare with other tracking approaches since no such algorithms were employed in the acoustic source tracking before. It is also unfair to simply implement other algorithms in our experiments since the performance of an algorithm highly depends on a bunch of parameters and different experiment may require different parameter setup.

6.6.1 Tracking performance under a simulated room environment

The room environment and experiment setup are illustrated in Fig. 2.9 on page 44. The two sources have a time-varying appearance: one is active from frame index 1 to 50, and the other from the frame index 30 to 80.

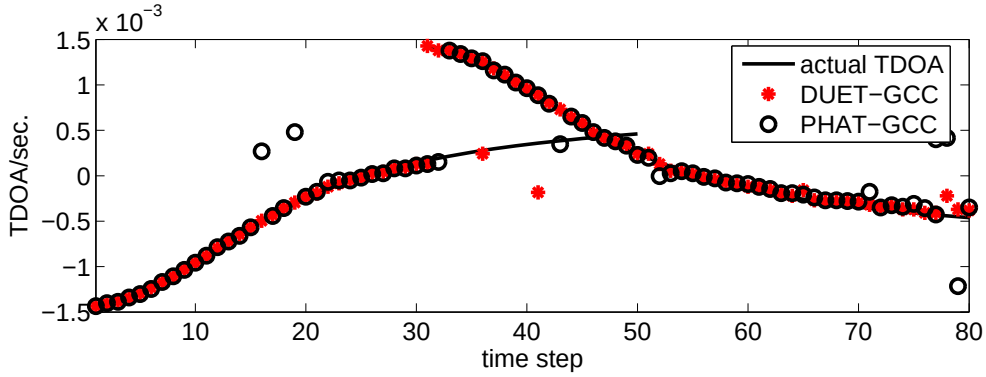
6.6.1.1 Tracking results from a single experiment

In the first experiment, the algorithm is implemented with different number of particles. All the wall reflection coefficients are set to 0.6, which leads to a reverberation time $T_{60} = 0.163$s. The SNR is set to 30dB. As mentioned at the beginning of this chapter, both the DUET-GCC method and the PHAT-GCC method introduced in Chapter 4 are employed to extract the TDOA measurements. The tracking performance from these two measurement extraction approaches will be compared. To avoid an exhaustive data association, the TDOAs are extracted from DUET-GCC method and PHAT-GCC method by setting the threshold values as 0.7 and 0.9 respectively to excludes the most false alarms. As investigated in Chapter 4, such threshold values are found able to keep a satisfactory probability of detection and reject the false alarms effectively. The corresponding probabilities of detection and false alarm can be found in Table 6.2. Due to the interference between the simultaneously active sources, the probability of detection is much lower than that in Chapter 5, where nonconcurrent multiple sources are considered. It is worth pointing out that the probabilities of detection and false alarm in Table 6.2 are obtained from the four microphone pairs, and the prior of false alarm in the tracking algorithm will set to a value around the probability of false alarm, as shown, p_f in Table 6.3.

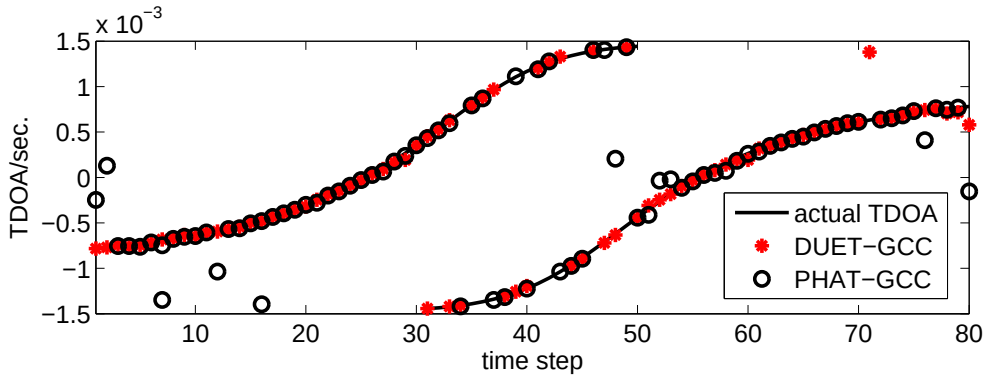
	DUET-GCC	PHAT-GCC
probability of detection P_D	0.763	0.710
false alarm rate P_F	0.059	0.134

Table 6.2: *Probabilities of detection and false alarm for DUET-GCC and PHAT-GCC methods respectively ($T_{60} = 0.163$s). The threshold for these two methods are 0.7 and 0.9 respectively.*

Figure 6.6 displays the TDOAs obtained from microphone pair 1 and microphone pair 2. Due to the reverberation and the interference between the source signals, it is very difficult to extract the TDOAs for all the sources when the sources are simultaneously active, as shown from time step 30 to time step 50 in Fig. 6.6. DUET-GCC method presents better TDOA measurements for simultaneously active sources when the TDOAs of two sources are spatially separated enough, e.g., TDOAs extracted from microphone pair 2. The TDOA estimates from source 2 are well extracted in microphone pair 1 during two source overlap time steps since it is closer to the microphone pair 1 and a higher SRR and SNR can be achieved. Very similar TDOA estimation performance can be found in microphone pair 3 and microphone pair 4. When the two sources are closely spaced, both methods fail to extract the TDOAs simultaneously for two



(a)



(b)

Figure 6.6: *TDOA estimates of (a) microphone pair 1; (b) microphone pair 2 from DUET-GCC and PHAT-GCC methods.*

sources due to the resolution problem (2-D histogram resolution for DUET-GCC method and peak resolution in GCC function for PHAT-GCC method). Although the probability of detection of DUET-GCC method is only a bit higher than that of PHAT-GCC method, the former is able to reduce the false alarm rate significantly.

For the parameters in the tracking algorithm, the birth prior P_{b} and the gamma parameter pair (α, β) are set according to the experimental study. $P_{\mathrm{b}} = 0.1$ and $(\alpha, \beta) = (4, 0.4)$ are found satisfactory in the experiments. Since the probabilities of false alarm for the DUET-GCC and PHAT-GCC based TDOA measurements are different, they are set to be $p_f = 0.05$ and $p_f = 0.1$ for DUET-GCC and PHAT-GCC measurement based tracking algorithm respectively. The prior of false alarm is chosen according to the probability of false alarm evaluated from all the TDOA measurements, as shown, in Table 6.2. Generally, fewer particles are needed to achieve a saisfactory estimation, since the optimal importance function is employed and particles are

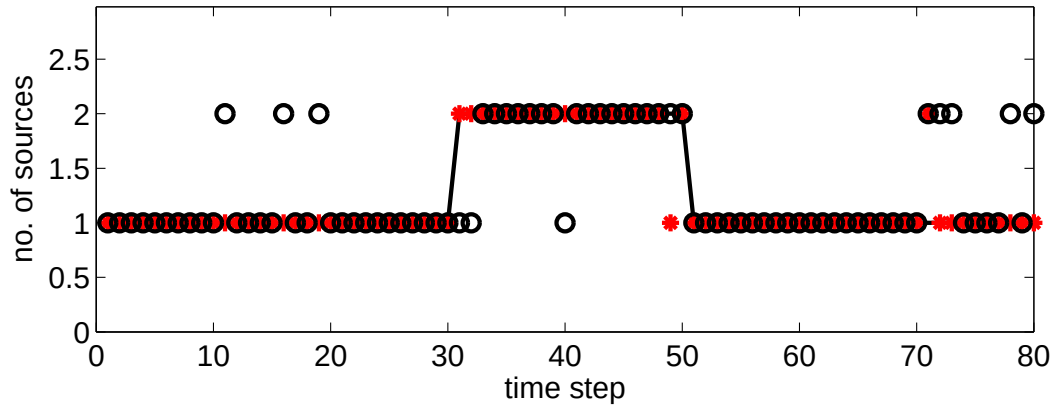
drawn efficiently [119]. In this experiment, 50 particles are used. All these parameters for the tracking algorithm are summarised in Table 6.3.

parameter	P_{b}	p_f	(α, β)	N
value	0.1	0.05/0.1	(4,0.4)	50

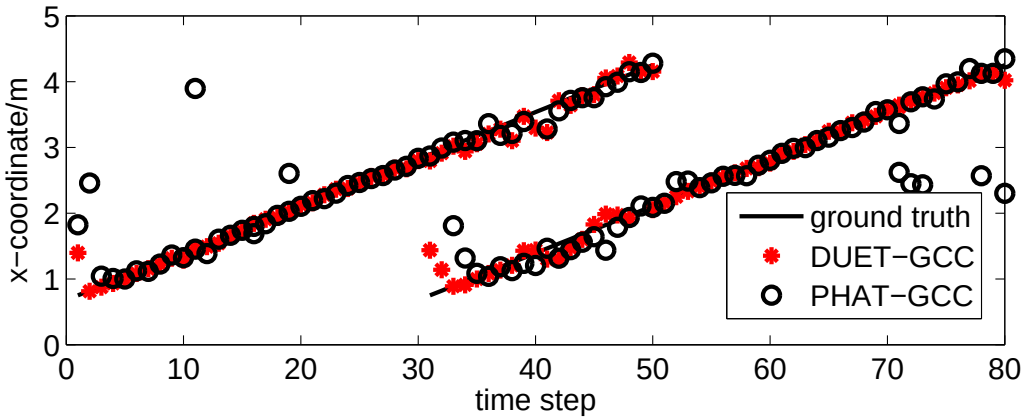
Table 6.3: *Parameter setup for the RBPF tracking algorithm. Note that the false alarm rate is set to be 0.05 and 0.1 for DUET-GCC and PHAT-GCC TDOA measurements respectively.*

Figure 6.7 shows the estimation result from a single trial of the DUET-GCC and PHAT-GCC measurement based approaches. It shows that the proposed algorithm is able to estimate the number of the active sources as well as the positions. Although there are large measurement missing at the time steps when two sources are simultaneously active, the algorithm is still able to preserve the track and lock on the sources with a satisfactory performance. The position tracking results are worse at the multiple source time steps because the TDOA measurements are not as accurate as they were extracted in the single source scenario. The tracking loss likely happens when all or most of microphone pairs fail to report the correct TDOA measurements. Particularly, both the approaches are able to track the trajectories of the sources. However, false detections are presented at some time steps in the PHAT-GCC measurement based approach due to heavy clutters.

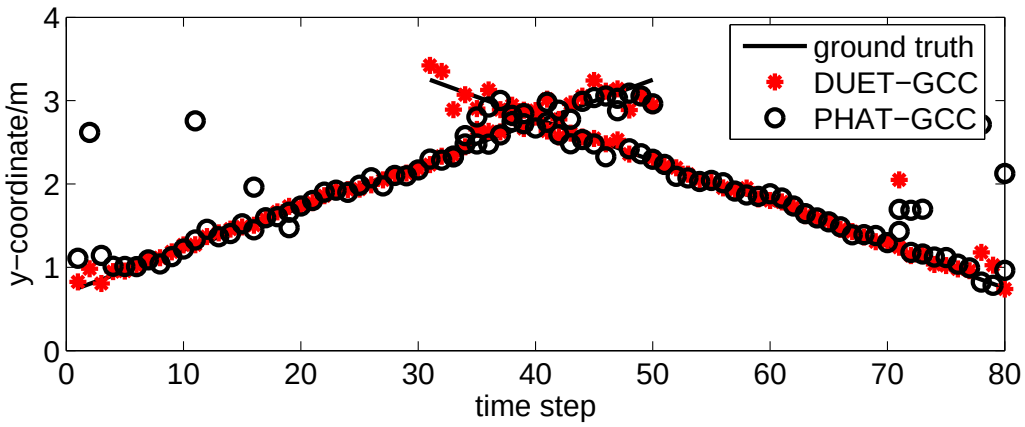
To give an exact tracking error over many implementations, the measures defined in Section 6.5.2, the probability of correct number estimation P_k, the cardinality error ϵ_k and the global mean deviation ξ_k are employed. Fig. 6.8 shows the average results over 100 Monte Carlo runs. It further shows the capability of the proposed tracking approach in tracking multiple time-varying number of sources. The errors from the cardinality estimation and the position estimation are small, even when the two sources are simultaneously active. The large error only presents at the time steps that the number of sources changes (complete missing of the detections can be regarded as a special case that the source number changes). Again, the DUET-GCC measurement based tracking performance is better than that of the PHAT-GCC measurement based tracking performance. For PHAT-GCC method, due to a large measurement missing at some time steps, it is very unlikely for the proposed tracking algorithm to detect the source. For example, at time steps 16, 31 and 71, the probability of correct number estimation is very small and all the other errors are large. Note that the global mean deviation can not fully illustrate the error of the position estimation since only the estimations with correct cardinality



(a)



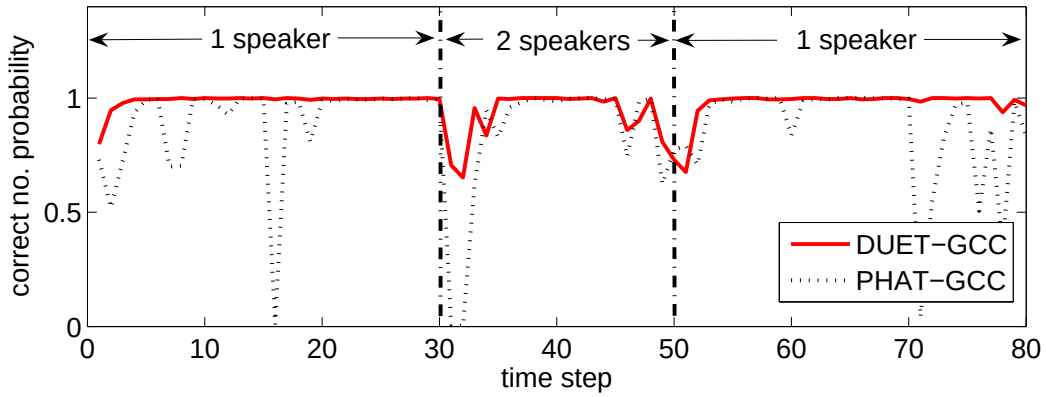
(b)



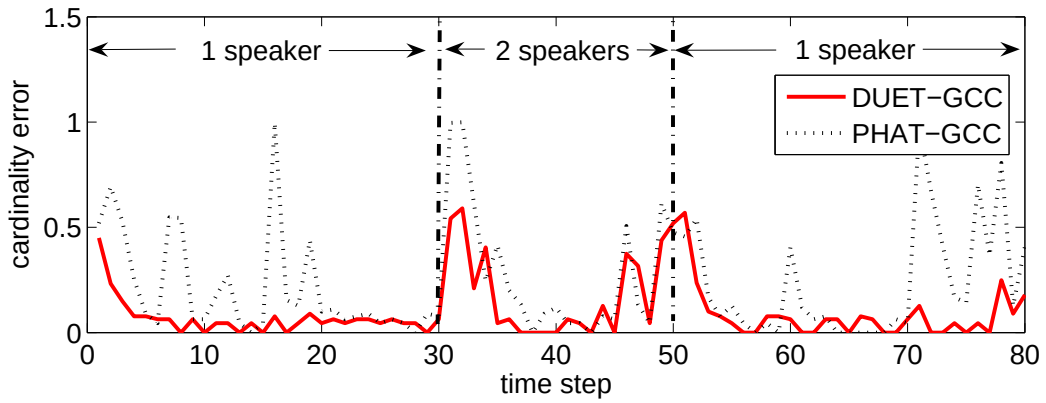
(c)

Figure 6.7: *Tracking result of a single trial under the reverberant environment ($T_{60} = 0.163s$). (a) Estimation of the number of the sources; (b) estimation results of the x-coordinate; and (c) estimation result of the y-coordinate.*

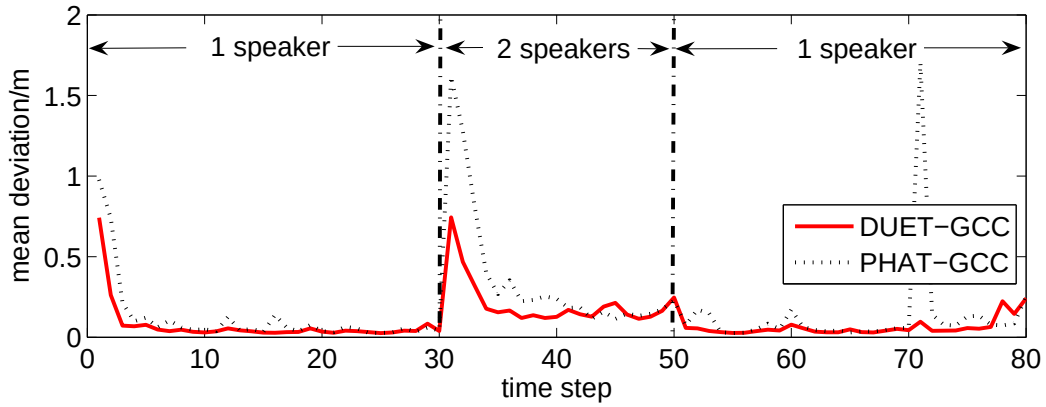
are employed. In some implementations, the algorithm is actually able to report the source position accurately but the position estimation is not counted in due to an over-estimation or under-estimation of the source number.



(a)



(b)



(c)

Figure 6.8: *Average tracking result of 100 Monte Carlo simulations under the reverberant environment ($T_{60} = 0.163s$). (a) Correct number estimation probability; (b) cardinality error; and (c) mean deviation.*

We also implement the algorithm with different number of particles to roughly get the upper error bound of the tracking performance. The average results for the DUET-GCC measurements (and the PHAT-GCC measurements in parathesis) based approaches versus different number of particles over 100 Monte Carlo simulations are shown in Table 6.4. The correct number probability P can be improved significantly when 50 particles are used. The cardinality error ϵ and global mean deviation ξ are also reduced sharply with such number of particles. Although the global mean deviation can be further reduced by increasing the number of the particles, but such improvement is trivial and no significant improvement can be found for the correct number probability. This is because the EKF source position state marginalisation can be refined by using more particles, however the performance of data association variable estimation is less affected. For both the DUET-GCC and PHAT-GCC measurement based approaches, 50 particles are thus used throughout the next experiments.

	method	10	20	50	100	200
Correct num. prob. P	DUET-GCC	0.910	0.956	0.969	0.965	0.967
	PHAT-GCC	0.729	0.824	0.869	0.872	0.858
Cardinality error ϵ	DUET-GCC	0.270	0.145	0.102	0.106	0.101
	PHAT-GCC	0.491	0.347	0.240	0.227	0.231
Global mean dev. ξ	DUET-GCC	0.188	0.134	0.103	0.092	0.0843
	PHAT-GCC	0.282	0.228	0.185	0.157	0.179

Table 6.4: *Tracking performance based on the DUET-GCC measurements and the PHAT-GCC measurements vs. different number of particles under the reverberant environment ($T_{60} = 0.163s$).*

6.6.2 Different simulated room environment

The algorithm is further implemented in a series of experiments to fully evaluate its performance. The experiments are set up with following system parameters:

- different simulated noisy environments, i.e., different SNRs.
- different simulated reverberant environments, i.e., different T_{60}s.

In the experiments, the algorithm is implemented in the anechoic, low reverberant and moderate reverberant environment respectively. The corresponding reverberation time T_{60}s are 0s, 0.163s and 0.289s respectively. Different noisy environments 0dB, 10dB and 20dB are also set to fully

experiments		T_{60}			SNR		
		0s	0.163s	0.289s	0dB	10dB	20dB
DUET-GCC	P_D	0.803	0.763	0.500	0.543	0.755	0.800
	P_F	0.006	0.059	0.433	0.472	0.079	0.012
PHAT-GCC	P_D	0.780	0.710	0.403	0.458	0.755	0.778
	P_F	0.034	0.134	0.539	0.483	0.088	0.037

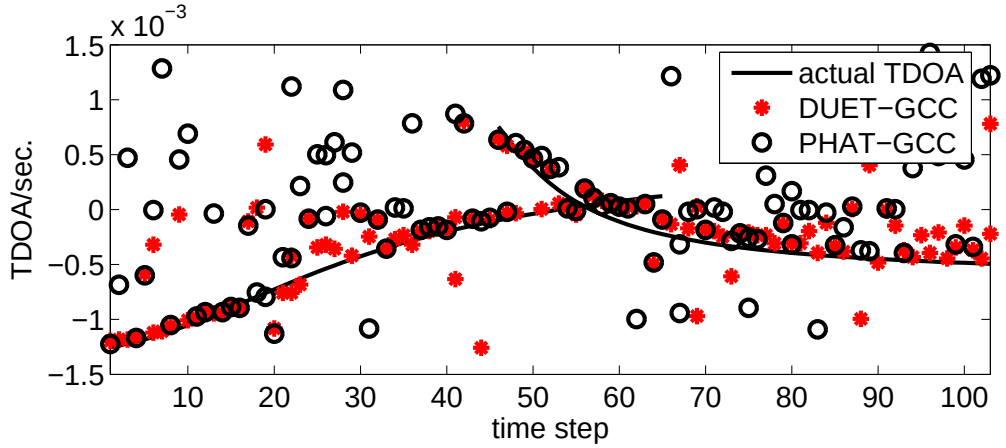
Table 6.5: *The probabilities of detection and false alarm under different reverberation time T_{60}s and SNRs.*

analyse the performance of our approaches. The probabilities of detection and false alarm under different adverse environments are illustrated in Table 6.5.

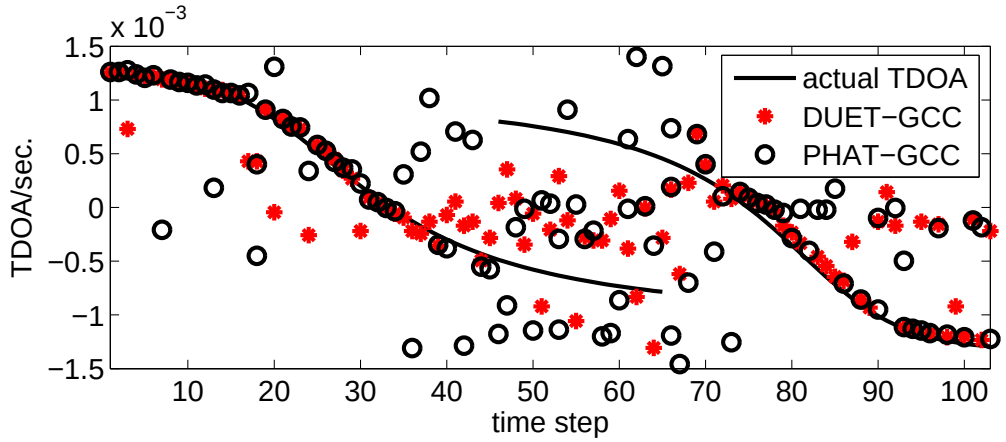
All the parameters of the tracking algorithm are set the same as them in Table 6.3, except that the priors of the false alarm are chosen as 0.5 and 0.55 for the DUET-GCC and PHAT-GCC based tracking respectively under the heaviest reverberant and noisy environments, i.e., $T_{60} = 0.289$s and SNR=0dB. Table 6.6 gives the average results based on the DUET-GCC measurements and PHAT-GCC measurements over 500 Monte Carlo experiments. The tracking performance degrades as the noise or reverberation become heavier. This is mainly because the probabilities of detection and false alarms become worse. The tracking results based on the DUET-GCC method present better performance under all the experiments since the DUET-GCC method preserves the detections well and meanwhile excludes the false alarms more effective than the PHAT-GCC method.

experiments	method	T_{60}			SNR		
		0s	0.163s	0.289s	0dB	10dB	20dB
Correct num. prob. P	DUET	0.964	0.969	0.780	0.826	0.951	0.963
	PHAT	0.920	0.869	0.686	0.641	0.815	0.921
Cardinality error ϵ	DUET	0.106	0.102	0.369	0.332	0.132	0.100
	PHAT	0.159	0.240	0.540	0.534	0.306	0.161
Global mean dev. ξ	DUET	0.086	0.103	0.381	0.256	0.109	0.093
	PHAT	0.100	0.185	0.494	0.439	0.136	0.112

Table 6.6: *Tracking performance based on the DUET-GCC measurements and the PHAT-GCC measurements under different adverse environments.*



(a)



(b)

Figure 6.9: *TDOA estimates of (a) microphone pair 4; (b) microphone pair 14 from DUET-GCC and PHAT-GCC methods in the real audio lab environment. Source 1 is active from time step 1 to 65, and then source 2 follows from time step 46 to 103.*

6.6.3 Real recording experiment

As in Chapter 5, the tracking algorithm is also implemented in a real audio lab environment. The detailed illustration of the whole recording environment can be found in Appendix A. The experiment setup is presented in Section 2.6.3, on page 45. The recorded signals are exactly the same as those in Chapter 5. Since the aim is to simulate time-varying number of multiple acoustic sources, the two recorded signals are overlapped between time step 45 to 65 to generate simultaneously active sources.

The measurements extracted from the microphone pair 4 (microphone 4 and 5) and microphone pair 14 (microphone 17 and 18) are presented in Fig. 6.9. The same as in Chapter 5, the ground

truth TDOAs are also roughly calculated from the proposed trajectories since the actual moving trajectories are not exactly known. Due to the heavy reverberation (the reverberation time T_{60} is as long as 0.8s as measured in Section 4.6.1 on page 110), the TDOA measurements are seriously deteriorated for both DUET-GCC and PHAT-GCC methods.

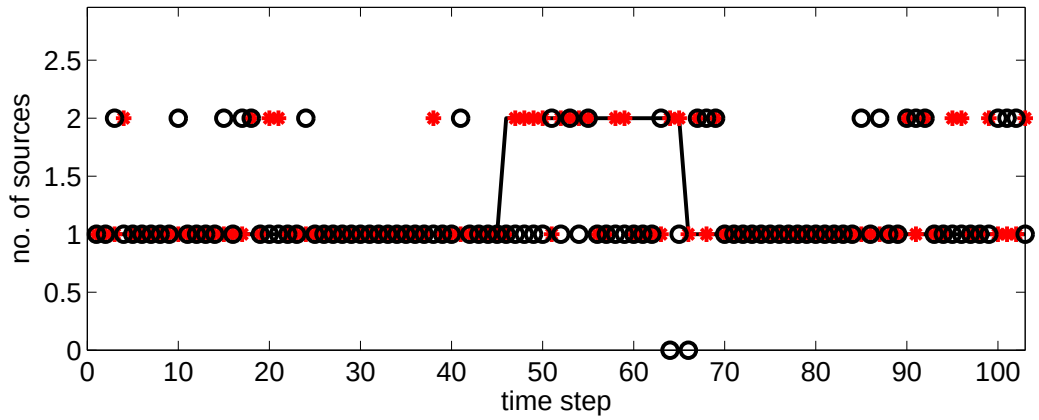
The parameters in the tracking algorithm are set the same as in the simulated experiments except the prior of the false alarm. Since the false alarms are heavier in the real audio lab experiments, the priors of false alarm are set as $p_f = 0.65$ and $p_f = 0.70$ for the DUET-GCC method and PHAT-GCC method respectively. The tracking results from a single experiment is shown in Fig. 6.10. Since the reverberation is much stronger, the tracking performance is worse than that in the simulated experiment. For the DUET-GCC measurement based tracking, the cardinality estimation is much better than that of that PHAT-GCC measurement. This is because the TDOA measurements based on DUET-GCC method are more accurate than PHAT-GCC method, particularly when the two sources are simultaneously active, e.g., from time step 46 to time step 65.

To fully illustrate the average tracking performance for the real recording signals, the measures introduced in Section 6.5.2 are presented in Fig. 6.11. The results show that both the cardinality and the position estimation performance based on the DUET-GCC measurements are better than that based on the PHAT-GCC measurements. The same as in the simulated environment, the performance is degraded at the time steps where source birth/death occurs. However, considering such a strong reverberant environment, the tracking performance from the DUET-GCC measurements is satisfactory. For the PHAT-GCC method, the tracking performance in the single source time steps is robust as well. The average errors is presented in Table 6.7.

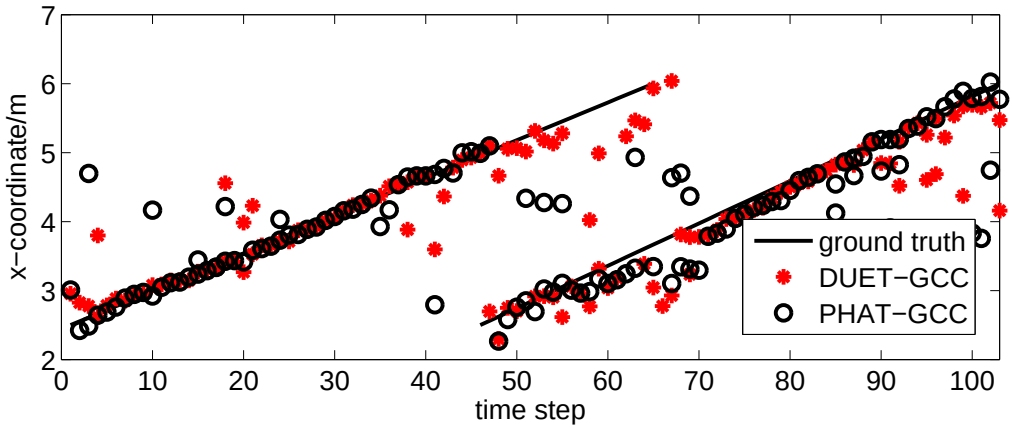
Other than from the tracking algorithm itself, the tracking errors also come from the error of the microphone positions and particularly, the inaccurate ground truth of source trajectories. The mean deviation is much higher than that in the simulated room experiments since the ground truth at each time step is estimated by assuming that the source motion speed is even and the sources follow the marked trajectories strictly (but actually this assumption is not true).

measurement	Correct num. prob. P	Cardinality error ϵ	Global mean dev. ξ
DUET-GCC	0.765	0.390	0.304
PHAT-GCC	0.686	0.509	0.382

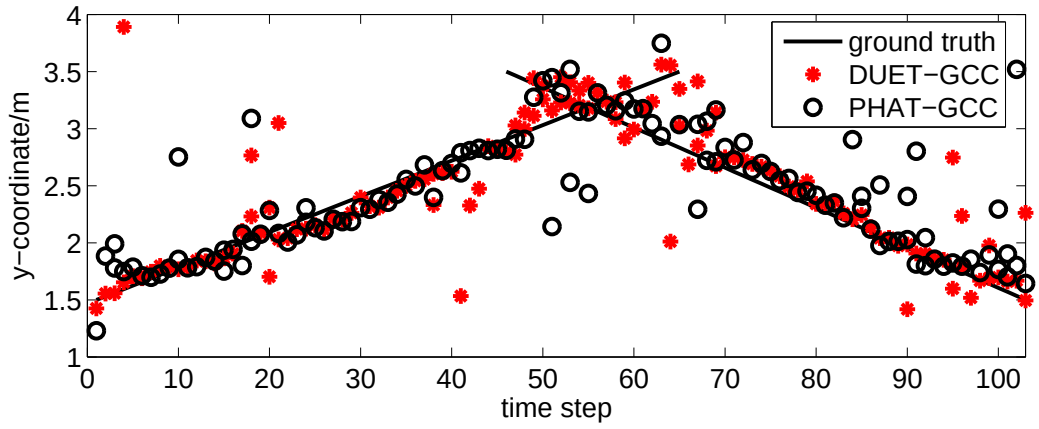
Table 6.7: *Average errors in the real audio lab environment.*



(a)



(b)



(c)

Figure 6.10: *Tracking result of the real recording signals. (a) Estimation of the number of the sources; (b) estimation results of the x-coordinate; and (c) estimation result of the y-coordinate.*

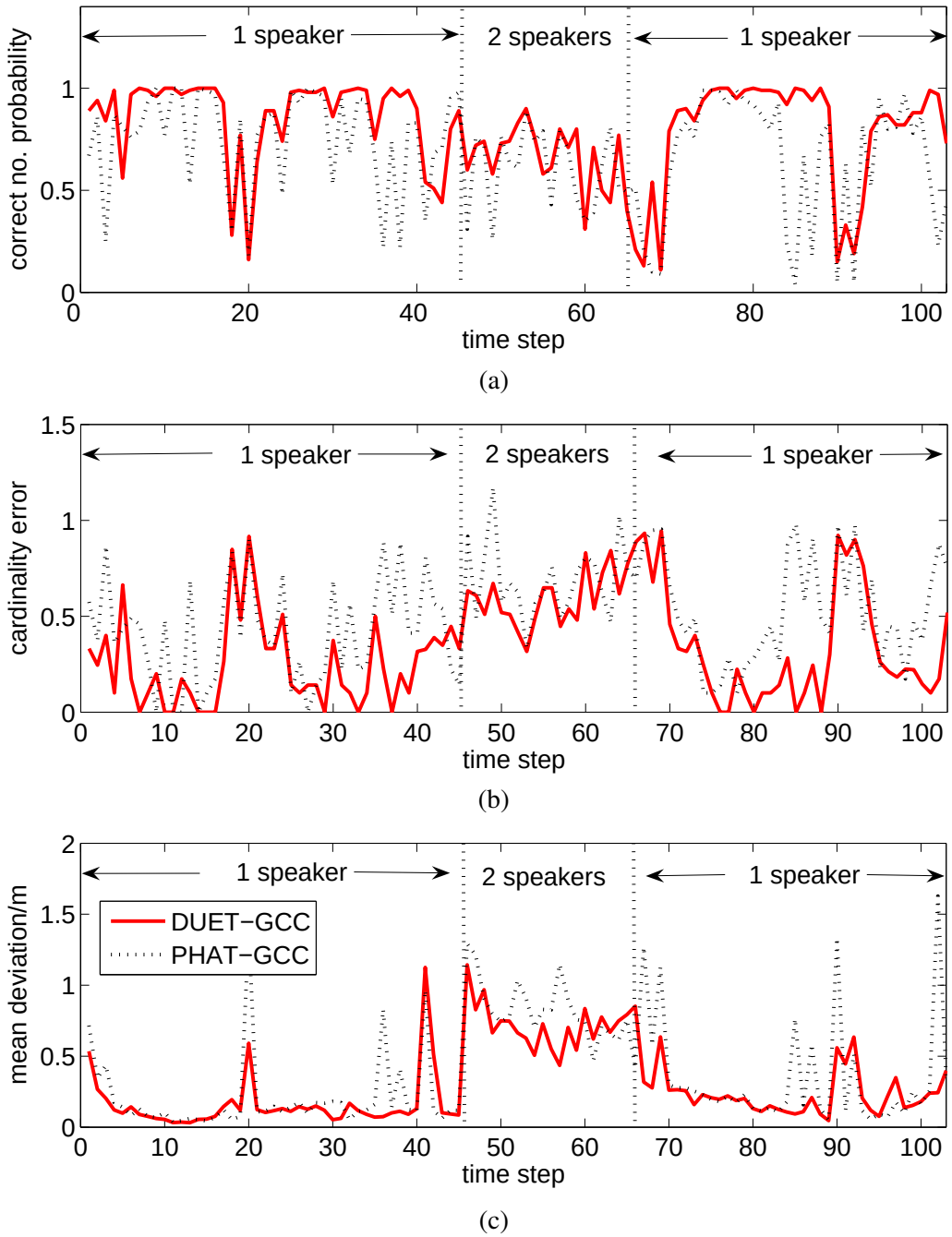


Figure 6.11: *Average tracking result of 100 Monte Carlo implementations in the real audio lab environment. (a) Correct number estimation probability; (b) cardinality error; and (c) mean deviation.*

6.7 Chapter summary

A Rao-Blackwellised particle filtering based random finite set approach is introduced and modified in this chapter to track the unknown and time-varying number of speakers. The multiple

source states are constructed by the position set and an additional association variable which indicates source dynamics. The Rao-Blackwellisation technique is employed to reduce the estimation variance and sampling efficiently, by which the position states are marginalized by using an extended Kalman filtering, and only the data association variable is handled by the particle filtering. Since the optimal importance function is derived, the particle are drawn effectively, and generally fewer particles are needed to achieve an accurate estimation.

Using the measurements extracted from two different approaches introduced in Chapter 4, the DUET-GCC and PHAT-GCC methods, the performance of the tracking approach is fully investigated in the simulated room environment as well as in the real audio lab environment. All the experiments show that the proposed speaker tracking approach is able to track multiple sources effectively. It is thus suitable for many related speech applications where the number of speaker is usually small. The experiment results also show that the tracking performance based on the DUET-GCC TDOA measurements are better than that based on the PHAT-GCC measurements. This is mainly because DUET-GCC separates the speech signals first, and is able to estimate the TDOA measurements from each individual source.

However, tracking multiple acoustic sources in the room environment is still a challenge problem due to the reverberation as well as the interference among the source signals. It is worth pointing out that for the tracking system developed in this thesis, the number of acoustic sources is assumed to be small. An interesting direction for future work is to investigate the tracking approach for large number (say more than two) of speakers. This unfortunately leads to following open questions. First, it requires more sophisticated approach to extract the TDOA measurements for multiple sources. This is not a trivial task since as mentioned at the beginning of this thesis, tracking the sources always requires a short frame length to keep the system locking on the source dynamics, and extract the TDOA measurements for multiple sources with such short frames will be very difficult. Second, the tracking approach developed in this chapter assigns different hypotheses between the source states and the measurements. As the number of sources increases, the computation will become more expensive to implement, particularly in the case that multiple source birth or death at a time step is considered.

Chapter 7
Conclusions and future work

The acoustic source tracking problem and solutions in the room environments have been investigated in this thesis. All the results obtained from previous chapters will be summarised in this chapter. Further, the conclusions will be drawn and the limitation of our work will be illustrated so as to determine the most promising directions for future work in this field.

7.1 Conclusions

The problem of acoustic source localisation and tracking in the room environment is extensively studied in the past decade. Although a series of EKF or PF can be regarded as standard modules for such problem, the existing solutions either did not consider multiple simultaneously active sources or cannot cope with the reverberation efficiently. Hence, some existing approaches should be improved and preferably, more advanced signal processing techniques should be exploited to adapt to such multi-target scenarios and the adverse environments. The final objective of this thesis is to develop a system which is able to track multiple unknown and time-varying number of acoustic sources. To formulate such a tracking system, the main tasks are from following two perspectives.

- Extract the TDOA measurements from the received signals emitted by an unknown number of multiple sources from the microphone pairs. Usually, the source signals can be simultaneously active and are distorted significantly by the adverse environments.
- Estimate the number of sources as well as the position of the sources from the extracted TDOA measurements. The measurements are also noise corrupted, and miss detection and false alarms are always presented.

This thesis has presented several improvements and extensions to the state of the art of multiple acoustic source tracking. Other than introducing the updated tracking algorithms, this thesis also proposes a new TDOA measurement extraction method and build a whole tracking system

appropriate for multiple source tracking in the room environment. Generally, the approaches developed in this thesis are able to track both the nonconcurrent or concurrent multiple sources, and are suitable for many speech applications where the number of simultaneously active speakers is small and the environment is with moderate noise and reverberation.

7.1.1 Outcomes of the thesis

The background knowledge of the acoustic source tracking problem has been fully reviewed and discussed in Chapter 2 and Chapter 3. Chapter 4 to Chapter 6 presents the contributions and work along this problem. The results can be summarised as follows.

- Chapter 4 first gives a definition of the probabilities of detection and false alarm. Based on this definition, a full investigation on the TDOA measurement performance of traditional PHAT-GCC method is presented. Particularly, a DUET-GCC method is developed to extract the TDOAs for multiple simultaneously active sources.

  The PHAT-GCC method assumes only one source signal impinges and thus only appropriate for single source TDOA measurement extraction. Since the speech mixtures can be assumed WDO in the time-frequency domain, DUET is introduced to separate the speech mixtures first, and the PHAT-GCC method is then applied to the spectrogram of each individual source. The TDOA measurement extraction for multiple simultaneously active sources is thus achieved.

  A number of parameters have been discussed for the PHAT-GCC and DUET-GCC methods. The microphone separation has a significant effect on the TDOA performance. A too small microphone separation (less than 0.3m) causes a resolution problem on the GCC function and 2-D histogram in the DUET-GCC method, while an overly large separation will reduce the correlation between the received signals. The best performance can be achieved when set the microphone separation around 0.5m.

  The performance of PHAT-GCC and DUET-GCC approaches under different noisy and reverberant environments is also investigated. Generally, DUET-GCC performs better than PHAT-GCC method when extracting the TDOA measurements for multiple simultaneously active sources under the anechoic and moderate noisy and reverberant environments. However, in the heavy noisy and reverberant environments, the WDO assumption is significantly deteriorated. The performance of DUET-GCC is thus degraded sharply.

- Chapter 5 addresses a special multiple source tracking case: nonconcurrent multiple source tracking. Under such a scenario, one source is active during a period, and the other follows. Two EKPF approaches are developed to track the sources, and to catch up with the position of the new source quickly. The core idea here is utilising an EKF to estimate the state coarsely. The particles can then be sampled around this posterior state estimation, rather than drawn according to the prior information in the traditional SIR-PF.

  The first EKPF approach uses a single TDOA measurement (corresponding to the largest GCC peak) from each microphone pair to estimate the state at the EKF step, while the other (multiTDOA-EKF) incorporates the reverberant measurement model which employs a set of TDOA measurements from each microphone pair to update the EKF. The simulated experiments, as well as the real recording experiments show that the approaches successfully locks on to the position of the new source. Due to incorporating the reverberant measurement model, the multiTDOA-EKF presents better performance than the single TDOA EKPF and traditional SIR-PF approaches.

- Chapter 6 considers the highest hierarchy of the acoustic source tracking problem: tracking a time-varying number of acoustic sources. Since the number of the sources is unknown and time-varying, it requires the tracker to estimate the number of the sources as well as the source positions.

  The DUET-GCC method developed in Chapter 4 is used to obtain the TDOA measurements for multiple sources. A random finite set (RFS) based Rao-Blackwellised PF is employed and modified to track the time-varying number of sources. Each particle has a RFS form encapsulating the states of all sources and is capable of addressing source dynamics: source survival, new source appearance and source deactivation. A data association variable is defined to depict the source dynamic and its relation to the measurements. The Rao-Blackwellisation step is used to decompose the state: the source positions are marginalised by using an EKF, and only the data association variable needs to be handled by a PF.

  The performance of the tracking approach is extensively studied under a number of experimental parameters, e.g., different SNRs and SRRs, and different number of particles. The simulated experiments illustrate that this tracking approach is able to successfully track the source positions and detect their activities. Also, utilising the DUET-GCC based TDOA measurements is more efficient than using the PHAT-GCC based TDOA

measurements. Further real room recording experiment shows that this approach works well in the real reverberant environment.

It is worth pointing out that all the tracking approaches developed in this thesis are implemented in the real audio lab environment. These real room experiments, as well as the simulated experiments fully illustrate that the approaches developed in this thesis are very efficient in tracking the sources as well as dealing with the adverse environments.

7.1.2 Limitation of the work

From the signal processing point of view, the major limitations of the work in this thesis are threefold.

- The measures to evaluate the TDOA performance in Chapter 4 are the probabilities of detection and false alarm. Although the tracking performance of a tracker is mostly determined by these two parameters, the affect from the divergence between the TDOAs and the ground truth is still unknown. In the single source scenario, RMSE can be employed to measure the distance between the TDOA estimate and the actual one easily. However, for a set of TDOAs generated by multiple sources, the TDOA measurement error is difficult to tell. Fig. 7.1 shows this problem. Assuming that there are two sources, ideally the number of TDOAs should be two as well. By using the reverberant measurement model described in Section 5.2.3 on page 122, usually more peaks in the GCC function or DUET 2-D histogram are picked. The cardinality of the final TDOA estimates are thus larger than that of the ground truth. How to evaluate the TDOA measurement error is thus a problem and how will the divergence error effect the tracking performance needs to be investigated.

- In Chapter 5, the incorporation of the amplitude information as the hypothesis prior is based on the proportion of the amplitude among all the amplitudes of TDOAs. It is a great interest to thoroughly investigate the amplitude information from the sources and clutter respectively. The probabilistic models may be employed to model the prior of the hypothesis. The amplitude information is thus able to be incorporated to build the likelihood in a full probabilistic sense.

- Another limitation of this work is that the computation complexity of the developed al-

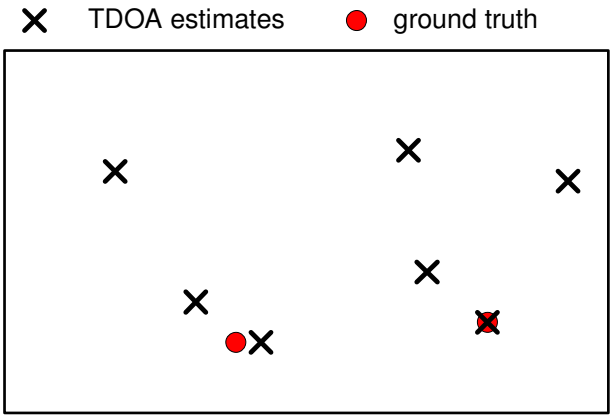


Figure 7.1: *Illustration of the TDOA estimates and ground truth from multiple sources. The cardinality of the TDOA estimates is larger than that of the ground truth. How to evaluate the TDOA measurement error is thus a problem.*

gorithms are not fully investigated. In addition, the tracking approach developed for multiple time-varying number of acoustic source tracking is only compared with the tracking approach at the measurement extraction level, but not with other trackers. The tracking algorithm proposed in [1] is also within a RFS Bayesian filtering framework. Given a specific experiment, the tracking performance comparison between these two approaches is worth being explored in future work.

7.2 Suggestions for future research

Obviously, the limitations discussed in the previous section and the future improvements mentioned in their respective chapters are interested perspectives to improve the proposed tracking approaches. Here we will suggest some directions for further research in terms of improving the solutions to the acoustic source tracking problem.

7.2.1 Joint TDOA and ILD tracking

The tracking approaches developed in this thesis only use the TDOA information from the received signal as its measurements. This is extravagant since the extracted TDOA measurements are used to replace the whole observed signals, which may contain other useful information for localising or tracking the sources. As mentioned in Section 2.3.3 on page 31, the Interaural level difference is also an important cue for acoustic source localisation problem. One natural

interest here is to develop a Bayesian framework for tracking acoustic sources exploiting the TDOA and ILD cues jointly.

However, such measurement combination is not straightforward. Although both the TDOA and ILD are highly related to the source position, and can easily be obtained, these two information extracted from the same microphone pair are highly correlated. When we fuse these two measurements, one practical question is: how much new information can be gained by adding the ILD measurements in? Authors in [126] particularly investigate this problem by comparing the source location Cramer-Rao lower bound (CRLB) using both the TDOA and gain ratio (can be regarded as ILD) measurements with that only using TDOA measurements. It shows that incorporating the energy measurement is able to reduce the estimation error when the source signal has a very narrow bandwidth or the sound propagation speech is relatively high. Since speech are wideband signals, tracking using the TDOA and ILD cues jointly is unknown and worth detailed investigations.

7.2.2 Tracking more simultaneously active sources

The maximum number of simultaneously active acoustic sources in this thesis is two. Theoretically, the DUET-GCC method is able to extract the TDOA measurements of more simultaneously active sources. However, due to reverberation and short frame length, it becomes more difficult for DUET to separate the speech mixtures as the number of the source increases.

One possible solution is to split the spectrogram into different frequency bands and then estimate the measurements at each individual frequency band. Consideration could also be given to combine other modality such as video information to track the sources.

Appendix A
Audio lab experiment details

As mentioned in Section 2.6.3, the algorithms proposed in this thesis are implemented for real recording signals in the audio lab. This appendix presents the detailed recording systems and a full illustration of the audio signal processing lab.

A.1 Recording system

The microphones are mounted on a set of T-bar stands, by which the microphone separation and height are adjustable. The microphones are omni-directional and are pressure receivers. Fig. A.1 shows the polar diagram of microphone response. The response is omni-directional at the direction range $0°$ to $180°$ under 4000Hz. This matches the omni-directional assumption in our tracking systems since only the frequencies between 0Hz to 4000Hz are considered.

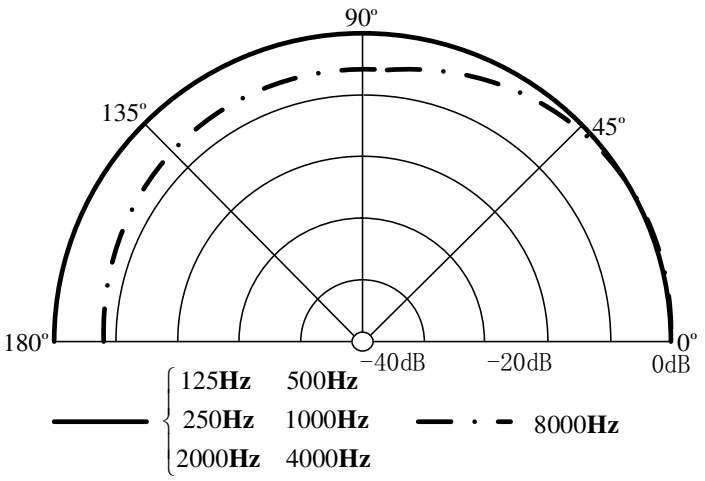


Figure A.1: *Polar diagram of the microphone response. Between the frequency band 0Hz to 4kHz which is interested in tracking problem, the microphone can be regarded as omni-directional.*

The detailed specifications of the microphones are summarised as follows:

- Transducer type: pressure receiver;

- polar pattern: omni-directional;
- Frequency range: 20 to 20KHz;
- Dynamic range (three options):
    1. 0 dB position: 129 dB,
    2. +6 dB position: 121 dB,
    3. -10 dB position: 127 dB;
- SNR (three options):
    1. 0 dB position: 83 dB,
    2. +6 dB position: 81 dB,
    3. -10 dB position: 77 dB;
- Operating temperature: -20 °C to +60 °C.

The 0 dB position is used for all the microphones. The received speech signals will then pass an A/D converter and will be recorded in the workstation.

The acoustic source used for all the recordings is an omnidirectional speaker amounted on a small trolley, as shown, in Fig. A.2. The source signal is picked from the TIMIT database [103]. Compared to a real person speaker, using this computer speaker has advantages that the experiments will be easier to reproduce, and the volume of the source signal is easier to control.

A.2 Lab details and ground truth

Figure 2.11 on page 46 presents the dimension of the audio lab and the experiment setup. The lab is with carpet floor and concrete block walls and ceiling, and the glass windows around the walls are covered by hard cardboard (with a thickness of 0.4cm roughly). The reverberation time in the room is 0.836s according to our estimation in Section 4.6.1, on page 110. This lab is thus very reverberant for the tracking problem. To reduce the reverberation, two doors behind the first microphone array are opened when recording the signals.

The motion trajectories of the source are the white lines marked in Fig. A.2. Unfortunately, the ground truth is unknown here due to a lack of motion capture system. However, it can

Figure A.2: *Audio lab environment and experiment setup. The acoustic source is amounted on a small trolley.*

be roughly estimated by assuming that the movement of the source is with an even speed and follows the marked trajectory strictly. The signal for multiple simultaneously active sources is generated by overlapping the recorded signal from two single sources via post-processing.

Appendix B
Definition of the SNR and SRR

Suppose the acoustic signal is processed frame by frame, and with a frame length of T. Given a speech frame

$$\mathbf{s}(k) = [s(kT), s(kT+1), \cdots, s(kT+T-1)]\,, \tag{B.1}$$

and Gaussian background noise

$$\mathbf{n}(k) = [n(kT), n(kT+1), \cdots, n(kT+T-1)]\,. \tag{B.2}$$

where $n(\cdot) \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian process with zero mean and variance σ^2. The acoustic propagation environment and the receivers are supposed to follow the assumptions in Section 2.1, on page 13.

B.1 Signal-to-noise ratio

The signal-to-noise ratio (SNR) in dB is defined as the logarithmic decibel scale of the power ratio between the speech signal and the background noise, given by

$$\text{SNR}_{\text{dB}} = 10\log_{10}\frac{P_{\text{signal}}}{P_{\text{noise}}}. \tag{B.3}$$

Normally the power of a signal is substituted by the square of root-mean-square (RMS) amplitude A, given as

$$\begin{aligned}\text{SNR}_{\text{dB}} &= 10\log_{10}\frac{A^2_{\text{signal}}}{A^2_{\text{noise}}}\\ &= 20\log_{10}\frac{A_{\text{signal}}}{A_{\text{noise}}},\end{aligned} \tag{B.4}$$

with

$$A_{\text{signal}} = \sqrt{\frac{\sum_{i=1}^{T} s(kT+i)^2}{T}} \qquad \text{and} \qquad A_{\text{noise}} = \sqrt{\frac{\sum_{i=1}^{T} n(kT+i)^2}{T}}. \tag{B.5}$$

Since the background noise is assumed to be Gaussian white noise, its RMS amplitude actually is its standard variance. The SNR can also be expressed as

$$\text{SNR}_{\text{dB}} = 20\log_{10}\frac{A_{\text{signal}}}{\sigma}. \tag{B.6}$$

B.2 Signal-to-reverberation ratio

While the SNR is easily defined, the signal-to-reverberation (SRR) ratio is in a more complex form since the reverberation part is determined by a set of room parameters. To simplify the derivation, we start from the direct path energy density [1]

$$P_{\text{direct}} = \frac{I}{c}, \tag{B.7}$$

where I is the wave intensity transported when the wave travels a certain distance c per second. Note that here the energy density P denote the energy per unit volume. Suppose that the acoustic power in the origin is $\bar{E} = \sum_{i=1}^{T} s(kT+i)^2$. After the wave travels a distance r, the wave intensity should be

$$I = \frac{\bar{E}}{4\pi r^2}. \tag{B.8}$$

The direct sound energy density is thus

$$P_{\text{direct}} = \frac{\bar{E}}{4\pi c r^2}. \tag{B.9}$$

Consider a room with following parameters: room volume V, absorption area $\mathcal{A}$, and average absorption rate α. After each reflection, the acoustic wave reduces its energy by $(1-\alpha)$ times of its original energy. The times the wave undergoes per second in the room is

$$N = \frac{c\mathcal{A}}{4V}. \tag{B.10}$$

The energy the acoustic wave lost after N reflections is thus

$$(1-\alpha)^{Nt} = \exp\left(Nt\ln(1-\alpha)\right), \tag{B.11}$$

[1] Please be aware that the original work of this part can be found in Chapter 5 of Kuttruff's book [20].

and the total energy remained is

$$E(t) = E_0 \exp\left(\frac{ct\mathcal{A}}{4V}\ln(1-\alpha)\right), \tag{B.12}$$

where E_0 is the certain energy created at time instant $t = 0$.

The energy lost per second can be obtained by differentiating equation (B.12) with respect to time, given as

$$\frac{dE(t)}{dt} = E(t)\frac{c\mathcal{A}}{4V}\ln(1-\alpha). \tag{B.13}$$

Considering the energy of the acoustic source, the total energy after the energy lost is

$$\frac{dE'(t)}{dt} = E(t)\frac{c\mathcal{A}}{4V}\ln(1-\alpha) + \bar{E}. \tag{B.14}$$

The left hand side will be 0 under steady state conditions, and the energy density can be obtained as

$$\begin{aligned}\frac{E(t)}{V} &= \frac{-4\bar{E}(t)}{c\mathcal{A}\ln(1-\bar{\alpha})} \\ &= \frac{\bar{E}}{VE_0}\int_0^\infty E(t)dt.\end{aligned} \tag{B.15}$$

To obtain the energy density contributed by the reverberation part, substitute the lower limit of integration by $1/N$, which is the time for the first reflections,

$$\begin{aligned}P_{\text{reverb}} &= \frac{\bar{E}}{VE_0}\int_{1/N}^\infty E(t)dt \\ &= -\frac{4\bar{E}(1-\alpha)}{c\mathcal{A}\ln(1-\alpha)} \\ &\doteq \frac{4\bar{E}(1-\alpha)}{c\alpha\mathcal{A}}.\end{aligned} \tag{B.16}$$

It finally leads to the expression of the SRR in logarithmic decibel scale, stated as

$$\begin{aligned}\text{SRR}_{\text{dB}} &= 10\log\frac{P_{\text{direct}}}{P_{\text{reverb}}} \\ &= 10\log\frac{\mathcal{A}\alpha}{16\pi r^2(1-\alpha)}.\end{aligned} \tag{B.17}$$

Since the absorption coefficient holds a relation with the reflection coefficient as $\alpha = 1 - \rho^2$,

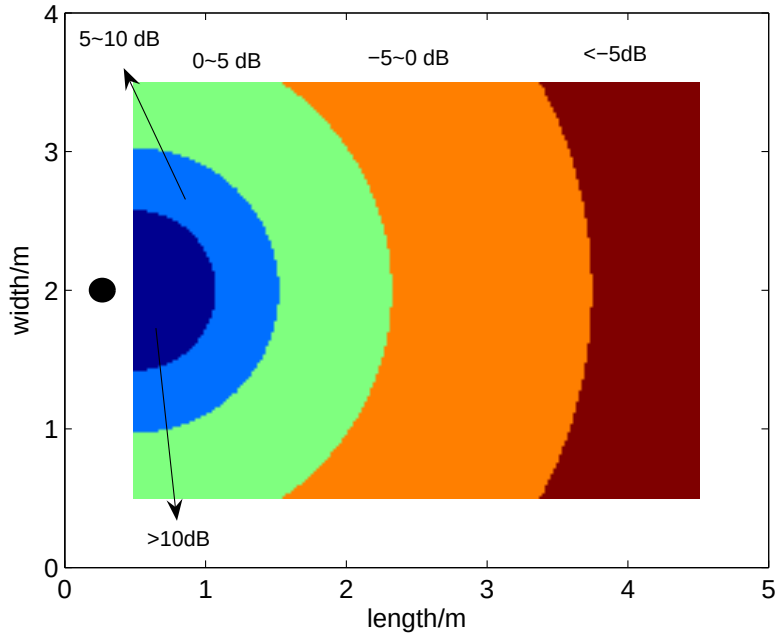


Figure B.1: *A sketch of the SRR distribution for the simulated room environment. The dark circle denote the position of the microphone receiver. The wall reflection coefficients for this plot is $\rho = 0.6$. Even in the same room environment, the SRR varies hugely; smaller than -5dB at the far end and larger than 10dB at the close end.*

the SRR can also be written as

$$\mathrm{SRR}_{\mathrm{dB}} = 10 \log \frac{\mathcal{A}(1-\rho^2)}{16\pi r^2 \rho^2}. \tag{B.18}$$

Given a room environment, the SRR of an acoustic source is actually determined by its distance to the microphone receiver r; SRR increases quadratically as the decreasing of the distance between the source and the sensor. This is also a reason why localising a dynamical source is difficult: the SRRs at different position vary hugely.

Figure B.1 gives an example of SRR distribution in the simulated room environment. The position of microphone receiver is $(0.3\ 2.0)$m. The wall reflection coefficients for calculating the SRR is $\rho = 0.6$. Here the SRR in a two dimensional plane is considered. The further the source signal is far away from the microphone, the smaller the SRR will be.

The simulated received signals with several different levels of SNRs and SRRs are also presented here. The original speech signal is shown in Fig. B.2. Fig. B.3 presents this speech

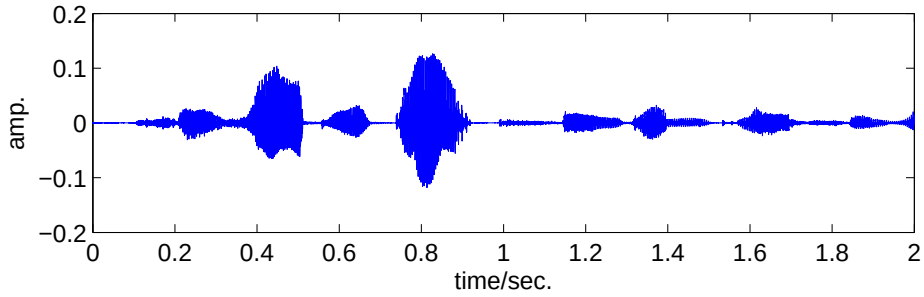


Figure B.2: *Original speech signal.*

signal corrupted by a white Gaussian noise with different levels of 10dB, 0dB, -5dB respectively. To keep the consistency of the SNR, the SNR is calculated based on each speech frame (which is 1024 samples in this thesis). The simulated signals with different SRRs, 10dB, 0dB and -10dB are presented in Fig. B.4. The parameters (source positions and wall reflection coefficients) for calculating different SRRs are described in Table 4.4, on page 105. Both the SNR and SRR have an obvious affect on the received signal. In the low SNR or SRR environments, the source signal is deteriorated significantly.

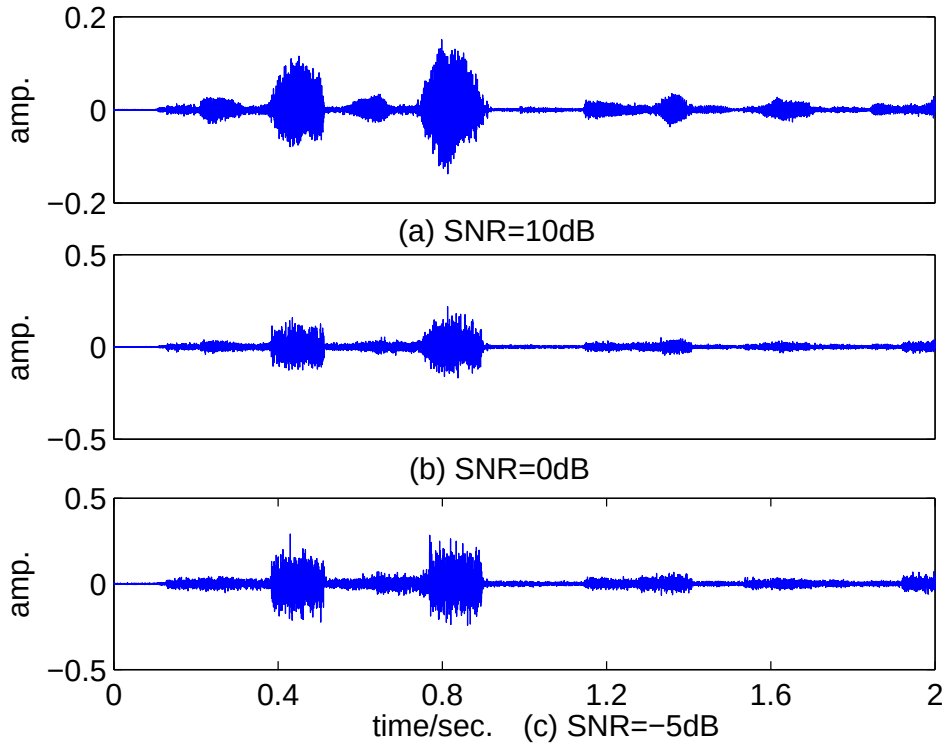


Figure B.3: *Speech signal corrupted by different level of White Gaussian noise.*

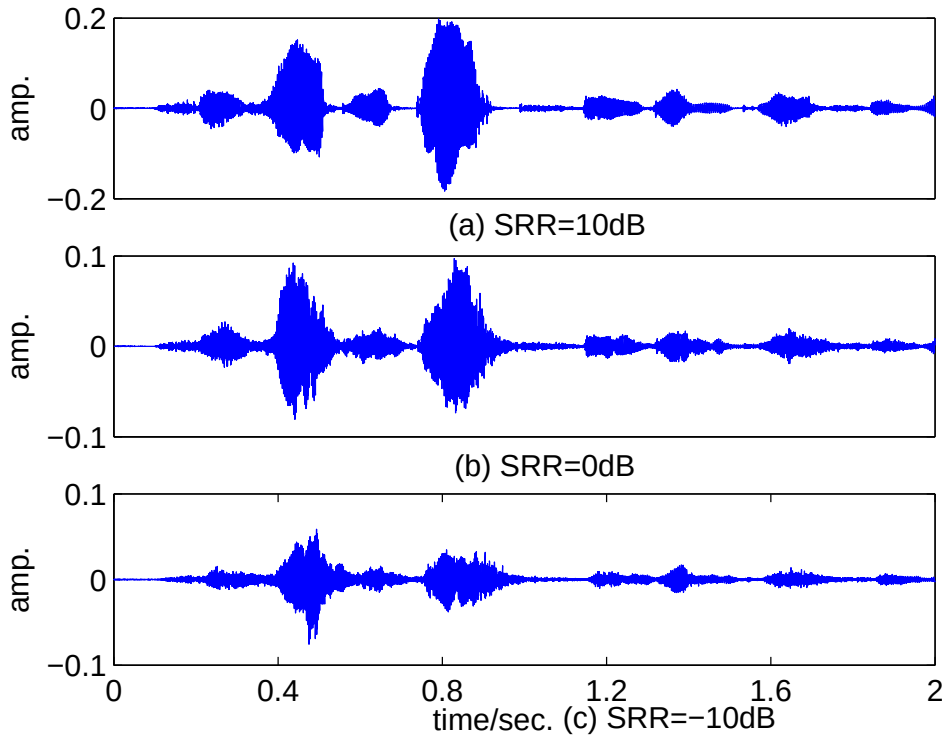


Figure B.4: *Speech signal corrupted by different level of reverberation.*

Appendix C
DUET-GCC and PHAT-GCC fails in the strong reverberant environment

For simplicity, only the first order of reflection is employed in the following derivation. Consider a noise-free scenario, a reflection component of the source signal can be regarded as another coherent source which is emitted at an imaging position. Suppose the original source signal and the imaging source signal are of the form in the frequency domain $S(\omega)$ and $S_{\text{imag}}(\omega)$ respectively. For the impulse response, suppose $H_1(\omega)$ and $H_2(\omega)$ are the direct path response to the two microphones separately, and $H_{11}(\omega)$ and $H_{22}(\omega)$ are the corresponding first order reflection response. The signals received across the ℓth microphone pair are thus

$$\begin{bmatrix} Z_{\ell,1}(\omega) \\ Z_{\ell,2}(\omega) \end{bmatrix} = \begin{bmatrix} H_1(\omega) & H_{11}(\omega) \\ H_2(\omega) & H_{22}(\omega) \end{bmatrix} \begin{bmatrix} S(\omega) \\ S_{\text{imag}}(\omega) \end{bmatrix}. \tag{C.1}$$

since these two source signals are coherent (actually the same one), we have

$$S_{\text{imag}}(\omega) = S(\omega). \tag{C.2}$$

Further, for the impulse response, we have following relationship $H_{11}(\omega) = \alpha_1 e^{-j\omega\tau_1} H_1(\omega)$ and $H_{22}(\omega) = \alpha_2 e^{-j\omega\tau_2} H_2(\omega)$. The signals received are thus

$$\begin{aligned} Z_{\ell,1}(\omega) &= (1 + \alpha_1 e^{-j\omega\tau_1}) H_1(\omega) S(\omega) \\ Z_{\ell,2}(\omega) &= (1 + \alpha_2 e^{-j\omega\tau_2}) H_2(\omega) S(\omega). \end{aligned} \tag{C.3}$$

Substitute the expression (C.3) into the PHAT-GCC function (4.9) on page 83, the function

turns out

$$\begin{aligned} R_\ell(\tau) &= \int_\Omega \frac{Z_{\ell,1}(\omega)Z^*_{\ell,2}(\omega)}{|Z_{\ell,1}(\omega)Z^*_{\ell,2}(\omega)|} e^{j\omega\tau} d\omega \\ &= \int_\Omega \frac{(1+\alpha_1 e^{-j\omega\tau_1})(1+\alpha_2 e^{-j\omega\tau_2})^* H_1(\omega)H_2^*(\omega)}{|(1+\alpha_1 e^{-j\omega\tau_1})(1+\alpha_2 e^{-j\omega\tau_2})^* H_1(\omega)H_2^*(\omega)|} e^{j\omega\tau-\tau} d\omega \\ &= \int_\Omega \frac{(1+\alpha_1 e^{-j\omega\tau_1})(1+\alpha_2 e^{-j\omega\tau_2})^*}{|(1+\alpha_1 e^{-j\omega\tau_1})(1+\alpha_2 e^{-j\omega\tau_2})^*|} e^{j\omega(\tau-\tau_\ell)} d\omega \\ &\neq \delta_{\tau_\ell}(\tau). \end{aligned} \tag{C.4}$$

If the reverberant components can not be ignored, the position of the highest peak in PHAT-GCC function will be meaningless. This is also the case for coherent source signals. Also, substitute the expression (C.3) into equation (4.21) on page 95, the ratio of two received signals in DUET formulation becomes

$$\begin{aligned} R_\ell(\omega) &= Z_{\ell,1}(\omega)/Z_{\ell,2}(\omega) \\ &= \frac{(1+\alpha_1 e^{-j\omega\tau_1})H_1(\omega)}{(1+\alpha_2 e^{-j\omega\tau_2})H_2(\omega)} \\ &= \frac{(1+\alpha_1 e^{-j\omega\tau_1})|H_1(\omega)|}{(1+\alpha_2 e^{-j\omega\tau_2})|H_2(\omega)|} e^{-j\omega\tau_\ell} \\ &\neq \frac{a_{\ell,1}}{a_{\ell,2}} e^{-j\omega\tau_\ell}. \end{aligned} \tag{C.5}$$

This means that both the spectrogram ration between two received signals no longer carries the gain-ratio and the time-delay information generated by the source, and thus the 2-D histogram cannot be formed.

According to the derivations in (C.4) and (C.5), the DUET-GCC and PHAT-GCC approaches can only present meaningful estimation if

$$\alpha_i \ll 1 \qquad \forall \quad i = 1, 2. \tag{C.6}$$

which is the case of the anechoic and low reverberant environments where the reverberant components can be almost ignored.

References

[1] 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," *Signal Processing, IEEE Trans. on*, vol. 54, no. 9, pp. 3291–3304, 2006.

[2] H. F. Silverman, W. R. P. III, and J. L. Flanagan, "The huge microphone array," *IEEE Concurrency*, vol. 6(4), pp. 36–46, 1998.

[3] H. F. Silverman, W. R. P. III, and J. L. Flanagan, "The huge microphone array," *IEEE Concurrency*, vol. 7(1), pp. 32–47, 1999.

[4] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "Loud: a 1020-node microphone array and acoustic beamformer," *In international Congress on Sound and Vibration*, 2007.

[5] R. J. Mailloux, *Phased Array Antenna Handbook.* Artech House, Boston, 1994.

[6] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE Trans. on*, vol. 11, pp. 826–836, Nov. 2003.

[7] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, Apr. 1997.

[8] U. Klee, Tobias, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[9] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 59625, 17 pages, 2006.

[10] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Trans. on*, vol. 50, pp. 174–188, Feb. 2002.

[11] D. Arnaud, F. N. de, and G. Neil, eds., *Sequential Monte Carlo Methods in Practice*. Springer-Verlag New York, 2001.

[12] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*, vol. 5, pp. 3021–3024, May 2001.

[13] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*, vol. 2, pp. 1777 – 1780, May 2002.

[14] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *Speech and Audio Processing, IEEE Trans. on*, vol. 5, pp. 45–50, Jan. 1997.

[15] B.-N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*, vol. 2, pp. 357–360, May 2004.

[16] B.-N. Vo, W.-K. Ma, and S. Singh, "Localizing an unknown time-varying number of speakers: a bayesian random finite set approach," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 1073–1076, Mar. 18–23, 2005.

[17] M. Fallon and S. Godsill, "Multi-target acoustic source tracking with an unknown and time varying number of targets," in *Hands-free Speech Communication and Microphone Arrays*, (Italy), pp. 77–80, May 2008.

[18] M. Fallon, *Acoustic Source Tracking using Sequential Monte Carlo*. Darwin College, University of Cambridge, 2008.

[19] M. R. Morelande, C. M. Kreucher, and K. Kastella, "A bayesian approach to multiple target detection and tracking," *Signal Processing, IEEE Trans. on*, vol. 55, pp. 1589–1604, May 2007.

[20] H. Kuttruff, *Room Acoustic*. London, Applied Science Publishers, 1973.

[21] L. E. kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics (4th Edition)*. John Wiley and Sons, 2000.

[22] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Jul. 1979.

[23] Y. Haneda, "Common acoustic pole and zero modeling of room transfer fucntions," *Speech Audio Process, IEEE Trans. on*, vol. 2, pp. 320–328, Apr. 1994.

[24] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, 1979.

[25] S. M. Kay, *Foundamentals of Statistical Signal Processing*. Prentice Hall, Englewood Cliffs, 1993.

[26] T. M. N. Strobel and R. Rabenstein, "Speaker localization using steered filtered-and-sum beamformers," *Proc. Erlangen Workshop on Vision, Modeling, and Visualization Erlangen*, vol. 11, pp. 195–202, 1999.

[27] M. Brandstein and D. Ward, *Microphone Arrays. Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[28] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–11, 2007.

[29] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP Journal on Applied Signal Processing*, vol. 2007, pp. 1–8, 2007.

[30] F. c. M. Jean-Marc Valin and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, pp. 216–228, oct. 2007.

[31] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Trans. on*, vol. 24, pp. 320–327, Aug. 1976.

[32] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *Acoustic, Speech and Signal Processing, IEEE Trans. on*, vol. 32, pp. 280–285, Dec. 1984.

[33] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using csp analysis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 921–924, May 7–10, 1996.

[34] P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 231–234, Apr. 21–24, 1997.

[35] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *Speech Audio Process, IEEE Trans. on*, vol. 5, pp. 288–292, May 1997.

[36] B. Yegnanarayana, S. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *Speech and Audio Processing, IEEE Trans. on*, vol. 13, pp. 1110–1118, Nov. 2005.

[37] V. Raykar, B. Yegnanarayana, S. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *Speech and Audio Processing, IEEE Trans. on*, vol. 13, pp. 751–761, Sept. 2005.

[38] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *Speech and Audio Processing, IEEE Trans. on*, vol. 4, pp. 148–152, March 1996.

[39] J. M. Tribolet, "A new phase unwrapping algorithm," *Acoustic, Speech and Signal Processing, IEEE Trans. on*, vol. 25, pp. 170–177, 1977.

[40] M. S. Brandstein, J. E. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone pair," *Computer, Speech, and Language*, vol. 9, pp. 153–169, Apr. 1995.

[41] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *Speech and Audio Processing, IEEE Trans. on*, vol. 12, pp. 520–529, Sept. 2004.

[42] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, pp. 384–391, Jan. 2000.

[43] I. O. L. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of IEEE*, vol. 60, pp. 926–935, Oct. 1972.

[44] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*, pp. 937–940, March 1999.

[45] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, pp. 1110–1124, 2003.

[46] P. Damaske, *Acoustic and Hearing*. Springer-Verlag Berlin, Heidelberg, 2008.

[47] D. Li and Y. H. Hu, "Energy-based collaborative source localization using acoustic microsensor array," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 321–337, 2003.

[48] X. Sheng and Y. H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *Signal Processing, IEEE Trans. on*, vol. 53, pp. 44–53, Jan. 2005.

[49] S. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 4, pp. 1109–1112, 18-23 March 2005.

[50] Z. C. W. Cui and J. Wei, "Dual-microphone source location method in 2-d space," in *Acoustics, Speech and Signal Processing. Proceedings of the IEEE International Conference on*, pp. 845–848, May 2006.

[51] Z. Liu, Z. Zhang, L. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays.," in *Acoustics, Speech and Signal Processing. Proceedings of the IEEE International Conference on*, vol. 2, pp. 761–764, May 2007.

[52] C. Faller and J. Merimaa, "Source localization in complex listening situations: slection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075–3089, Nov. 2004.

[53] D. Li and S. E. Levinson, "A Bayes-rule based hierarchical system for binaural sound source localization.," in *Acoustics, Speech, and Signal Processing, Proceedings of the IEEE International Conference on*, vol. 5, pp. 521–524, 2003.

[54] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Trans. on*, vol. 52, pp. 1830–1847, July 2004.

[55] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 728–739, May 2008.

[56] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ild and itd," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 68–77, Jan. 2010.

[57] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "sequential monte carlo fusion of sound and vision for speaker tracking," in *Proceedings of International Conference on Computer Vision*, pp. 741–746, 2001.

[58] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *IEEE International Conference on Image Processing*, pp. 25–28, 2003.

[59] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," in *Proceedings of IEEE*, vol. 92, pp. 485–494, March 2004.

[60] H. Asoh, F. Asano, T. Yoshimura, K. Yamamoto, Y. Motomura, N. Ichimura, I. Hara, and J. Ogata, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Proceedings of International Conference on Information Fusion*, pp. 805–812, 2004.

[61] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *Audio, Speech and Language Processing, IEEE Trans. on*, vol. 15, pp. 601–616, Feb. 2007.

[62] A. M. Johansson, E. A. Lehmann, and S. Nordholm, "Real-time implementation of a particle filter with integrated voice activity detector for acoustic speaker tracking," in *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, pp. 1004–1007, Dec. 4–7, 2006.

[63] S. S. Blackman, *Multiple-target Tracking with Radar Applications*. Artech House, 1986.

[64] L. D. Stone, C. A. Barlow, and T. L. Corwin., *Bayesian Multiple Target Tracking.* Artech House, 1999.

[65] J. L. David L. Hall, ed., *Handbook of multisensor data fusion*. CRC Press, 2001.

[66] R. P. S. Mahler, *Statistical Multisource-Multitarget Information Fusion.* Artech House, 2007.

[67] S. Haykin, *Adaptive Filter Theory (4th Edition)*. Prentice Hall, 2001.

[68] D. Simon, *Optimal State Estimation*. John Wiley and Sons, 2006.

[69] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 371–374, Apr. 21–24, 1997.

[70] I. Potamitis and G. Kokkinakis, "Speech separation of multiple moving speakers using multisensor multitarget techniques," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 37, no. 1, pp. 72–81, 2007.

[71] G. Tobias, K. Ulrich, M. J. W., I. Shajith, W. Matthias, and F. Christian, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *Proceedings of the 9th International Conference on Spoken Language Processing*, vol. 5, pp. 2594–2597, 2006.

[72] N. T. Pham, W. Huang, and S. H. Ong, "Tracking multiple speakers using CPHD filter," in *Proceedings of the 15th international conference on Multimedia*, no. 529–532, 2007.

[73] S. S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking System*. Artech House, 1999.

[74] X. Rong Li and V. P. Jilkov, "Survey of maneuvering target tracking. part 1: Dynamic models," *Aerospace and electronic systems, IEEE Trans. on*, vol. 39, pp. 1333–1364, Oct. 2003.

[75] E. A. 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 *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 98–101, Oct. 21–24, 2007.

[76] A. M. Johansson and E. A. Lehmann, "Evolutionary optimization of dynamics models in sequential monte carlo target tracking," *EEE Transactions on Evolutionary Computation*, vol. 13, pp. 879–894, August 2009.

[77] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1527–1529, Apr. 1986.

[78] K. Aspelin, "Establishing pedestrian walking speeds," Portland State University. 2005. www.westernite.org.

[79] T. R. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.

[80] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic press, 1970.

[81] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition.," in *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[82] R. E. Kalman, "A new approach to linear filtering and presiction problems," *Transactions of the AMSE-Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[83] H. Tanizaki and R. S. Mariano, "Nonlinear filters based on taylor series expansions," *Communications in Statistics - Theory and Methods*, vol. 25, no. 6, pp. 1261–1282, 1996.

[84] H. Tanizaki, *Nonlinear Filters: Estimation and Applications*. New York: Springer-Verlag, 2nd ed., 1996.

[85] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[86] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[87] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, pp. 1032–1044, 1998.

[88] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[89] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and bayesian missing data problems," *Journal of the American Statistical Association*, vol. 89, pp. 278–288, 1994.

[90] A. F. M. Smith and A. E. Gelfand, "Bayesian statistics without tears: A sampling-resampling perspective," *The American Statistician*, vol. 46, no. 2, pp. 84–88, 1992.

[91] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. London: Chapman & Hall, 1994.

[92] G. Casella and C. P. Robert, "Rao-blackwellization of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[93] J. F. G. D. Freitas, M. Niranjan, A. H. Gee, and A. Doucet, "Sequential monte carlo methods to train neural network models," *Neural Computation*, vol. 12, no. 4, pp. 955 –993, 2000.

[94] A. Doucet, A. Logothetis, and V. Krishnamurthy, "Stochastic sampling algorithms for state estimation of jump markov linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 1, pp. 188–202, 2000.

[95] K. P. M. S. J. R. Arnaud Doucet, Nando de Freitas, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 176–183, 2000.

[96] R. Mahler, "Multi-target bayes filtering via first-order multi-target moments," *Aerospace and Electronic Systems, IEEE Trans. on*, vol. 39, pp. 1152–1178, Oct. 2003.

[97] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes (2nd Edition)*. Springer-Verlag New York, Berlin, Heidelberg, 2003.

[98] B.-N. Vo, S. Singh, and A. Doucet, "Sequential monte carlo methods for multitarget filtering with random finite sets," *Aerospace and Electronic Systems, IEEE Trans. on*, vol. 41, pp. 1224–1245, Oct. 2005.

[99] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," vol. 11, no. 6, pp. 791–803, 2003.

[100] J. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *Acoustic, Speech and Signal Processing, IEEE Trans. on*, vol. 30, pp. 998–1003, Dec. 1982.

[101] J. Chen, J. Benesty, and Y. A. Huang, "Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments," *EURASIP Journal on Applied Signal Processing*, no. 1, pp. 25–36, 2005.

[102] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, pp. 177–204, Jan. 2005.

[103] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993.

[104] M. Jian, A. C. Kot, and M. H. Er, "Performance analysis of time delay estimation in a multi-path environment," *in Proc. IEEE International conference of digital signal processing*, vol. 2, pp. 919–922, March 1997.

[105] D. M. Green and J. M. Swets, *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons Inc., 1966.

[106] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[107] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 375–378, 1997.

[108] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–19, 2006.

[109] S. Makino, T. W. Lee, and H. Sawada, eds., *Blind Speech Separation*. Springer, 2007.

[110] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. IV–IV, 2006.

[111] M. I. Mandel and D. P. W. Ellis, "Em localization and separation using interaural level and phase cues," in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on*, pp. 275–278, 21-24 Oct. 2007.

[112] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged mutliple sensors," *Singal Processing*, vol. 87, pp. 1833–1847, 2007.

[113] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan, "The unscented particle filter," *Technical Report*, Cambridge University Engineering Department, Aug. 2000.

[114] N. G. Branko Ristic, Sanjeev Arulampalam, *Beyond the Kalman filter: Particle Filters for Tracking Applications*. Artech House, 2004.

[115] E. A. Lehmann, *particle filtering methods for acoustic source localisation and tracking*. PhD thesis, The Australian National University, 2004.

[116] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localisation in a room environment," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2008, pp. 1–14, 2007.

[117] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized monte carlo data association for multiple target tracking," *Proceedings of the Seventh International Conference on Information Fusion*, vol. 1, pp. 583–590, 2004.

[118] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, pp. 2–15, January 2007.

[119] M. Vihola, "Rao-blackwellised particle filtering in random set multitarget tracking," *Aerospace and Electronic Systems, IEEE Trans. on*, vol. 43, pp. 689–705, April 2007.

[120] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying rao-blackwellised monte carlo data association," in *Proc. IEEE 15th Workshop on Statistical Signal Processing*, pp. 253–256, Aug. 2009.

[121] X. Zhong and J. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 293–296, 2008.

[122] R. W. Sittler, "An optimal data association problem in surveillance theory," *Military Electronics, IEEE Transactions on*, vol. 8, pp. 125–139, Apr. 1964.

[123] J. Hoffman and R. Mahler, "Multitarget miss distance and its applications," in *Proc. Fifth International Conference on Information Fusion*, vol. 1, pp. 149–155, 2002.

[124] J. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *Systems, Man and Cybernetics, IEEE Trans. on, Part A: System and Humans*, vol. 34, no. 3, pp. 327–336, 2004.

[125] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," vol. 56, pp. 3447–3457, Aug. 2008.

[126] K. C. Ho and M. Sun, "Passive source localization using time differences of arrival and gain ratios of arrival," *Signal Processing, IEEE Trans. on*, vol. 56, pp. 464–477, Feb. 2008.

Appendix D
Publications

D.1 Conference papers

- X. Zhong and James R. Hopgood, "Time-frequency Masking based Multiple Acoustic Source Tracking Applying Rao-Blackwellised Monte Carlo Data Association", *In Proceedings of Statistical Signal Processing, IEEE Workshop on*, pp. 253-256, 2009.

- X. Zhong and James R. Hopgood, "Nonconcurrent Multiple Speakers Tracking based on Extended Kalman Particle Filter", *In Proceedings of Acoustic, Speech and Signal Processing, IEEE International Conference on*, pp. 293-296, 2008.

- X. Zhong and James R. Hopgood, "Acoustic source tracking using joint PHAT-GCC TDOA and amplitude measurements", **To be submitted** to *Acoustic, Speech and Signal Processing, IEEE International Conference 2011*.

D.2 To be submitted Journal papers

- X. Zhong and James R. Hopgood, "Tracking a Time-varying Number of Acoustic Sources Applying a Random Finite Set Rao-Blackwellised Particle Filtering", **To be submitted** to *Audio, Speech, and Language Processing, IEEE Transactions on*.

- X. Zhong and James R. Hopgood, "Multiple Acoustic Source Tracking: a Joint Time-delay and signal strength Perspective", **To be submitted** to *Audio, Speech, and Language Processing, IEEE Transactions on*.

NONCONCURRENT MULTIPLE SPEAKERS TRACKING BASED ON EXTENDED KALMAN PARTICLE FILTER

Xionghu Zhong and James R. Hopgood

Institute for Digital Communications,
School of Engineering and Electronics, University of Edinburgh, UK
x.zhong@ed.ac.uk and James.Hopgood@ed.ac.uk

ABSTRACT

Acoustic reverberation introduces multipath components into an audio signal, and therefore changes the source signal statistical properties. This causes problems for source localisation and tracking since reverberation generates spurious peaks in the time delay functions, and makes the subsequent location estimator hard to track the motion trajectory. Previous time delay based tracking methods, such as the extended Kalman filter and the particle filter, are sensitive to reverberation and are unable to follow sharp changes in the source positions. In this paper, the extended Kalman filter and the particle filter are combined to solve this problem. One of the advantages of this approach is that the optimal importance function can be obtained after extended Kalman filtering. Thus, the position samples are distributed in a more accurate area than using a prior importance function. Experiment results show that the proposed algorithm outperforms the sequential importance resampling particle filter by reducing the estimation error and following the switch of speakers quickly under a moderate reverberant environment (reverberation time $T_{60} < 0.3s$).

Index Terms— source tracking, reverberation, particle filter, extended Kalman filter

1. INTRODUCTION

Locating and tracking an acoustic source in a reverberant environment is an increasingly important research area in many applications such as teleconferencing, multimedia, hearing aids and hands-free teleconferencing systems. One popular way for this problem is the so-called indirect method, wherein the time difference of arrival (TDOA) of microphone pairs is estimated by, for example, employing generalised cross-correlation (GCC) function [1] or adaptive eigenvalue decomposition (AED) algorithm [2]. The TDOA is then used to triangulate the position using a maximum likelihood criterion [3]. This triangulation can also be achieved by a position estimator like the Kalman filter [4] or linear intersection algorithm [5]. If there is merely a time delay between the received signals and the TDOA estimates have a Gaussian distribution, and traditional indirect methods are able to track the source correctly.

However, the presence of reverberation and different kinds of noise in real-life often violate these assumptions. Thus, the performance of these TDOA estimation methods is seriously deteriorated. Recently, particle filtering is introduced into the acoustic source tracking problem to reduce the errors brought by false time delay estimates caused by the multipath reverberant components [6] and [7]. It is assumed that the received signal can be modeled by a free-field model, in which the reverberant signal is separated into direct path and multipath components. The later part is regarded as a noise term. The motion model of the speaker is then defined, and the likelihood function is constructed based on the assumption of a Gaussian distribution. Finally, the posterior distribution of the location is estimated using a particle filter. A full description of this method can be found in [6] and [7]. This sequential importance resampling (SIR) particle filter (PF) suffers from a tracking lag or even a track loss in following a sharp change of the position, which is a common case for nonconcurrent multiple speakers tracking.

The extended Kalman filter (EKF) is used in acoustic source localisation and tracking in [4]. The speaker's position is updated employing an EKF, wherein the observation and the states are associated with the TDOAs and the speaker's position separately. The results in [4] reveal that the EKF provides source accuracy superior to the linear intersection techniques. However, one drawback is that the EKF can't cope with the reverberant environment well. In this paper, we combine the particle filter and extended Kalman filter to room acoustic source tracking problem. The combined algorithm is an extended Kalman particle filter (EKPF) [8], through which the optimal importance function can be derived by estimating the posterior distribution of the states using an EKF. Thus during each iteration, samples are relocated with both the knowledge of the former state estimates and the current observations. These samples are more accurately distributed than using a prior importance function in [7] which only takes the past states into account. Furthermore, we can derive the variance of the Gaussian distribution in the likelihood function by one step prediction of the observation rather than empirical studies. These factors make the EKPF method more appropriate to the reverberant environments and complicated motion trajectories.

The rest of this paper is organised as follows. In section 2, a model for the reverberant signal and localisation problem is formulated. Section 3 summarizes the EKPF algorithm and exploits it for a tracking problem. The simulation experiments and the performance comparison with the SIR particle filter are described in section 4. Our conclusions are presented in section 5.

2. SIGNAL MODEL AND SOURCLOCALISATIONON

Let $\boldsymbol{p}_{m,i}, \mathbf{x}_t \in \mathbb{R}^3$ denote the position of ith microphone of mth microphone pair and the position of the source, respectively. The discrete time signal from a single source received can be modeled as

$$x_{m,i}(t) = s(t) * h(\boldsymbol{p}_{m,i}, \mathbf{x}_t) + n_{m,i}(t) \quad (1)$$

where $s(t)$ is the source signal, $h(\boldsymbol{p}_{m,i}, \mathbf{x}_t)$ is the overall impulse response cascading the room and the microphone channel response, $n_{m,i}(t)$ is additive noise which often assumed to be uncorrelated with the source signal and from different sensors, and $*$ denotes convolution. To formulate TDOA estimates, we rewrite the impulse

response in terms of the direct path and multipath components as

$$x_{m,i}(t) = \frac{1}{r_{m,i}} s(t - \tau_{m,i}) + s(t) * g(\boldsymbol{p}_{m,i}, \mathbf{x}_t) + v_{m,i}(t) \quad (2)$$

where $r_{m,i}$ is the distance between the source and microphone, $\tau_{m,i}$ is the direct path time delay, and $g(\boldsymbol{p}_{m,i}, \mathbf{x}_t)$ is the new impulse response which is defined as the original response minus the direct path component. In fact, this model is a free field model in that it regards the reverberation part as a noise term.

The signal model contains the parameter of interest, namely the time delay $\tau_{m,i}$. The time difference of arrival of the microphone pair can be expressed as

$$\tau_m(\mathbf{x}_t) = \tau_{m,1} - \tau_{m,2} = \frac{\|\mathbf{x}_t - \boldsymbol{p}_{m,1}\| - \|\mathbf{x}_t - \boldsymbol{p}_{m,2}\|}{c} \quad (3)$$

where c is the sound velocity, and $\|\cdot\|$ is the Euclidean norm denoting the distance between two positions. Given a series of time delay estimates $\hat{\tau}_m(t)$, the maxmum likelihood (ML) criterion [3] for location can be estimated as

$$\hat{\mathbf{x}}_t = \underset{\mathbf{x}_t}{\operatorname{argmin}} \sum_{m=1}^{M} \left(\hat{\tau}_m(t) - \tau_m(\mathbf{x}_t)\right)^2 \quad (4)$$

The evaluation of $\mathbf{x}_t$ at each time step involves the optimization of a non-linear function and necessitates the use of search methods, since no close form solution exists to equation (4).

3. EKPF FOR TRACKING

3.1. Extended Kalman filter

For the operation of the Kalman filter, we present the process and observation equation involved in the state space model. The process equation is governed by a random walk motion model, wherein the speaker is moving only under the control of the process noise

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta T \mathbf{v}_t \quad (5)$$

where ΔT is the time period between two neighbouring steps, and $\mathbf{v}_t$ is white noise with variance $\mathbf{Q}_t$ representing the velocity component of the motion. According to the motion model, the transition probability density function (PDF) then can be expressed as

$$p(\mathbf{x}_t^{(i)} | \mathbf{x}_{t-1}^{(i)}) = \mathcal{N}(\mathbf{x}_t^{(i)}; \mathbf{x}_{t-1}^{(i)}, \mathbf{Q}_t) \quad (6)$$

To be related with the observations, we approximate the time delay $\tau_m(\mathbf{x}_t)$ with a first-order Taylor expansion [4]. That is:

$$\tau_m(\mathbf{x}_t) = \tau_m(\mathbf{x}_{t-1}) + \mathbf{c}_m^T(t)[\mathbf{x}_t - \mathbf{x}_{t-1}] + \bar{n}_t \quad (7)$$

where superscript T denotes transposepose, $\bar{n}_t = O(\mathbf{x}_t)$ is the higher order of the time delay expansion, and $\mathbf{c}_m^T(t)$ is the coefficient vector of Taylor expansion

$$\mathbf{c}_m^T(t) = \frac{1}{c}\left[\frac{\mathbf{x}_t - \boldsymbol{p}_{m,1}}{\|\mathbf{x}_t - \boldsymbol{p}_{m,1}\|} - \frac{\mathbf{x}_t - \boldsymbol{p}_{m,2}}{\|\mathbf{x}_t - \boldsymbol{p}_{m,2}\|}\right]^T_{\mathbf{x}_t = \bar{\mathbf{x}}_{t-1}} \quad (8)$$

with $\bar{\mathbf{x}}_{t-1}$ denoting the state at the last time step. Defining

$$\mathbf{C}_t = \begin{bmatrix} \mathbf{c}_1^T(t) \\ \mathbf{c}_2^T(t) \\ \vdots \\ \mathbf{c}_M^T(t) \end{bmatrix}, \hat{\boldsymbol{\tau}}_t = \begin{bmatrix} \hat{\tau}_1(t) \\ \hat{\tau}_2(t) \\ \vdots \\ \hat{\tau}_M(t) \end{bmatrix}, \boldsymbol{\tau}(\mathbf{x}_t) = \begin{bmatrix} \tau_1(\mathbf{x}_t) \\ \tau_2(\mathbf{x}_t) \\ \vdots \\ \tau_M(\mathbf{x}_t) \end{bmatrix} \quad (9)$$

then following equation (7), define:

$$\mathbf{y}_t = \hat{\boldsymbol{\tau}}_t - \boldsymbol{\tau}(\bar{\mathbf{x}}_{t-1}) + \mathbf{C}_t \bar{\mathbf{x}}_{t-1} \quad (10)$$

such that the observation equation can be written as

$$\mathbf{y}_t \approx \mathbf{C}_t \mathbf{x}_t + \mathbf{n}_t \quad (11)$$

Here $\hat{\boldsymbol{\tau}}_t$ is estimated by GCC method, and $\mathbf{n}_t$ is the measurement noise which is assumed to be zero-mean Gaussian process with a variance of $\mathbf{R}_t$ representing the higher order expansion of the time delay vector.

The key idea of the EKF is equation (7), the implementation of a minimum square error (MMSE) estimator through Taylor series expansion of the nonlinear functions around the estimates. Regarding equations (5) and (11) as process and observation equation respectively, the Gaussian approximation of the posterior distribution of the states by EKF is easily derived as following:

$$\bar{\mathbf{x}}_{t|t-1} = \bar{\mathbf{x}}_{t-1|t-1} \quad (12a)$$

$$\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1|t-1} + \Delta T^2 \mathbf{Q}_t \quad (12b)$$

$$\bar{\mathbf{y}}_{t|t-1} = \mathbf{C}_t \bar{\mathbf{x}}_{t|t-1} \quad (12c)$$

$$\mathbf{S}_t = \mathbf{R}_t + \mathbf{C}_t \mathbf{P}_{t|t-1} \mathbf{C}_t^T \quad (12d)$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{C}_t^T \mathbf{S}_t^{-1} \quad (12e)$$

$$\bar{\mathbf{x}}_{t|t} = \bar{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \tau(\bar{\mathbf{x}}_{t|t-1})) \quad (12f)$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{C}_t \mathbf{P}_{t|t-1} \quad (12g)$$

Evidently, the filtered distribution of the states is $p(\mathbf{x}_t | \mathbf{x}_{0:t-1}\, \mathbf{y}_{1:t})$ $= \mathcal{N}(\mathbf{x}_t; \bar{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t})$. This distribution is used as the importance function in the particle filter in next subsection.

3.2. Tracking algorithm

As the introduction of the particle filter can easily be found in many open literature [8],[9], here we only summarize, without deduction, the particle filter algorithm combining with EKF to track the source trajectory.

Extended Kalman particle filter

1. Initialization: $t = 0$
   • For $i = 1, \ldots, N$, draw the position samples (particles) $\mathbf{x}_0^{(i)}$ from the prior $p(\mathbf{x}_0)$.
2. For $t = 1, 2, \ldots$
   • For $i = 1, \ldots, N$:
   Importance sampling step
   - Update the particles with the EKF according to (12).
   - Sample $\hat{\mathbf{x}}_t^{(i)} \sim q(\hat{\mathbf{x}}_t^{(i)} | \mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}) \triangleq \mathcal{N}(\hat{\mathbf{x}}_t; \bar{\mathbf{x}}_{t|t}^{(i)}, \mathbf{P}_{t|t}^{(i)})$
   - Set $\hat{\mathbf{x}}_{0:t}^{(i)} \triangleq (\mathbf{x}_{0:t-1}^{(i)}, \hat{\mathbf{x}}_t^{(i)})$ and $\hat{\mathbf{P}}_{0:t}^{(i)} \triangleq (\mathbf{P}_{0:t-1}^{(i)}, \mathbf{P}_{t|t}^{(i)})$
   - Evaluate the importance weights

$$\omega_t^{(i)} \propto \frac{p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)}) p(\hat{\mathbf{x}}_t^{(i)} | \mathbf{x}_{t-1}^{(i)})}{q(\hat{\mathbf{x}}_t^{(i)} | \mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t})} \quad (13)$$

where $q(\cdot)$ is the importance function and will be given by equation (15).

   • For $i = 1, \ldots, N$:
   (a) Normalizing the importance weight

$$\widetilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{i=1}^{N} \omega_t^{(i)}} \quad (14)$$

(b) Selection step

Multiply/Discard particles $(\hat{\mathbf{x}}_{0:t}^{(i)}, \hat{\mathbf{P}}_{0;t}^{(i)})$ with high/low importance weights $\widetilde{\omega}_t^{(i)}$.

3. Output: MSE estimate of the state $E(\mathbf{x}_t) = \sum_{i=1}^{N} \widetilde{\omega}_t^{(i)} \hat{\mathbf{x}}_t^{(i)}$.

Importance function: There are several choices for selecting the importance function (proposal distribution) [9]. In this paper, we use optimal importance function, which is conditional upon the trajectory $\mathbf{x}_{0:t-1}^{(i)}$ and the observations $\mathbf{y}_{0:t}$. This is formed as

$$q(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}) = p(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}\, \mathbf{y}_{1:t}) \tag{15}$$

where $p(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}\, \mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{x}_t^{(i)}; \bar{\mathbf{x}}_{t|t}^{(i)}, \mathbf{P}_{t|t}^{(i)})$. $\bar{x}_{t|t}^{(i)}$ and $\mathbf{P}_{t|t}^{(i)}$ are the filtered state and variance using EKF respectively.

Likelihood function: Because of reverberation and noise, the likelihood model $p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$ can no longer be expressed in a simple way. Let K be the number of potential delays obtained from the time-delay estimation function. Using one-step prediction of the observation and following the approaches used in [6], the likelihood function of the pth microphone pair can be

$$f_p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) = \sum_{\kappa=1}^{K} q_\kappa \mathcal{N}(\mathbf{y}_t^{(i)}; \mathbf{y}_{t|t-1}^{(i)}, \mathbf{S}_t^{(i)}) + q_0 \tag{16}$$

where $\mathbf{y}_{t|t-1}^{(i)}$ and $\mathbf{S}_t^{(i)}$ is the prediction mean and variance of observation respectively. $q_\kappa < 1$, $\kappa = 1, \ldots, K$ is the prior probability denoting that the κth potential time delay is associated the true position, and $q_0 < 1$ denotes the probability that none of the delays will contribute to the true source. We assume that the measurements across all microphone pairs are independent. If P sensor pairs are used, the complete likelihood function becomes

$$p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) = \prod_{p=1}^{P} f_p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) \tag{17}$$

4. SIMULATION EXPERIMENTS

In this section, the performance of the algorithm is illustrated on two typical motion trajectories; motion as a line or "switch-speaker". For the line trajectory case, there is only one speaker moving along the diagonal line, which is marked as trajectory 1 in Figure 1. The switch-speaker case involves a source change at the time center of the whole voice period; the motion orientation and the break position are denoted in trajectory 2. The length of the audio file for the both cases is $7.6s$, and the corresponding reverberant signal at each of microphones are generated using the image method [10]. In our experiment 4 microphone pairs each with a separation of $40cm$ at symmetrical position were employed. The room dimesion is $5m \times 5m \times 2.7m$ with background noise yielding a SNR level of 30dB. Different reflection coefficients are set from 0 to 0.9 with an increment of 0.1 to simulate various reverberant environments. The reverberation time T_{60} corresponding to the different reflection coefficients can be found in table 1. The audio signal is split into 120 frames for each trajectory. The whole experimental setup is depicted in Fig. 1.

Table 1. Reflection coefficient β and its corresponding reverberation time T_{60}.

β	0	0.1	0.2	0.3	0.4
T_{60}	0	0.11	0.12	0.13	0.15
β	0.5	0.6	0.7	0.8	0.9
T_{60}	0.19	0.23	0.28	0.41	0.56

Here we give the tracking results for both cases under a reflection coefficient of 0.5, that is $T_{60} = 0.19s$. The tracking algorithm is run with $N = 100$ particles, and particles are initialised around the center position of the room. q_0 was set to 0.4 according the setting in [7] . Fig. 2 shows the speech signal from the single source and the tracking result for line trajectory, which is represented by trajectory 1. Both the SIR particle filter and the EKPF track the source trajectory well.

Fig. 3 shows the speech signal from two nonconcurrent speakers denoted by trajectory 2 and the estimated result. This result demonstrate that EKPF is able to track the trajectory with a satisfactory accuracy, and quickly locks on to the source when the source switches.

For testing the performance of the algorithm in different reverberant environment, a Monte Carlo experiment with 50 runs is implemented under the different reflection coefficients. Fig. 4 gives the root mean square error (RMSE) [7] obtained from each trajectory with the SIR particle filter and the EKPF. As depicted in the figure, both the algorithms do well for line trajectory, but the tracking result of EKPF is much better for switch-speaker trajectory. This shows that our method is more effective for the sharp change of the position or the switch of the speakers in the moderate reverberant environment (reflection coefficients ≤ 0.6). However the performance degrades quickly when the reflection coefficient is greater than 0.6. This is because GCC algorithm collapses under the strong reverberation environment, and thus the observed TDOAs are far away from the true time delays.

5. CONCLUSIONS

A new approach to source tracking in a reverberant environment is presented in this paper. By linearizing the time delay function and using an extended Kalman filter, the optimal importance function can be derived. The particles thus can be relocated in a more accu-

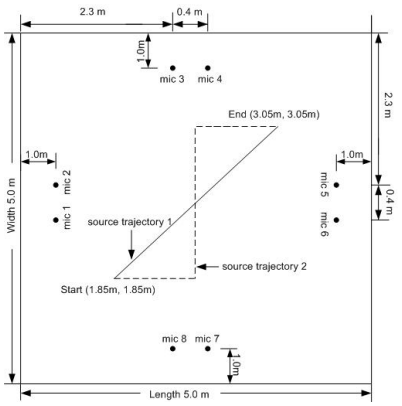


Fig. 1. Experiment setup. Black dots numbered from 1 to 8 are the position of microphones, solid line and dash-dotted line represent the line trajectory and switch-speaker trajectory respectively.

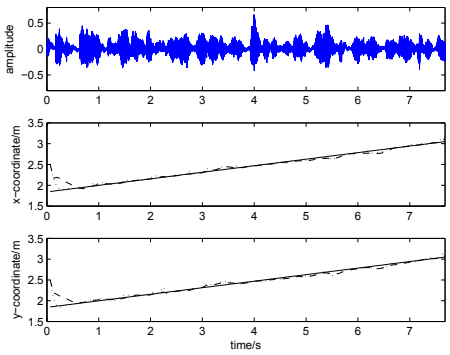


Fig. 2. Line trajectory (trajectory 1) estimation result under the reflection coefficient 0.5. Solid lines are true position, dashed lines and dotted lines denote the estimates by PF and EKPF respectively.

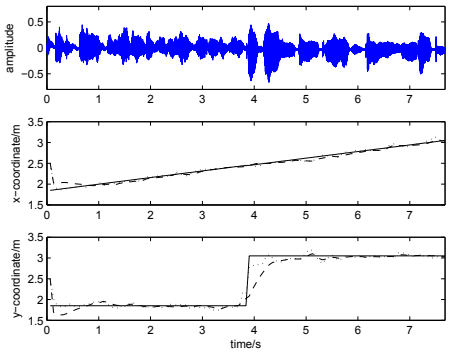


Fig. 3. Switch speaker (trajectory 2) estimation result under the reflection coefficient 0.5. Solid lines are true position, dashed lines and dotted lines denote the estimates by PF and EKPF respectively.

rate area, and helps the algorithm easily to recover from any tracking loss and detect the switch of speakers. The simulation results show that the tracking performance is robust against the reverberation and background noise, and even in a complicated motion case. As an initial stage of multiple simultaneously active sources tracking, nonconcurrent speakers tracking provide a lot of knowledge for our future work.

6. REFERENCES

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

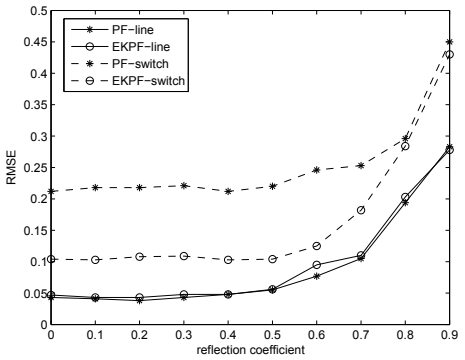


Fig. 4. Average RMSE for proposed method and general PF with different trajectories *vs.* different reflection coefficients. Solid lines and dashed lines are for line trajectory (trajectory 1) and switch trajectory (trajectory 2) respectively; circle and star denote the estimates based on EKPF and PF separately.

[3] M. Brandstein and D. Ward, *Microphone Arrays. Signal Processing Techniques and Applications*, Berlin, Germany: Springer-Verlag, 2001.

[4] Ulrich Klee, Tobias, and John McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[5] M.S. Brandstein, J.E. Adcock, and H.F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE trans. on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[6] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3021–3024, May 2001.

[7] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[8] R. van der Merwe, A. Doucet, N. de Freitas, and E. A. Wan, "The unscented particle filter," Tech. Rep., Cambridge University Engineering Department, Aug. 2000.

[9] Godsill S. Doucet, A. and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[10] J. B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.

TIME-FREQUENCY MASKING BASED MULTIPLE ACOUSTIC SOURCES TRACKING APPLYING RAO-BLACKWELLISED MONTE CARLO DATA ASSOCIATION

Xionghu Zhong and James R. Hopgood

Institute for Digital Communications, Joint Research Institute for Signal and Image Processing,
School of Engineering, The University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK
x.zhong@ed.ac.uk and James.Hopgood@ed.ac.uk

ABSTRACT

A Rao-Blackwellised particle filtering approach for tracking multiple simultaneously active and time-varying number of speakers is investigated. A novel measurement extraction method appropriate for the scenario of multiple sources is proposed based on a time-frequency masking technique, in which each source is represented separately by a joint gain-ratio and time-delay histogram. An existing Rao-Blackwellised particle filtering and data association approach is then employed to track the sources. The position of the sources are marginalised by an extended Kalman filter, and it is only the data association that needs to be processed using a particle filter. The tracking capability of the proposed algorithm is demonstrated in different simulated room environments and compared with the tracking performance obtained by applying the observation from the more commonly used generalised cross correlation function.

Index Terms— Time-frequency masking, particle filtering, Rao-Blackwellisation, data association.

1. INTRODUCTION

Locating and tracking speakers in a reverberant environment is an important research topic in many applications such as multimedia, hearing aids and hands-free teleconferencing systems. Existing localisation and tracking techniques usually rely upon the time delay of arrival (TDOA) estimates [1, 2], which is typically extracted by employing the generalised cross-correlation (GCC) function [3]. However, in the scenario that multiple simultaneously active speakers exist, the performance of the GCC based TDOA observation becomes severely degraded in that: 1) cross-correlation based TDOA estimation is not theoretically appropriate for multiple speech sources since it assumes that only one impinging wavefront arrives [3], thus it may not yield sharp peaks to present the TDOAs accordingly; and 2) the GCC function always has a time resolution problem that if the moving speakers are closely spaced in the room, it is almost impossible to report multiple TDOAs accurately.

The binary time-frequency (TF) masking method is a powerful tool for separating the sources from the received mixtures [4]. By assuming that the sources are W-disjoint on the TF domain, one can cluster the TF spectrogram bins of each source to form a two dimensional (2-D) time-delay and gain-ratio histogram (as shown, for example, in Fig. 2(a)). The TF masking based TDOAs are more reliable than the GCC based observations since: 1) the TF masking method is appropriate for the observation extraction of multiple sources, even when these sources are simultaneously active; 2) by clustering the TF bins, we obtain the observation from the source signal but get rid of those bins generated by noise. For this reason calculating the TDOAs using TF masks has an advantage to yield robust estimates even at low SNRs; and 3) because the TDOAs are calculated by phase unwrapping, it also has the advantage of achieving higher accuracy than the TDOAs estimated by the GCC method.

Various multisensor multitarget tracking techniques have been introduced for the multiple speakers tracking problem in the past few years [2]. Particle filtering (PF) methods are found to be appealing under the cases that the state/measurement equations are nonlinear and the noises are non-Gaussian [2]. Our aim in this paper is to track the unknown and time-varying number of speakers by using a Rao-Blackwellised particle filtering (RBPF) data association technique [5, 6]. In essence, the target states are assumed to follow a linear Gaussian model and can be integrated out in a closed-form, and thus the particle filtering is only necessary to be applied for the data association problem. Since this marginalisation replaces the finite particle set in the PF approaches, the estimation variance can be reduced, and fewer particles are needed for the same accuracy. The TDOA measurement function is nonlinear, but fortunately, the speakers moving in the room can be always supposed to be slow-paced, and thus the extended Kalman filter (EKF) is sufficient to track the positions. To implement the RBPF data association method for our tracking problem, the measurement function is linearised to form an EKF step, and the source dynamics are described by the birth, survival, or death models. A latent state is then introduced to identify the hypothesis associations. The particle filter is applied to the latent state to evaluate the importance of different associations. The number of simultaneously active speakers is usually assumed to be small so that the computation is still affordable by employing this multiple hypothesis tracking (MHT) based method.

2. TF MASKING BASED TDOA ESTIMATION

The multiple sources signal model is presented in this section. A TDOA extraction approach based on time-frequency masking is then proposed for the scenario of multiple sources. Let $\boldsymbol{p}_{\ell,i}$ and $\mathbf{x}_{m,t}$ denote the position of ith microphone in ℓth microphone pair and the position of mth source at time t, respectively. The free-field model of the received signals from M_t multiple sources can be written as

$$\begin{aligned} z_{\ell,i}(t) &= \sum_{m=1}^{M_t} \frac{1}{4\pi r_{\ell,i}^m(t)} s_m(t-\tau_{\ell,i}^m(t)) \\ &\quad + \sum_{m=1}^{M_t} s_m(t) * g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t}) + n_{\ell,i}(t) \\ &= \sum_{m=1}^{M_t} \frac{1}{4\pi r_{\ell,i}^m(t)} s_m(t-\tau_{\ell,i}^m(t)) + v_{\ell,i}(t), \end{aligned} \quad (1)$$

978-1-4244-2710-9/09/$25.00 ©

where $s_m(t-\tau^m_{\ell,i}(t))$ is the pure delayed source signal; $g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t})$ is the room impulse response (RIR) which is defined as the whole response minus the direct path component; $v_{\ell,i}(t)$ is the noise term denoting the sum of the channel noise $n_{\ell,i}(t)$ and the reverberation signal $s_m(t) * g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_{m,t})$ for all m, with $*$ denoting the convolution; $r^m_{\ell,i}(t) = \|\mathbf{x}_{m,t} - \boldsymbol{p}_{\ell,i}\|$ is the distance between the source and microphone where $\|\cdot\|$ denotes the Euclidean-norm; $\tau^m_{\ell,i}(t) = r^m_{\ell,i}(t)/c$ is the direct path time-delay with c representing the speed of sound.

The time index t and microphone pair index ℓ are suppressed to simplify the expression where there is no danger of ambiguity arising, and let $a_{m,i} = 1/4\pi r_i^m$. The discrete short time Fourier transform (STFT) of the received signal (1) can be written as

$$Z_i(k,\nu) = \sum_{m=1}^{M_t} a_{m,i} e^{-j\nu\omega_0\delta_i^m} S_m(k,\nu) + V_{m,i}(k,\nu), \quad (2)$$

where k and ν are the sliding window index and the frequency index respectively; ω_0 is the discrete frequency spacing parameter; and Z, S and V are the STFT of the received signal, source signal and noise term respectively. The TF bin ratio is defined as [4]

$$R(k,\nu) = Z_1(k,\nu)/Z_2(k,\nu). \quad (3)$$

Ignoring the effect of noise, we can easily obtain the gain-ratio (GR) and time-delay (TD) estimates for each TF bin as

$$\hat{a}_{k,\nu} = |R(k,\nu)| \quad \text{and} \quad \hat{\tau}_{k,\nu} = -1/(k,\nu\omega_0)\angle R(k,\nu), \quad (4)$$

with $|\cdot|$ and $\angle\cdot$ denoting the amplitude and the phase of the estimates respectively. Given the GR resolution parameter A and TD resolution parameter D, define the TF bin indicator function as

$$\Lambda_{A,D}(k,\nu) = \begin{cases} 1, & \text{if } |\hat{a}_{k,\nu} - \zeta A| \le A, |\hat{\tau}_{k,\nu} - \eta D| \le D; \\ 0, & \text{otherwise,} \end{cases}$$

where ζ and η are any integers. The function Λ indicates whether the GR and TD of a TF bin locate around a given parameter $(\zeta A, \eta D)$. Following [4], a 2-D histogram is constructed as

$$h(\zeta A, \eta D) = \sum_{k,\nu} \Lambda_{A,D}(k,\nu)\, |Z_1(k,\nu) Z_2(k,\nu)|^\gamma, \quad (5)$$

where $|Z_1(k,\nu)Z_2(k,\nu)|^\gamma$ is a weighting term for some γ. Detailed discussion about the different choices of γ can be found in [4]. Here $\gamma = 0$ is picked to equalise the importance of all the TF bins. Thus, at the frame index k, the TDOA for the nth peak in the 2-D histogram, h, of the ℓth microphone pair can be estimated as

$$\mathrm{y}^\ell_{n,k} = E(\hat{\tau}_{k,\nu} | k, \nu \in \mathcal{S}_n), \quad (6)$$

where $E(\cdot)$ denotes the expectation, and $\mathcal{S}_n$ is the set of TF bins defined by the bins around the nth peak on the 2-D histogram of the corresponding microphone pair.

Assuming that N^ℓ_k peaks can be enumerated in the 2-D histogram, and L microphone pairs are employed, the complete observation set can be expressed as

$$\mathcal{Y}_k = \bigcup_{\ell=1}^{L} \mathbf{y}^\ell_k \quad \text{with} \quad \mathbf{y}^\ell_k = \{\mathrm{y}^\ell_{1,k}, \cdots, \mathrm{y}^\ell_{N^\ell_k,k}\}. \quad (7)$$

The TF masking measurement extraction is a separation-based approach. Unlike the GCC function which is unable to differentiate between all the different source signals, the TDOAs obtained by TF masking approach are estimated from the TF bins of individual source signals. Thus, it is able to give the TDOAs for multiple sources, even when these sources are simultaneously active.

3. RBPF MULTIPLE SOURCES TRACKING AND DATA ASSOCIATION

3.1. Rao-Blackwellisation formulation

The position set of the speakers on frame step k can be presented by $\mathcal{X}_k = \{\mathbf{x}_{1,k}, \cdots, \mathbf{x}_{M_k,k}\}$, with $\mathbf{x}_{m,k}$, $m = 1, \ldots, M_k$ representing the state of mth source, and M_k denoting the number of active speakers on the frame step k. The motion dynamics are modelled by the Langevin motion model [1]. The association variable $\boldsymbol{\theta}_k$ is defined as $\boldsymbol{\theta}_k = (\lambda_k, b_k, d_k)$. $\lambda_k \in \mathbb{N} \cup \emptyset$ is a data association event: with value 0 denoting that the observation is associated to a clutter; m associated to the mth source; and $\emptyset$ means that the microphone pair fails to produce any measurement. b_k and d_k are the birth and death indicators of the source separately. The interest here is to estimate the joint posterior distribution $p(\mathbf{x}_{m,1:k}, \boldsymbol{\theta}_{1:k} | \mathrm{y}^\ell_{n,1:k})$, which can be decomposed to the conditional source distribution $p(\mathbf{x}_{m,1:k} | \boldsymbol{\theta}_{1:k}, \mathrm{y}^\ell_{n,1:k})$ and the association posterior density $p(\boldsymbol{\theta}_{1:k} | \mathrm{y}^\ell_{n,1:k})$, as follows

$$p(\mathbf{x}_{m,1:k}, \boldsymbol{\theta}_{1:k} | \mathrm{y}^\ell_{n,1:k}) = \underbrace{p(\mathbf{x}_{m,1:k} | \boldsymbol{\theta}_{1:k}, \mathrm{y}^\ell_{n,1:k})}_{\text{EKF approximation}} \underbrace{p(\boldsymbol{\theta}_{1:k} | \mathrm{y}^\ell_{n,1:k})}_{\text{PF}}. \quad (8)$$

Supposing that the position state $\mathbf{x}_{m,k}$ can be marginalised out, according to the Rao-Blackwellisation theory, only the particle sampling for the latent variable $\boldsymbol{\theta}_k$ is required [5, 6]. Given an importance distribution of the association hypothesis variable $q(\boldsymbol{\theta}_k | \boldsymbol{\theta}^{(j)}_{1:k-1}, \mathrm{y}^\ell_{n,1:k})$, the jth particle weight of the Rao-Blackwellisation particle filter can be updated as

$$\omega^{(j)}_k \propto \omega^{(j)}_{k-1} \frac{p(\mathrm{y}^\ell_{n,k} | \boldsymbol{\theta}^{(j)}_{1:k}, \mathrm{y}^\ell_{n,1:k-1}) p(\boldsymbol{\theta}^{(j)}_k | \boldsymbol{\theta}^{(j)}_{1:k-1})}{q(\boldsymbol{\theta}^{(j)}_k | \boldsymbol{\theta}^{(j)}_{1:k-1}, \mathrm{y}^\ell_{n,1:k})}. \quad (9)$$

A target birth is assumed to happen with a probability of $p(b^{(j)}_k)$, and is independent of any existing sources. The common initial Gaussian distribution with the initial state and variance $\mathbf{m}_0$ and $\mathbf{P}_0$ is given if a birth occurs. After assigning the new measurement with a source, the life-time $t^{(j)}_d$ of the source can be modelled by a gamma distribution [5], i.e., $t^{(j)}_d \sim \text{gamma}(\phi|\alpha, \beta)$. The probability that the source is dead at current time step t_k is [5]

$$p(d^{(j)}_k) = P(t^{(j)}_d \in [t_{k-1} - t^{(j)}_{m_k}, t_k - t^{(j)}_{m_k}] | t^{(j)}_d \geqslant t_{k-1} - t^{(j)}_{m_k}), \quad (10)$$

where $t^{(j)}_{m_k}$ is the time that the last association with source $m^{(j)}_k$, and t_{k-1} denotes that the source is alive on the previous time step.

3.2. Optimal importance function

One assumption is made to avoid an exponential growth of the hypotheses: at most one source can be born/die at a given time. The optimal importance distribution can be derived as

$$\begin{aligned} \boldsymbol{\theta}^{(j)}_k &\sim q(\boldsymbol{\theta}^{(j)}_k | \boldsymbol{\theta}^{(j)}_{1:k-1}, \mathrm{y}^\ell_{n,1:k}) \\ &= \frac{p(\mathrm{y}^\ell_{n,k} | \boldsymbol{\theta}^{(j)}_{1:k}, \mathrm{y}^\ell_{n,1:k-1}) p(\boldsymbol{\theta}^{(j)}_k | \boldsymbol{\theta}^{(j)}_{1:k-1})}{p(\mathrm{y}^\ell_{n,k} | \boldsymbol{\theta}^{(j)}_{1:k-1}, \mathrm{y}^\ell_{n,1:k-1})}. \end{aligned} \quad (11)$$

If the nth, $n = 1, \cdots, N^\ell_k$ measurement $\mathrm{y}^\ell_{n,k}$ is from the mth target, it follows the measurement function

$$\begin{aligned} \mathrm{y}^\ell_{n,k} &= \mathbf{h}(\mathbf{x}_{m,k}, \boldsymbol{p}_\ell) + \mathbf{w}^\ell_{n,k} \\ &= c^{-1}(\|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_{m,k} - \boldsymbol{p}_{\ell,2}\|) + \mathbf{w}^\ell_{n,k}. \end{aligned} \quad (12)$$

Since this measurement model is nonlinear, the local linearisation method such as extended Kalman filter (EKF) is needed to evaluate the likelihood, given by

$$p(\mathrm{y}_{n,k}^{\ell}|\mathbf{x}_{m,k}, \mathrm{y}_{n,1:k-1}^{\ell}) = \mathrm{EKF}(\mathrm{y}_{n,k}^{\ell}; \widetilde{\mathbf{H}}_{m,k}\mathbf{x}_{m,k}, \widetilde{\mathbf{R}}_{m,k}), \quad (13)$$

where $\mathrm{EKF}(\cdot)$ denotes the implementation of EKF. The derivation of the EKF model matrix $\widetilde{\mathbf{H}}$ and $\widetilde{\mathbf{R}}$ and its particle filtering implementation can be found in [7].

The measurement can be generated either by a source or by a clutter due to the reverberation/noise. An uniform distribution over the possible TDOA interval is given in the case that the measurement is from a clutter,

$$p(\mathrm{y}_{n,k}^{\ell}) = \mathcal{U}_{[-\tau_{\max}, \tau_{\max}]}(\mathrm{y}_{n,k}^{\ell}) = \frac{1}{2\tau_{\max}}, \quad (14)$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay which can only happen when the microphone pair and the source lie exactly on an extended line.

Since the time-varying number of sources is considered in this paper, the birth and the death processes are used to model the activity and inactivity of a source respectively. The processes are assumed to be independent from each other and the survival sources. The prior of the hypothesis association $p(\boldsymbol{\theta}_k^{(j)}|\boldsymbol{\theta}_{1:k-1}^{(j)})$ can thus be written as

$$p(\boldsymbol{\theta}_k^{(j)}|\boldsymbol{\theta}_{1:k-1}^{(j)}) = p(\lambda_k^{(j)}|b_k^{(j)}, d_k^{(j)}, \lambda_{1:k-1}^{(j)}) p(d_k^{(j)}) p(b_k^{(j)}), \quad (15)$$

where $p(b_k^{(j)})$ and $p(d_k^{(j)})$ are already defined in the section 3.1. The association distribution $p(\lambda_k^{(j)}|b_k^{(j)}, d_k^{(j)}, \lambda_{1:k-1}^{(j)})$ is dependent only on the number of sources $\overline{M}_k$ at the time step k, given by [6]

$$p(\lambda_k^{(j)}) = \begin{cases} p_f, & \text{if } \lambda_k^{(j)} = 0; \\ (1-p_f)P_{\overline{M}_k}, & \text{if } \lambda_k^{(j)} = \emptyset; \\ \frac{(1-p_f)(1-P_{\overline{M}_k})}{\overline{M}_k}, & 1 \leqslant \lambda_k^{(j)} \leqslant \overline{M}_k, \end{cases} \quad (16)$$

where $P_{\overline{M}_k} = (1-P_d)^{\overline{M}_k}$ denotes the probability that the microphone pair fails to report any correct measurements; P_d represents the probability of detection; and p_f is the prior of the clutter. The denominator in equation (11) can be derived as

$$p(\mathrm{y}_{n,k}^{\ell}|b_k^{(j)}, d_k^{(j)}, \boldsymbol{\theta}_{1:k-1}^{(j)}, \mathrm{y}_{n,1:k-1}^{\ell}) = \sum_{\lambda,b,d} p(\mathrm{y}_{n,k}^{\ell}|\boldsymbol{\theta}_{1:k}^{(j)}, \mathrm{y}_{n,1:k-1}^{\ell}) p(\lambda_k^{(j)}) p(d_k^{(j)}) p(b_k^{(j)}). \quad (17)$$

Because the dimension of the states needed to be estimated by the particle filtering is reduced by integrating out the position states using an EKF, fewer particles is needed to achieve the same accuracy. The top-level procedure of the tracking algorithm is depicted in the Alg. 1. The complete data association algorithm can be found in [6].

Algorithm 1: Top-level procedure of the RBPF.

Initialisation: $\omega_0^{(j)} \leftarrow 1/N$; $\mathcal{X}_0^{(j)} \leftarrow \emptyset$.
Over all the measurements:
for $k \leftarrow 1$ **to** K **do**
  for $j \leftarrow 1$ **to** N **do**
    - compute the likelihood in (13) and (14) and filtered states $\hat{\mathcal{X}}_{k-1}^{(j)}, \hat{\mathbf{x}}_0$;
    - compute the association priors in (10) and (16);
    - evaluate the importance distribution in (11);
    - draw the samples according to the importance distribution,
    if *birth/death* **then**
      $\mathcal{X}_k^{(j)} \leftarrow \hat{\mathcal{X}}_{k-1}^{(j)} \cup \hat{\mathbf{x}}_0 \quad \mathcal{X}_k^{(j)} \leftarrow \hat{\mathcal{X}}_{k-1}^{(j)} \setminus \mathbf{x}_m$
    end
    if *survival* **then**
      $\mathcal{X}_k^{(j)} \leftarrow \hat{\mathcal{X}}_{k-1}^{(j)}$;
    end
    - update the importance weight $\omega_k^{(j)}$.
  end
  Output the estimates.
  Resample $(\mathcal{X}_k, \omega_k)$ *if necessary.*
end

4. SIMULATION EXPERIMENTS

The dimension of the simulated office room is $5\times5\times3\mathrm{m}^3$. Four microphone pairs each with a separation of 0.1m are organised around the center of the walls. Here, only the two dimensional tracking problem is considered, so that the height of the microphones and sources are assumed to be known and at the same height of 1.5m. The audio signals are split into 120 frames with a frame length of 128ms, at a sampling frequency of 8000Hz. The sources are thus moving at a velocity of 0.2m/s, comparable with the source velocities in [1, 2]. All the reverberations in the room are simulated using the imaging method [8], and the signal-to-noise ratio (SNR) is set to be 35dB. Two speakers have time-varying appearance: one is active from frame index 1 to 100, and the other from the frame index 70 to 120. The detailed experiment setup are depicted in Fig. 1.

For each microphone pair, the two largest peaks in the histogram are picked to estimate the TDOAs. Fig. 2(b) gives the ground truth and observed TDOAs across all the microphone pairs. Due to the reverberation, the false peaks may be observed; some of these false peaks are even higher than the peaks from the real sources; see Fig. 2(a). The parameters for the tracking algorithm are set as $p_{\mathrm{b}} = 0.08$, $p_f = 0.3$, $P_d = 0.95$, $\alpha = 2$ and $\beta = 1$. 20 particles are used to evaluate the algorithm. Fig. 3 gives the tracking result for a single experiment under the reverberation time $T_{60} = 0.19$s. The estimated positions are visualised employing a 'winner-particle'approach [6]. The results demonstrate that our approach is able to track the posi-

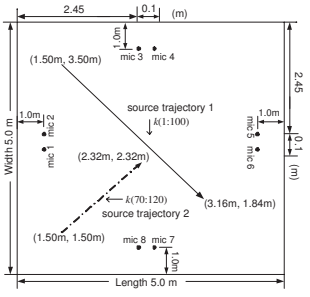
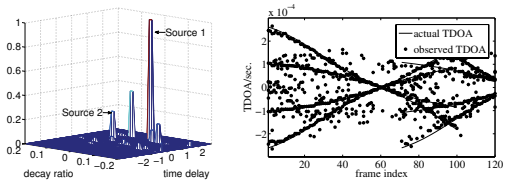


Fig. 1. Experiment setup. Number 1 to 8 dots are the microphone positions, solid and dash-dotted line are the trajectories respectively.



(a) (b)

Fig. 2. (a) 2-D histogram of two sources in the reverberant environment; (b) TF masking based TDOA estimates from the four microphone pairs in the reverberant environment (T_{60}=0.19s).

tion of multiple time-varying speakers with a satisfactory accuracy, even in the reverberant environment.

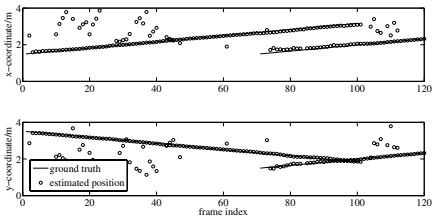


Fig. 3. Tracking result under the reverberation time (T_{60}=0.19s).

The Wasserstein distance (WD) [9] is introduced to examine the average tracking performance, and the result is also compared with the performance based on the GCC measurements. All the experiment setup are the same for the GCC method except that the microphone distance is set as 50cm following the microphone separations in [2]. Fig. 4 shows the average WD distance based on the TF masking method and GCC method in the anechoic environment and reverberant environment, where the reverberation time T_{60} is 0s and 0.19s respectively. All of the experiments are implemented under 500 Monte Carlo trials. The results demonstrate that the TF masking based method is outperforming the GCC based method, especially when multiple sources simultaneously exist.

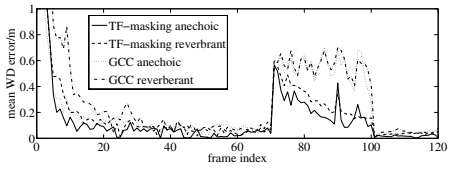


Fig. 4. Average position error under the anechoic (T_{60}=0s) and reverberant (T_{60}=0.19s) environment.

Table 1 gives the average WD over all the time steps under 500 Monte Carlo trials in different reverberant environment. The wall reflection coefficients is varied to generate different reverberations. The corresponding reverberation time T_{60} can be found in the table 1. The result shows that the TF masking based approach outperforms the GCC based tracking algorithm in the moderate reverberant environment. However, both of the approaches degrade sharply as long as T_{60} exceeds 0.4s. This is because the observations are distorted drastically in the strong reverberation environment. Thus RBPF algorithm is failed to formulate any appropriate tracks.

Table 1. Average WD *vs.* different reverberation time

T_{60}	0	0.15	0.19	0.23	0.28	0.41
TF	0.096	0.117	0.127	0.152	0.173	0.341
GCC	0.188	0.194	0.211	0.241	0.291	0.415

5. CONCLUSION AND FUTURE WORK

Using the time-frequency masking technique, a new TDOA estimation method is developed to extract the observations under the scenario of multiple simultaneously active sources. The RBPF based data association approach is then introduced to handle the unknown and time-varying number of sources tracking problem. Simulations show that the proposed method can correctly report the positions of each source. Knowing that the TF masking method does not only provide the TDOAs, but also gives the decayed source energy information, which may be useful to discriminate the TDOAs from the sources/clutters, it will be interesting in our future work to investigate a tracking algorithm incorporating the source energy information to enhance the data associations, which is extremely necessary in the reverberant environment. An application of the proposed approach in a real room environment will also be considered.

6. REFERENCES

[1] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE trans. on*, vol. 11, no. 6, pp. 826-836, Nov. 2003.

[2] Wing-Kin Ma, Ba-Ngu Vo, Sumeetpal S. Singh, and Adrian Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *Signal Processing, IEEE Trans. on*, vol. 54, no. 9, pp. 3291-3304, Sep. 2006.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE trans. on*, vol. 24, no. 4, pp. 320-327, Aug. 1976.

[4] Özgür Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE trans. on*, vol. 52, no. 7, pp. 1830-1847, July 2004.

[5] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-Blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2-15, January 2007.

[6] M. Vihola, "Rao-blackwellised particle filtering in random set multitarget tracking," *Aerospace and Electronic Systems, IEEE trans. on*, vol. 43, no. 2, pp. 689-705, April 2007.

[7] Xionghu Zhong and J.R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 293-296, 2008.

[8] J. B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, Jul. 1979.

[9] J.R. Hoffman and R.P.S. Mahler, "Multitarget miss distance via optimal assignment," *Systems, Man and Cybernetics, IEEE Trans. on, Part A: System and Humans*, vol. 34, no. 3, pp. 327- 336, 2004.