Binaural Sound Source Localisation in Complex Conditions

Xiang Wu

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY OF THE AUSTRALIAN NATIONAL UNIVERSITY

THE AUSTRALIAN NATIONAL UNIVERSITY

April, 2019

Declaration

The contents of this thesis are the results of original research carried out by my self, under the supervision of Prof. Thushara D. Abhayapala, and Prof. Wen Zhang and Dr. Dumidu S. Talagala. These have not been submitted for a higher degree to any other university or institution.

Some of the work in this thesis has been published or has been submitted for publication in referee journal papers and conference proceedings. The following is a list of these publications.

Publications

D. S. Talagala, X. Wu, W. Zhang and T. D. Abhayapala,"Binaural localization of speech sources in the median plane using cepstral HRTF extraction," *2014 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, 2014, pp. 2055-2059.

X. Wu, D. S. Talagala, W. Zhang and T. D. Abhayapala,"Binaural localization of speech sources in 3-D using a composite feature vector of the HRTF," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 2654-2658.

X. Wu, D. S. Talagala, W. Zhang and T. D. Abhayapala,"Spatial feature learning for robust binaural sound source localization using a composite feature vector," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 6320-6324.

X. Wu, D. S. Talagala, W. Zhang and T. D. Abhayapala,"Individualized Interaural Feature Learning and Personalized Binaural Localization Model," *Applied Acoustics Journal*, vol. 9, no. 13, 2682, June 2019.

X. Wu, D. S. Talagala, W. Zhang and T. D. Abhayapala,"Two-State Binaural Head Rotation Model and Multiple Sources Localization," *IEEE Trans. Audio, Speech, Language Process.*, to be submitted.

Xiang Wu
College of Engineering and Computer Science
The Australian National University
ACT 0200
Australia

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Prof. Thushara Abhayapala, Dr. Wen Zhang and Dr. Dumidu Talagala for their guidance, insight and encouragement. Prof. Thushara gave me the opportunity to pursue this doctorate, as well as for his continuing support to me for many years. I take this opportunity to sincerely acknowledge his guidance and support. Prof. Wen introduced me into the audio and acoustics area, and I learnt many valuable learning, research and writing skills from her, which causes me to benefit lifetime. I would like to thank her for her patience, for spending lots of time and energy on revising my works. My sincere thanks also goes to Dr. Dumidu Talagala for his motivation, enthusiasm and support. I will never forget those hours that we discussed about my project and thesis on skype.

Secondly, I would like to thank Dr. Ravish Mehra and Dr. Glenn Dickins, for offering me the internship opportunities at Oculus and Dolby. I am very glad to have those chances to working on those exciting projects.

I would also like to thank my family, especially for my parents and my wife, who has encouraged and believed in me through these years. Thank you for supporting me and giving me faith when I was stressed and depressed.

Last but not least, I want to thank the editors from Elite Editing for their help on proofreading this thesis.

Abstract

There has been a growing interest in the reproduction of human spatial hearing behaviours, arising from the development of spatial audio signal processing techniques. To accurately localise single or multiple sound sources using humanoid apparatus, it is essential to be able to exploit the spatial-related features of the human subject filtering effect, which requires an understanding of both the feature characteristics and the mapping relationship to the source locations. In this thesis, we analyse and evaluate the localisation feature characteristics of binaural signal, and explore a method for constructing a localisation mapping model.

As a result of the reflecting and diffracting of human-like apparatus, sound waves are filtered before being captured by the eardrum, and the filtering effects result in various behaviours in the frequency domain. This thesis first summarises the characteristics of those behaviours and evaluates their importance to localisation. We analyse and evaluate the correlation between source location and three main interaural cues, which are interaural level differences, interaural time difference and interaural phase difference. Then, we explore the process to exploit those features using, and develop a novel feature vector by combining the most valuable spectra. Following this, by employing mutual information as the evaluation metric for frequencies selection, we propose a new feature location mapping model that embeds the feature evaluation process. The new mapping uses a multiple-tree structured model based on the random forest that shows high tolerance to noise. Through computational simulations and practical experiments, the model presents an improvement in both accuracy and robustness according to the comparison of the angular error and localisation correct rate. Finally, by combining our localisation method with the recent proposed direct path transfer function estimation method based on a convolutive transfer function model, we design a binaural localisation system for an unknown environment.

The remainder of this thesis demonstrates the possibility of using the active localisation cues in a binaural system. Based on observations of human active head rotation behaviour, we investigate the effect of dynamic features in binaural localisation. The analysis shows that head rotation enriches the variation of localisation features, which resolve the problem of cone-of-confusion and simplifies vertical-wise

localisation in a 3-D space. In addition, we develop a multiple-source localisation method based on the head rotation process, which indicates that dynamic features would be the solution to many localisation problems caused by the limitation on the number of signal channels.

List of Symbols

α	Sound wave incident angle
β	Absorption coefficient of reflecting surface
$\boldsymbol{R}$	Listener's head movement matrix
$\mathbf{v}^p, \mathbf{v}^m$	Interaural phase and magnitude feature vector
$\mathrm{x}, \mathrm{y}, \mathrm{z}$	Cartesian coordinates in 3-D space
Θ	Sound source location
θ, ϕ	Azimuth and elevation in vertical polar coordinate system, respectively
ϑ, φ	Azimuth and elevation in interaural polar coordinate system, respectively
a	Listener's head radius
f	Frequency index
$h'_{\mathrm{l}}, h'_{\mathrm{r}}$	Room impulse response of left and right ear
$H_{\mathrm{l}}, H_{\mathrm{r}}$	Head-related transfer function of left and right ear
$h_{\mathrm{l}}, h_{\mathrm{r}}$	Head-related impulse response of left and right ear
n	Additive noise component
r	The distance between sound source to the observation point
s	Source signal
t	Time index
$x_{\mathrm{l}}, x_{\mathrm{r}}$	Received binaural signals of left and right ears

List of Acronyms

BRIR	Binaural Room Impulse Response
CTF	Convolutive Transfer Function
DNN	Deep Neural Network
DP-RTF	Direct PathRelative Transfer Function
DRR	Direct-to-residual ratio
EM	Deep Neural Network
GCC	Generalised Cross-correlation
GCC-PHAT	Generalised Cross-correlation Phase Transform
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
HRIR	Head-related Impulse Response
HRTF	Head-related Transfer Functions
IID	Interaural Intensity Difference
ILD	Interaural Level Difference
ILPD	full-spectrum ILD and low-frequency IPD vectors
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
ITF	Interaural Transfer Function
KNN	k-Nearest Neighbours
MI	Mutual Information
MPC	Minimum Phase Component
MTF	Multiplicative Transfer Function
MUSIC	Multiple Signal Classification
OOB	Out-of-bag
PETs	probability estimation trees
PHAT	Phase Transform
PPAM	Probabilistic Piecewise Affine Mapping
PSD	Power Spectral Density
RF	Random Forest
RIR	Room Impulse Response

RTF	Relative Transfer Function
SCA	Source Cancellation Algorithm
SNR	Signal-to-Noise Ratio
STFT	Short-time Fourier Transform
TDOA	Time Difference of Arrival
VP	Vertical-polar

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

Human beings can detect and analyse the direction, distance and movement of sound sources via their outstanding auditory systems, which are essential for perceiving the surrounding environment and avoiding potential dangers in regions that their vision cannot reach. The ability to exploit spatial information from audio signals is known as sound source localisation, which has various applications. For instance, in speech recognition, a localisation system can help specify the target speaker from multiple sound sources and improve the correctness of the recognition [1, 2]. In artificial intelligence, spatial auditory systems can help robots sense the environment and localise objects when their vision is blocked [3, 4]. Moreover, in the virtual reality field, which has dramatically developed in recent years, investigations of humans spatial auditory systems play a significant role in understanding humans perceptual behaviour and improving customers immersion experiences [5–7]. Thus, sound localisation techniques are necessary for humancomputer interaction.

Recently, there has been growing interest in humans hearing mechanisms, arising from the development of personalised media devices, such as virtual reality goggles. Great effort has been invested in understanding humans spatial auditory perceptual behaviour, which is known as auditory scene analysis [8], and the approach to reproduce the spatial hearing model based on humans biological and psychological mechanism, which is known as computational audio scene analysis [9]. Reproducing the human hearing model is achieved by constructing computational models that incorporate the two received signals and the known listeners physical acoustical features. This thesis aims to propose novel binaural localisation methods for complex acoustical conditions with consideration of both the humanoid subjects spatial features and the microphones placement.

Since the “duplex theory’ was proposed by Lord Rayleigh in the early twenti-

eth century [10], many hearing models have been developed. With the publication of Jeffresss coincidence-based model [11], the binaural hearing model was widely accepted and developed. At the same time, the interaural masking effect was independently observed and described by [12] and [13]. In 1983, Lyon first adopted Jeffress's model in binaural localisation [14], which calculated the spectral coherence between the ears to obtain the time delays caused by different sound sources. Later, Bodden proposed a multiple-source localisation model incorporating both time difference and level difference between the ears [15]. Meanwhile, psychoacoustic studies indicated that the main localisation cues include interaural time difference (ITD), interaural level difference (ILD) and spectral cues [16]. Further, the more detailed acoustical feature characteristics of listeners' anatomical structures were described by head-related transfer functions (HRTFs) in the frequency domain, and many localisation algorithms based on HRTF were established, such as the matched filtering approach [17], source cancellation algorithm [18, 19] and reference signal approach [20].

In contrast, many attempts have been made to resolve the localisation problem by adopting existing array signalprocessing techniques. The array signal processing-based source localisation technique has been widely used in numerous applications with various signal media, such as sonar, radar and telecommunications, and those algorithms are capable of adapting to process acoustic signals, with different assumptions of signal wave propagation. For instance, methods such as multiple signal classification (MUSIC) [21, 22] and generalised cross-correlation [23] have been proposed for pinpointing acoustic sound sources based on the variation between channel transfer functions caused by different source locations. Based on these methods, one may consider binaural localisation a special case of array signal processing with two receivers. Talagala and his colleagues [24] proposed a localisation algorithm based on MUSIC and employed the complex behaviour of the HRTF as additional spatial information to estimate the source positions.

In addition, the investigation in HRTF revealed the nature of humans binaural sound localisation in 3-D space. The researchers established that the monaural HRTF spectral features that caused by pinna provide the primary cues for vertical localisation [16, 25, 26]. The effect of the torso in vertical plane localisation was first investigated by Gardner [27], while Algazi analysed elevation-related low-frequency cues in 2001 [28]. Further, Duba summarised the elevation-related interaural feature characteristics [29]. Based on this research, a few localisation methods have been developed for 3-D space localisation. Keyrouz *et al.* extended the maximum

cross-correlation method to the 3-D space [20, 30]. Recently, with the development of artificial intelligence and machine learning, applying the statistical-based learning approach to localisation problems has become a promising solution. For instance, Weng *et al.* adopted a non-parametric tree-based learning method to describe the mapping between the interaural cues and source locations with fewer restrictions on its spatiotemporal characteristics and environment structure [31]. Deleforge *et al.* proved local linear bijective mapping between the interaural cues and source location on the binaural manifold, and derived a statistical localisation model using an expectation-maximisation (EM) algorithm [32, 33]. However, because of the complexity of human listeners' biological structures, there is not yet a closed-form function to map the source vertical position and spatial cues; thus, localising sound sources in 3-D space with binaural configurations is a challenging topic.

Another concern regarding binaural localisation is the effect of the surrounding acoustical conditions, such as for enclosed spaces with reverberation. In such a scenario, the reflected sound waves are directional, and will interfere with the estimation of the true direction of the source; thus, binaural de-reverberation is particularly necessary for binaural applications [34–36]. Studies in psychoacoustics have shown that the sound source location sensed by human auditory systems is dominated by the first wave front, which is known as the Haas effect or precedence effect [37, 38], and a serious of methods have been proposed using such phenomenon for localising sources in a reverberant environment [3, 39, 40]. Alternatively, numerous binaural de-reverberation approaches have been proposed based on signal processing techniques [36, 41–44] and applying coherence-based filtering or selection procedures to reduce the reverberation noise [36, 43]. However, most of the above methods would require a pre-training procedure for an ideal de-reverberation outcome; otherwise, the residues may still affect the localisation performance.

More recently, the active binaural localisation model has attracted increasing attention. The previous models mentioned above assumed that the listener's position and posture were static, while the active binaural model introduced mobility of the subject to obtain more localisation features. The idea was inspired by observations of human listeners, who adjust their head posture to obtain a more accurate source position [45–47], and several methods have been proposed to adopt this mechanism to enable better localisation performance in different conditions [48–51]. Although those implementations have justified the effectiveness and robustness of the active model, there remains room for improvement in system design and algorithm optimisation.

1.2 Problem Statement

Based on the discussion above, the main problem considered by this thesis can be summarised as follows:

The challenge in binaural localisation is to create a model to map the speech source direction in the 3-D space with interaural features and to be robust to complex acoustical conditions and external interferences.

1.3 Thesis Structure

Figure 1.1 illustrates the general research contents and their relationship to the thesis structure. A brief introduction to the main content in each chapter is listed as follows.

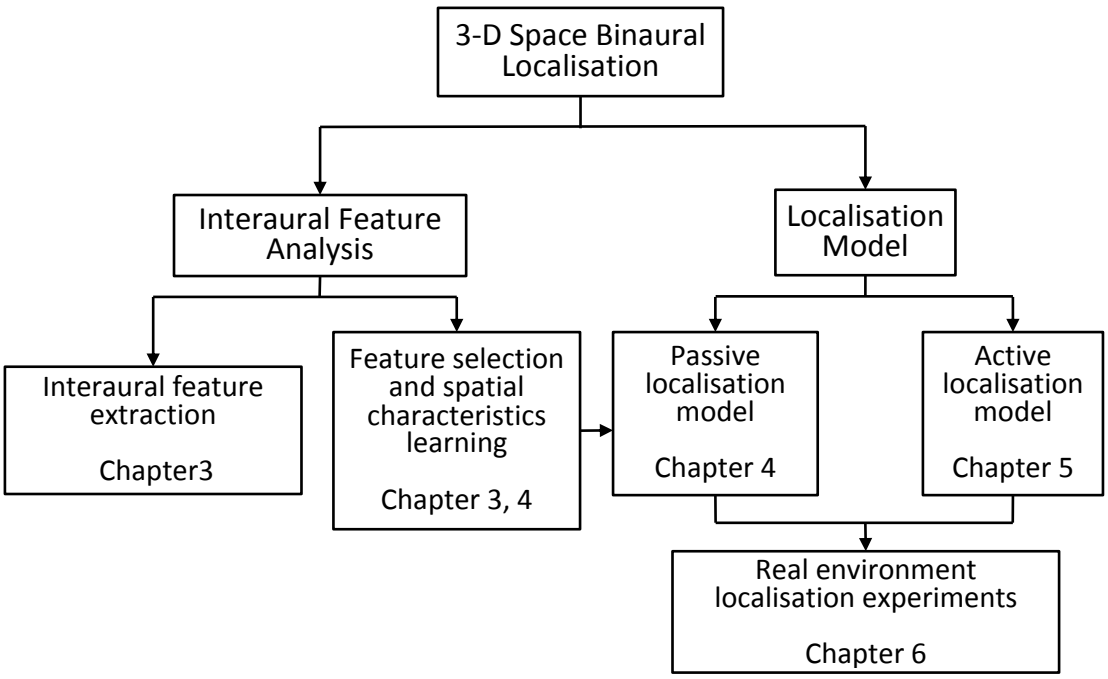


Figure 1.1: Main thesis research area and structure.

Chapter 2 provides a literature review on binaural localisation, including the fundamental concept of binaural signal processing. By introducing the famous duplex theory [10], the basic binaural localisation system is explained with commonly used coordinate systems. In addition, an overview of the widely accepted binaural localisation features is included. In particular, the HRTF is formally defined and the relationship between those features and the HRTF is derived. In addition,

this chapter introduces the modelling of binaural signals with external acoustical conditions, such as enclosed spaces.

Chapter 3 investigates the extraction of spatial features, especially for elevation estimation, in the presence of additive noise. The approach uses cepstral transformation and truncation to extract the elevation-related features. In addition, we propose a composite features vector that simultaneously takes advantage of the magnitude and phase of the interaural cues based on prior empirical knowledge in 3-D space localisation. The Euclidean distance matchingbased localisation method is adopted with the proposed feature vector, which can provide a robust estimation of the source directional position in a noisy environment by selecting the most valuable frequency ranges of the interaural cues.

Chapter 4 develops a feature selection algorithm and proposes an advanced probabilistic localisation model based on the observations from Chapter 3. The dependency between interaural features and the source direction is measured and analysed by mutual information (MI). The localisation model is obtained by the random forest (RF) algorithm, which incorporates not only the dependency between the source direction and interaural features, but also the internal conditional relationship between features. The model can accurately localise the source position in an unfamiliar reverberant and noisy environment by cascading a de-reverberation pre-processor.

Chapter 5 proposes an active localisation model that includes the head movement of the subject. This chapter starts by investigating the interaural feature behaviours with three types of head movement. We then develop a two-state localisation model based on Gaussian progress using only ITDs. Further, we introduce a novel mixed-source separation algorithm based on cepstral speech features in the head rotation model, which enables the model to localise multiple sources with a complex acoustical condition.

Chapter 6 demonstrates the experiment results for the methods proposed in Chapter 4 and Chapter 5. The experiment is implemented using an array of 30 loudspeakers and a KEMAR dummy head recording system. The experiment results indicate validation of the proposed methods in practical reverberation and noise conditions.

Chapter 7 presents the conclusion of the thesis and discusses possible directions for future work.

1.4 Thesis Contributions

This thesis makes several contributions to binaural localisation and understanding the human spatial auditory mechanisms. In general, this thesis analyses the relationship between interaural features and the directional source position, and proposes robust 3-D space localisation methods for various acoustical conditions. The principle contributions are listed below:

- Proposes a novel approach using a new composite feature vector of extracting elevation-related features. The proposed approach demonstrates the application of the envelope of ILD and high-frequency interaural phase difference (IPD) in vertical domain localisation.

- Presents a feature learning algorithm to extract elevation-related information in advance. This algorithm shows that the most valuable frequency region of interaural cues varies with the incident angle of the sound source. By selectively taking the interaural magnitude and phase features, the robustness of sound source localisation result can be improved.

- Propose a new statistical conditional model to describe the mapping between interaural features and source directions. The model takes the features' internal relevance into consideration, which further increase the robustness of the localisation performance. By combing with Direct-Path relative transfer function based dereverberation method, an improved localisation performance can be achieved in common in-door environment.

- Investigates the variation of ITD caused by different types of head rotations, and proposes an ITD-only dependent active localisation model. The study derives the relationship between the ITD variation and the vertical position of the sound source during the head movement.

- Indicates that the ITD variation caused by head rotation will significantly improve the accuracy and robustness of localisation results, and can be applied for multiple-source separation and localisation with minor adjustments to the system.

Chapter 2

Background: Spatial Hearing and Interaural Cues

Overview: *This chapter aims to provide background knowledge for the thesis. It introduces the mechanism of human binaural hearing, especially for sound source localisation. It also introduces the important localisation cues, such as ITD, IPD and ILD. Further, this chapter defines the HRTF and explains the relationship between the interaural cues and HRTF. In addition, it discusses the commonly used models for describing and simulating a realistic listening environment, and introduces the spatial feature extraction methods based on those models. It also provides a literature review of the different localisation mapping approaches, and highlights the problems with current binaural localisation.*

2.1 Introduction

Humans can perceive and analyse the surrounding environment by using two ears with remarkable sound source separation and localisation ability. In early research, the term cocktail party processing was created to describe humans capacity to intentionally select individual sound sources in a complex listening environment [52]. Previous research has demonstrated that the binaural system, which refers to the setup of two microphones placed on each side of a subjects head, plays a significant role in analysing auditory scenes and improving speech intelligibility in noisy and reverberating environments. In addition, the binaural system has demonstrated the possibility of prominent performance in sound localisation and separation, though it is not fully understood how such performance can be achieved using a similar system setup. Primarily, this chapter aims to review the fundamental mechanisms of binaural localisation of a sound source, alongside commonly used terminologies and assumptions. In addition, it provides a review of existing binaural localisation

methods and the recent related studies.

As a branch under sound source localisation, binaural localisation tackles the issue of estimating the position of the sound source based on the audio data captured by the binaural system only. Figure 2.1 summarises the pipeline of the binaural localisation system. A source signal $s(t)$ emitted from position $\mathbf{\Theta}$ is first filtered by the environment which includes the room impulse response (RIR) and the listener's anatomical structure (such as torso, head and pinna)and is then captured by the microphones. Such a process is described and modelled by a propagation model, and the most widely used model assumes the following condition:

- *Far field*: The distance of the sound source to the listener subject is greatly larger than the distance between the microphones, so that sound waves at the listener can be considered planar.

Normally, the sound source distance should be larger than 1.5 to 2 m, considering the average human head width. This assumption simplifies the problem, since the distance will not be considered, and the propagation model can be formulated as:

$$\begin{aligned} x_{\mathrm{l}}(\Theta, t) &= s(t) * a_{\mathrm{l}}(\Theta, t) + n(t) \\ x_{\mathrm{r}}(\Theta, t) &= s(t) * a_{\mathrm{r}}(\Theta, t) + n(t) \end{aligned} \tag{2.1}$$

where $x_{\mathrm{l}}(t)$ and $x_{\mathrm{r}}(t)$ represent the signals received by the left and right eardrums, respectively. The filtering effects of the environment and anatomical structure are modelled by the convolution between the source signal and the left and right position-dependent impulse responses, a_{l} and a_{r}, which is known as binaural room impulse response (BRIR), plus a background noise component $n(t)$. The listener's brain then processes those received signals and estimates the source position $\hat{\mathbf{\Theta}}$ based on a variety of spatial features. Therefore, the binaural auditory localisation refers to pinpointing the position information regarding the direction and distance of a sound source about the listener from the perceived acoustic signals. Lord Rayleighs duplex theory [10] was the first comprehensive analysis of the physics of binaural perception, and his theory remains valid to this day, with some extensions. Blauert [16] well summarised that the human listener's spatial hearing ability relies on both binaural cues (e.g., ITD and ILD) and monaural cues (e.g., spectral cues). In advance, the binaural auditory system plays an important role in separating and selectively focusing on individual sound sources in a cluttered acoustical environment with multiple sources, noise and reverberation, which is described as the well-known cocktail party problem [53]. Although localising sound sources is a

simple task for the brain, replicating and building a humanoid binaural localisation system with only two microphones is still a challenging problem. The primary goal of this chapter is to provide an understanding of the basic mechanisms underlying binaural spatial cues for far-field localisation and to discuss the main challenges in accurately estimating the source position in noisy and reverberant environments.

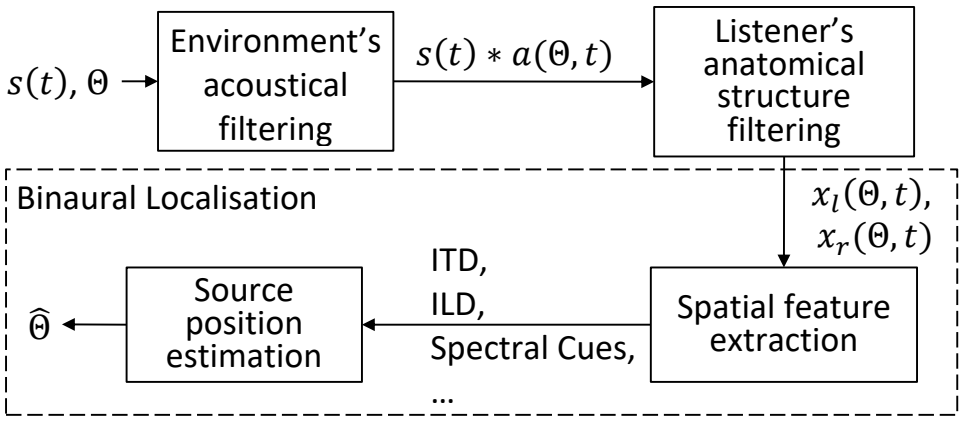


Figure 2.1: Binaural localisation system.

The chapter is organised as follows. First, two commonly used spatial coordinate systemsthe interaural-polar coordinate system and vertical plane coordinate systemare defined in Section 2.2, and the advantages of using either system are discussed. Second, the HRTF is introduced, as well as the localisation cues mentioned above. Third, Section 2.4 describes the modelling of an indoor acoustical environment including reverberation. Fourth, we briefly review the state-of-the-art localisation approaches in Section 2.5. Finally, the summary of this chapter is provided in Section 2.6.

2.2 Spatial Coordinate Systems

In spatial hearing, the relative position of a sound source to the listener is specified by direction and distance. In this thesis, we focus on pinpointing the source direction in a far-field scenariothat is, the distance is greater than 1.5 m, where the sound wave is approximated by the plane wave. Usually, the head centre at the level of entrances of the ear is set as the origin of the coordinate system, and the sketch in Figure 2.2 illustrates such placement of a binaural system with an origin O. Notably, the positive directions of x, y and z axes define the right, front and up of the subject,

respectively, and the planes Oxy, Oyz and Oxz define the horizontal, median and frontal planes, respectively. In addition, the plane parallel to the median plane is named the sagittal plane, and the plane parallel to the horizontal plane is named the transverse plane.

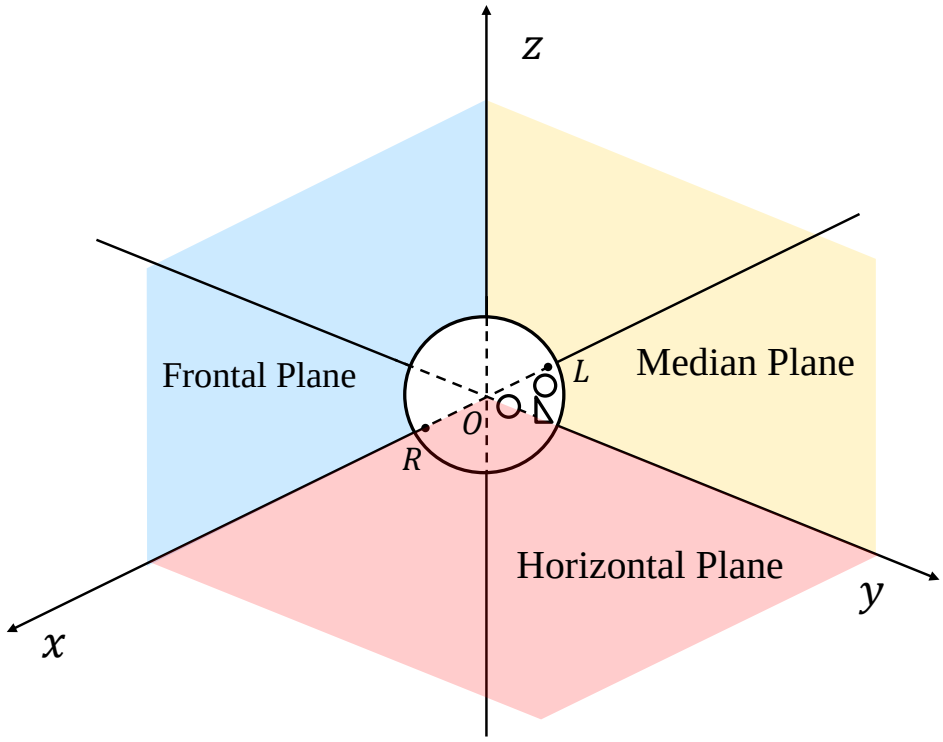


Figure 2.2: Binaural system setup.

With the placement shown in Figure 2.2, the directional vector of the source can be described by the Cartesian coordinates as $\mathbf{\Theta}_{\mathrm{C}} = (\mathrm{x}, \mathrm{y}, \mathrm{z})$ on the unit sphere, where $\sqrt{\mathrm{x}^2 + \mathrm{y}^2 + \mathrm{z}^2} = 1$. Otherwise, the source direction is usually specified by a polar-based coordinate system with two angles, azimuth and elevation. In the literature, the various polar-based coordinate systems define the lateral and elevation angle in different ways for specific applications. In those systems, the interaural-polar system and vertical-polar system are most commonly used. Both polar-based systems will be introduced in the following content.

2.2.1 Interaural-polar Coordinate System

Figure 2.3 displays the interaural-polar coordinate system. In such a coordinate system, we use a 2-D vector, $\mathbf{\Theta}_{\mathrm{IP}} = (\vartheta, \varphi)$, to denote the source direction, where

ϑ represents the azimuth and φ represents the elevation. By defining the source direction vector $\overrightarrow{OS}$ as the origin to the sound source, the ϑ is the angle between $\overrightarrow{OS}$ and the median plane, with $-90° \leq \vartheta \leq +90°$, and the φ is the angle from the positive y-axis to the projection of $\overrightarrow{OS}$ into the median plane, with $-90° < \vartheta \leq +270°$. The transformation between the Cartesian coordinates and interaural-polar coordinates on the unit sphere is:

$$\begin{cases} \mathrm{x} = \sin(\vartheta) \\ \mathrm{y} = \cos(\vartheta)\cos(\varphi) \\ \mathrm{z} = \cos(\vartheta)\sin(\varphi) \end{cases}$$

In the interaural-polar coordinate system, each azimuth value represents a specific sagittal plane, and the elevation defines a circle of positions that have the same distance differences to the left and right ear on that sagittal plane. Therefore, the interaural-polar coordinate system would offer great convenience to investigating the cone-of-confusion problem. The widely used CIPIC HRTF database [54] has adopted this coordinate system for its sampling scheme.

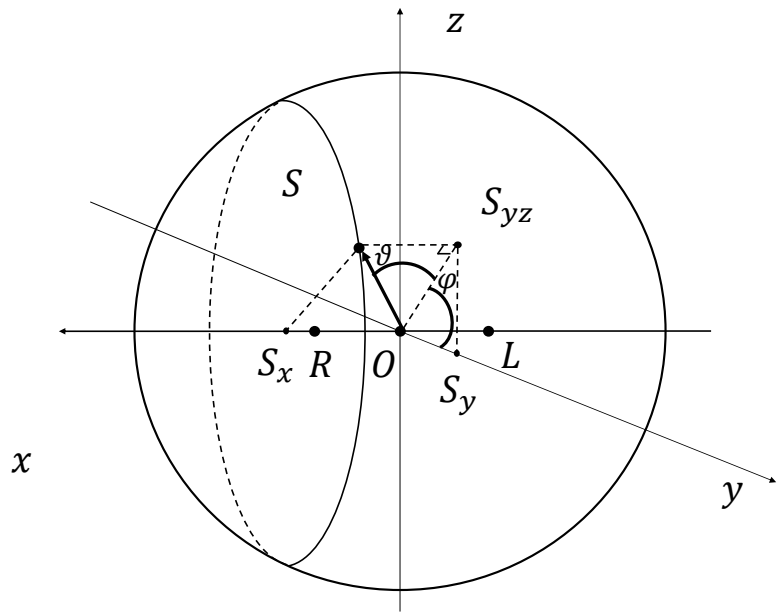


Figure 2.3: Interaural polar coordinate system. S, L, and R indicates the source, left and right ear, respectively.

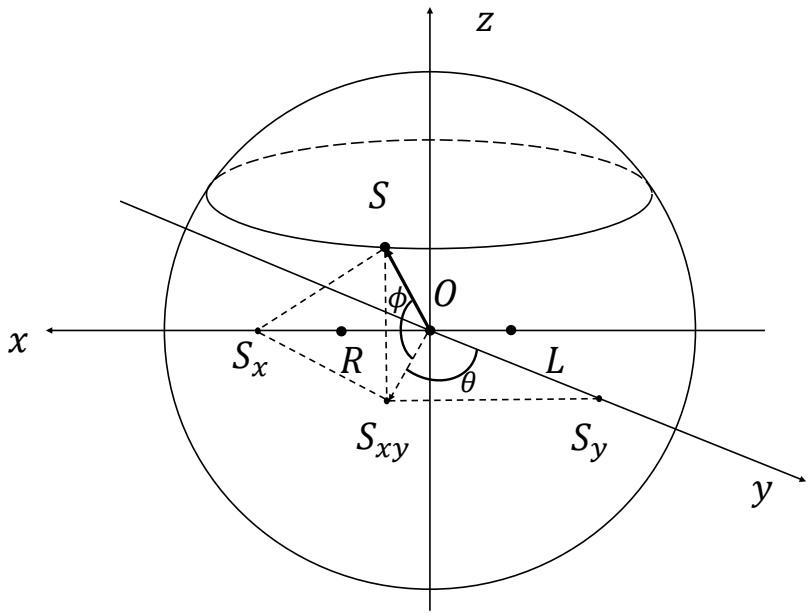


Figure 2.4: Vertical polar coordinate system.

2.2.2 Vertical Polar Coordinate System

Figure 2.4 shows the vertical-polar coordinate system. The angle between the horizontal projection of directional vector $\overrightarrow{OS}$ defines the azimuth θ with $0° \leq \theta < 360°$, and the angle between $\overrightarrow{OS}$ and the horizontal plane defines the elevation ϕ with $-90° \leq \phi \leq +90°$. Hence, the source direction vector is represented by $\mathbf{\Theta}_{\mathrm{VP}} = (\theta, \phi)$. The transformation between the Cartesian coordinates and vertical-polar coordinates on the unit sphere is:

$$\begin{cases} x = \cos(\phi)\sin(\theta) \\ y = \cos(\phi)\cos(\theta) \\ z = \sin(\phi) \end{cases}$$

In the vertical-polar coordinate system, each elevation value defines a specific transverse plane. By adopting this coordinate system, the yawing and pitching movement of the subject can be easily achieved by applying the turning angle on θ and ϕ directly [33]. Therefore, the vertical-polar coordinate system would be beneficial to investigate the dynamic cues in binaural localisation. The widely used MIT KEMAR dummy head HRTF database [55] has adopted this coordinate system for its sampling scheme.

In the following content of this thesis, we use $\mathbf{\Theta}$ to represent the source direction vector in a general sense, while $\mathbf{\Theta}_{\mathrm{C}}$, $\mathbf{\Theta}_{\mathrm{IP}}$ and $\mathbf{\Theta}_{\mathrm{VP}}$ specify the source direction

described in the Cartesian coordinate system, interaural-polar system and vertical-polar system, respectively.

2.3 Localisation Cues and Head-related Transfer Function

The task of localising a far-field sound source with the binaural setup uses spatial cues to estimate the direction parameters mentioned above. In the course of this thesis, we mainly consider three types of spatial cues extracted from the received signals: ITD, ILD and spectral cues. This section introduces the basic concepts of those spatial cues.

2.3.1 Interaural Difference

To reveal the directional relationship between the incidence sound wave and the interaural time delay, we define the plane through the source and the left and right ears as incident surface SLR in Figure 2.3 and 2.4. In a far-field scenario, the sound waves can be approximated as the plane waves propagating on the incident surface and scattered by the subject head. Figure 2.5 shows the curved path of the sound

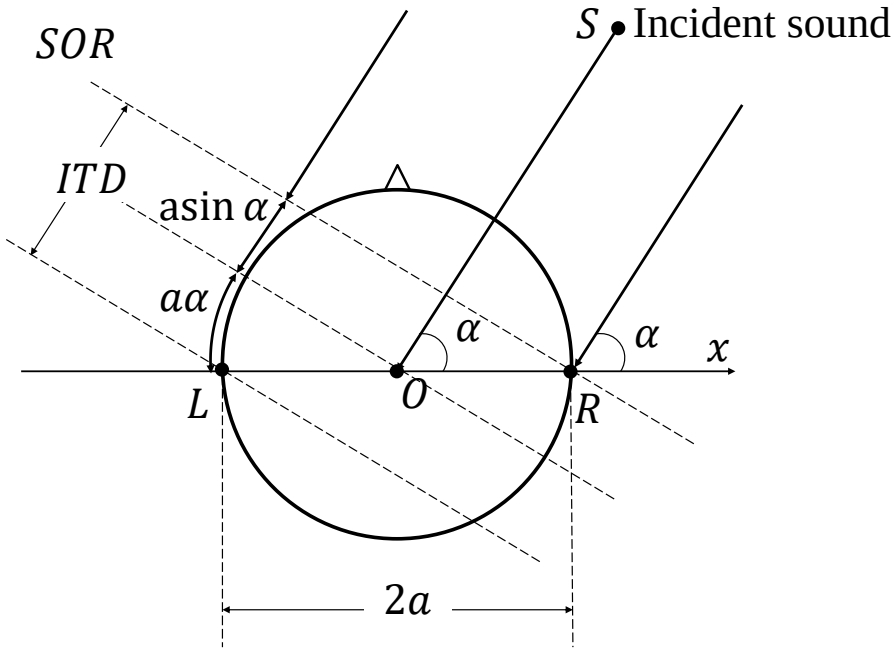


Figure 2.5: Incident sound wave and head scattering.

around the spherical head on the surface SLR. We define the incident angle as the angle between the ray from the head centre to the sound source, and the positive direction of the x-axis. The incident angle is denoted as α, while r represents the radius of the head.

Interaural Time Difference

With the spherical model, Woodworth and Schlosberg proposed the famous Woodworth formula [56] in 1954 as:

$$\tau(\alpha) = \frac{r}{c}(\cos\alpha + \alpha) \tag{2.2}$$

where τ indicates the ITD and c indicates the sound speed. Given that the spherical head is perfectly symmetrical, the incident angle α is set between 0° and 90°, and (2.4) can also be used to describe the ITD with the sound source in other quadrants. Figure 2.7 displays the calculated ITD using (2.2) with head radius $a = 0.07$

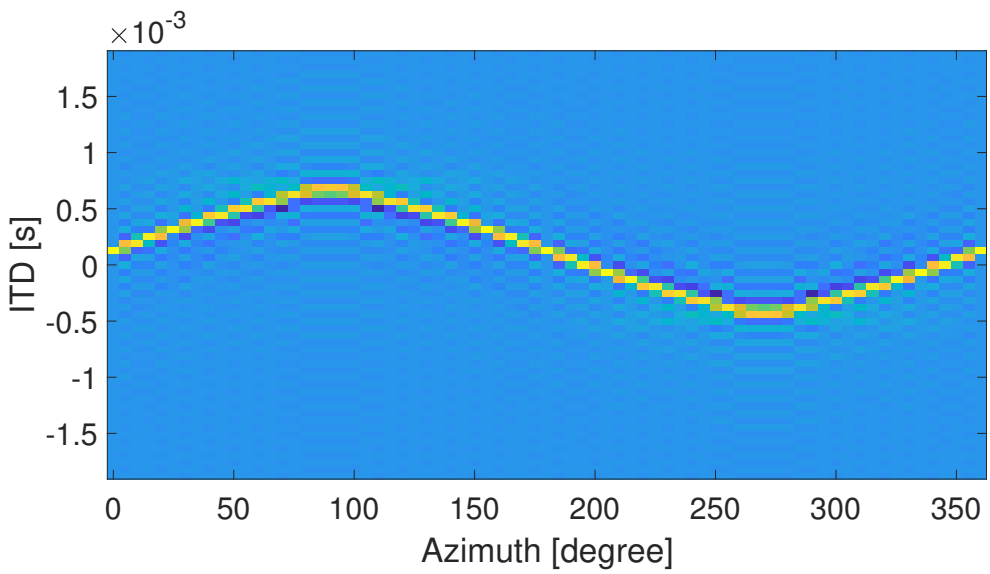


Figure 2.6: ITD on the horizontal plane. The simulated spherical head diameter $a = 0.07$ m.

m. Assuming the sound source is on the horizontal plane, the incident angle α is equivalent to the azimuth and the vertical-polar coordinate system is applied; thus, $0° \leq \theta < 360°$ and $\phi = 0°$. In the figure, the ITD and corresponding azimuth angle show a monotonous relationship at the frontal regionthat is, $\theta < 90°$ or $\theta \leq 270°$and the back regionthat is, $90° \leq \theta < 270°$. By exploiting this relationship, one may localise the sound source in the frontal or back hemisphere. However, ambiguity

exists when using ITD to distinguish the sound source location between the front and back, which is known as frontback confusion and will be extended to a cone of confusion in the 3-D space.

Interaural Phase Difference

Blauert [16] further proved that the IPD plays an essential role in estimating ITD below around 1.5 kHz, and the corresponding ITD can be obtained via IPD as [57]:

$$\tau_{\mathrm{p}}(\alpha, f) = \frac{\Delta\Phi}{2\pi f} = \frac{\Phi_{\mathrm{L}} - \Phi_{\mathrm{R}}}{2\pi f} \tag{2.3}$$

where $\tau_{\mathrm{p}}(\alpha, f)$ denotes the ITD derived from IPD, f indicates the frequency, and $\Delta\Phi = \Phi_{\mathrm{L}} - \Phi_{\mathrm{R}}$ represents the phase difference of sound pressures between the left and right ears. However, when the head diameter is larger than a wave length with a frequency above 1.5 kHz, the $\Delta\Phi$ would exceed 2π, and the τ_{p} would be ambiguous for the localisation.

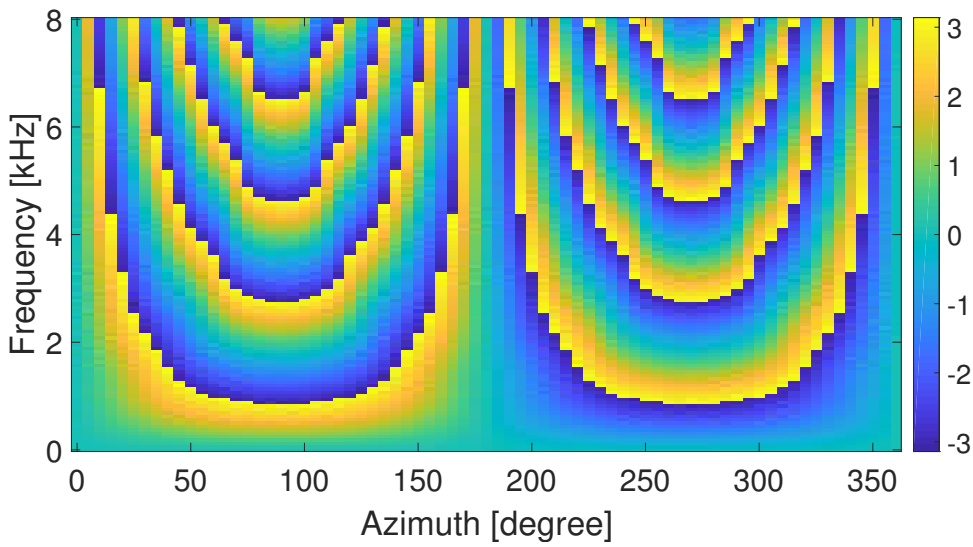


Figure 2.7: IPD on the horizontal plane. The simulated spherical head radius $a = 0.07$ m.

Figure 2.7 demonstrates the calculated IPD spectra with different azimuths on the horizontal plane. The source direction is distinguishable within the right side that is, $0° \leq \theta < 180°$and the left side that is, $180° \leq \theta < 360°$but the localisation still suffers from frontback confusion because of the symmetry of the sphere head model.

Interaural Level Difference

Another important clue frequently adopted by numerous localisation approaches is the ILD, which reflects the shadowing effect of the head. In Figure 2.5, the sound wave coming from the right side must travel around the head to reach the left ear, and, during this process, the amplitude of the sound pressure at the left ear is reduced compared with the right. The amplitude difference caused by this phenomenon leads to frequency- and direction-dependent ILD cues, which can be defined as in [57]:

$$ILD(\alpha, f) = 20\ \log_{10}\left|\frac{P_{\mathrm{L}}(\alpha, f)}{P_{\mathrm{R}}(\alpha, f)}\right| \tag{2.4}$$

where $P_{\mathrm{L}}(\alpha, f)$ and $P_{\mathrm{R}}(\alpha, f)$ are the sound pressures at the left and right ear with frequency f, respectively. Figure 2.6 shows the theoretic ILD with different frequencies for a spherical head model on the horizontal plane, measured in longitude.

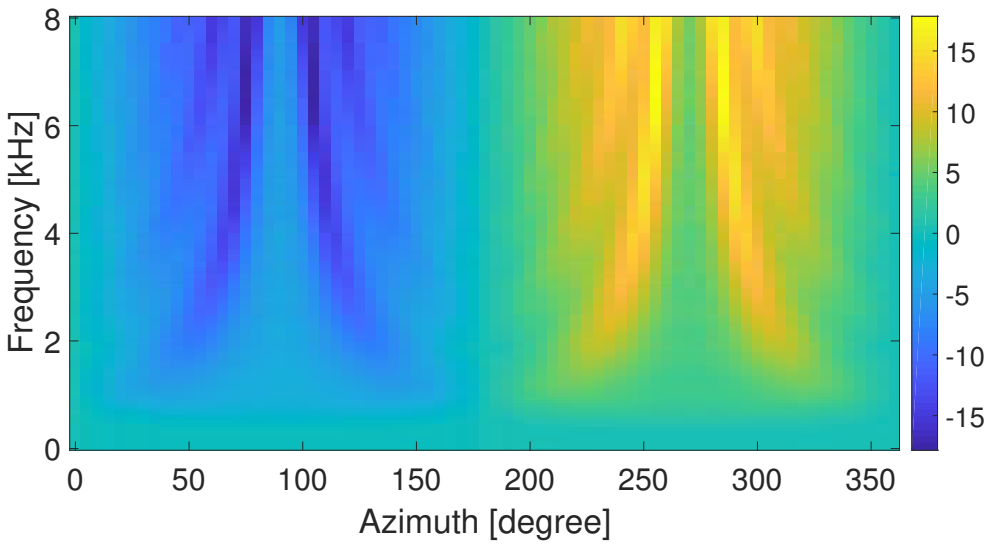


Figure 2.8: ILD on the horizontal plane. The simulated spherical head diameter $r = 0.07$ m.

In Figure 2.8, the ILD increases with increasing frequency, and a notable variation appears above 1 kHz, where the half-wavelength is smaller than the diameter of the head. The sign of ILD distinguishes the sources on the left and right side of the median plane, with the positive value indicating that the sound wave comes from the left side, and the negative value indicating that the sound wave comes from the right side, with the definition in (2.3). It should be noted that the maximum ILD does not occur at the position where the contralateral ear is opposite the source, e.g., $\alpha = 90°$ or $270°$. This phenomenon is because the multipath diffractions of

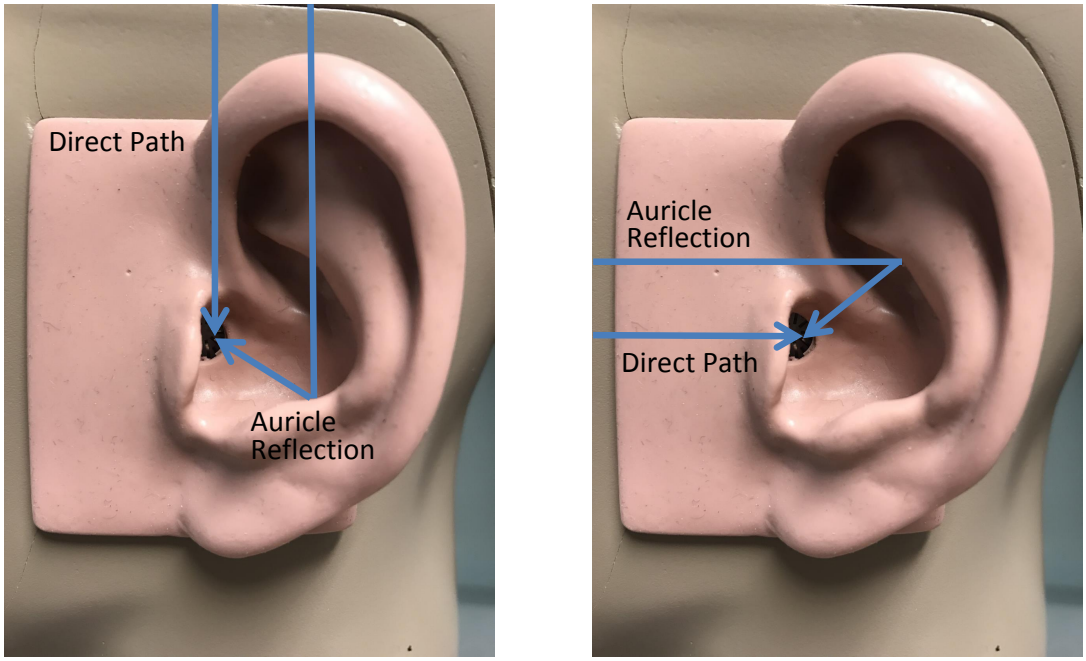
sound waves around the head lead to an in-phase interference in the sound pressure at the ear furthest from the source, which also causes multiple ridges and valleys at high frequencies . Still, the ILD also shows frontback ambiguity.

2.3.2 Head-related Transfer Function

As presented above, the spherical head model would suffer from not only the problem of front-back ambiguity, but also the cone-of-confusion in 3D space. The confusion cone is defined as a cone surface around the interaural axis [57]. The sound sources on this cone with same azimuth in interaural-polar coordinate system have same interaural differences with spherical head [54], which causes the confusion on distinguishing different elevation. However, the psychoacoustic experiments show that human listeners can distinguish sound sources located at the front and back, and estimate the elevation of sound sources [16, 57, 58]. This capability is owing to the human anatomical structure's filtering effect that is, the reflection and diffraction of torso, head and pinna. As a result of the frontback non-symmetry of the human subject [58], those filtering effects vary with the relative vertical locations of the sources, as shown in Figure 2.9, and the auditory system adopts these variations to help localise the sound source in a 3-D space. Academically, the filtering effects in the free-field because of the listener's anatomical structure are described by a linear time-invariant process, named HRTF, which is defined by:

$$\begin{aligned} H_{\mathrm{L}}(\boldsymbol{\Theta}, d, f, r) &= \frac{P_{\mathrm{L}}(\boldsymbol{\Theta}, d, f, r)}{P_{\mathrm{O}}(d, f)} \\ H_{\mathrm{R}}(\boldsymbol{\Theta}, d, f, r) &= \frac{P_{\mathrm{R}}(\boldsymbol{\Theta}, d, f, r)}{P_{\mathrm{O}}(d, f)} \end{aligned} \tag{2.5}$$

where P_{L} and P_{R} denote the sound pressure at the left and right ear, respectively, and $P_{\mathrm{O}}(d, f)$ denotes the sound pressure at the centre point of the head in the free-field. From (2.5), the HRTF is a transfer function that depends on the source location described by $(\boldsymbol{\Theta}, d)$ and the ear position described by the head radius r in the frequency domain. In particular, for a far-field scenario (i.e., $d > 1.5$ m), the HRTF difference caused by distance is negligible and the HRTFs are re-expressed as $H_{\mathrm{L}}(\boldsymbol{\Theta}, f, r)$ and $H_{\mathrm{R}}(\boldsymbol{\Theta}, f, r)$. Consequently, the IPD and ILD in (2.3) and (2.4)



(a) Sound from above (b) Sound from front

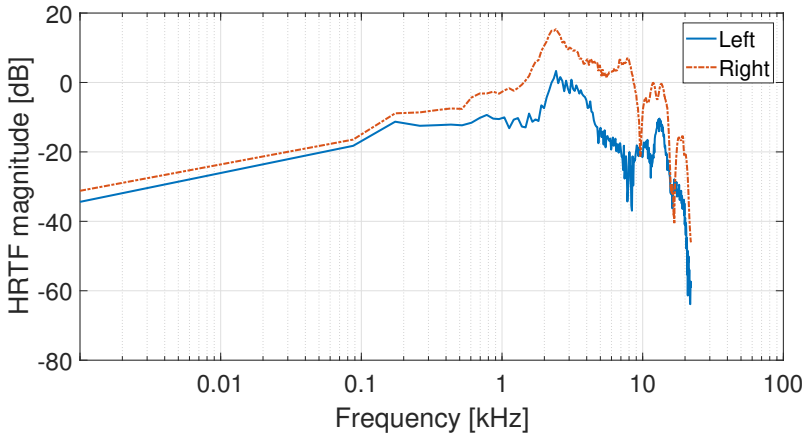
Figure 2.9: The Sound signals from up and frontal region filtered by the pinna

can be derived from HRTF as:

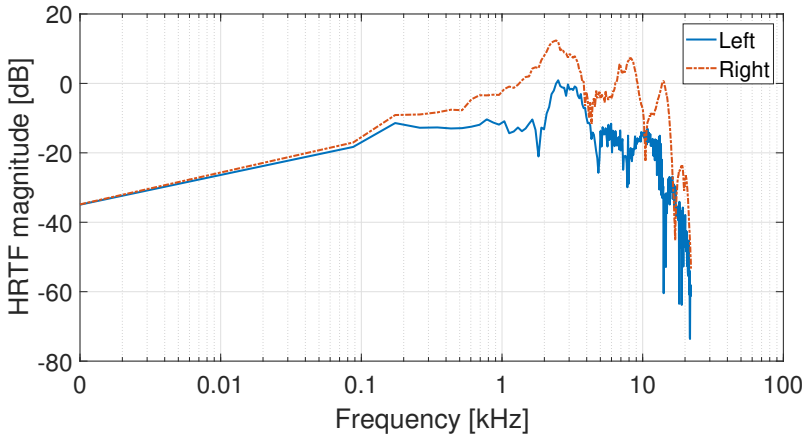
$$\begin{aligned} ILD(\mathbf{\Theta}, f, r) &= 20\log_{10}\left|\frac{H_{\mathrm{L}}(\mathbf{\Theta}, f, r)}{H_{\mathrm{R}}(\mathbf{\Theta}, f, r)}\right| \\ IPD(\mathbf{\Theta}, f, r) &= \angle\frac{H_{\mathrm{L}}(\mathbf{\Theta}, f, r)}{H_{\mathrm{R}}(\mathbf{\Theta}, f, r)} \end{aligned} \tag{2.6}$$

where $\angle$ indicates the wrapped angle calculation operation for the complex value. We can then obtain the interaural cues from HRTF data to localise a sound source with not only the azimuth, but also the elevation.

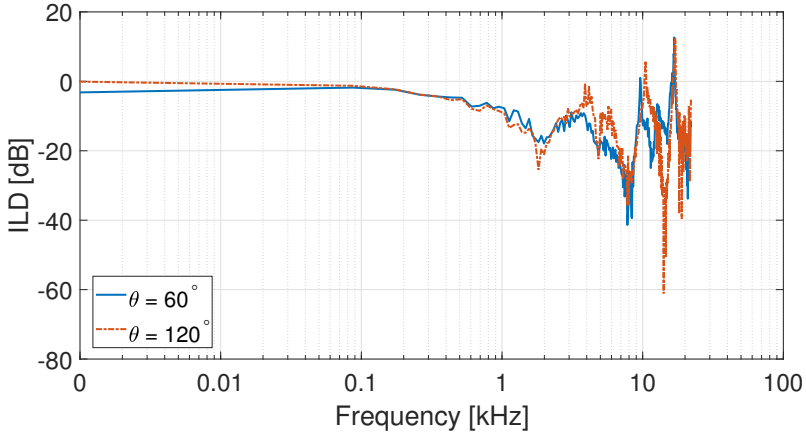
Figure 2.10 (a) and (b) presents an example of HRTFs from a KEMAR dummy head [55] at a position in the frontal (i.e., $\theta = 60°, \phi = 0°$) and back (i.e., $\theta = 120°, \phi = 0°$) regions. The two sources are located on a confusion cone, as mentioned above, and their corresponding HRTFs share high similarity on the envelope at low-frequency regions. However, unlike the spherical head model, the non-symmetric anatomical structure contributes to detailed differences in the frequency region above 2 kHz, and those differences are also reflected in the interaural cues. Figure 2.10 (c) demonstrates a more intuitive comparison of the ILD magnitude differences between the two positions, which become the key features to distinguish the front and back, and we will exploit these features to estimate the elevation of the sources.



(a) HRTF at $\theta = 60°, \phi = 0°$



(b) HRTF at $\theta = 120°, \phi = 0°$



(c) ILD at $\theta = 60°$ and $\theta = 120°$

Figure 2.10: Kemar dummy head HRTFs and ILD comparison between sound sources located at front and back region.

2.4 Room Acoustics and Spatial Signal Preprocessing

Another challenge in sound source localisation is the interference of reflections in enclosed spaces [41, 44]. In an indoor environment, the signals received by the ears are not only from the source directly, but also from the reflections of the surfaces, such as the walls, floor, ceiling and other objects in the room. Although the modelling of the enclosed sound field is the main content of room acoustics, the effects of reverberation should not be neglected when designing the localisation system, and evaluating the localisation performance of the system is such environments is important in practical applications [44]. Therefore, we use the existing modelling methods to generate various reverberant environments for testing purposes.

2.4.1 Reverberant Binaural Signal Modelling

In a reverberant room, the received signal can be modelled as the convolution of the original signal and in the time domain by (2.1). Specially, they can be decomposed into the direct component ($< 2ms$) and the reverberant component ($\geq 2ms$) [36]. The decomposed signals can be expressed by:

$$\begin{aligned} x_{\mathrm{L}}(t, \boldsymbol{\Theta}) &= \sum_{t=0}^{\mathrm{T_d}} h_{\mathrm{L}}^{0}(t, \boldsymbol{\Theta}) * s(t) + \sum_{t=\mathrm{T_d}+1}^{\infty} h_{\mathrm{L}}'(t, \boldsymbol{\Theta}) * s(t) \\ x_{\mathrm{R}}(t, \boldsymbol{\Theta}) &= \sum_{t=0}^{\mathrm{T_d}} h_{\mathrm{R}}^{0}(t, \boldsymbol{\Theta}) * s(t) + \sum_{t=\mathrm{T_d}+1}^{\infty} h_{\mathrm{R}}'(t, \boldsymbol{\Theta}) * s(t) \end{aligned} \tag{2.7}$$

where $h_{\mathrm{i}}^{0}(t, \boldsymbol{\Theta})$ and $h_{\mathrm{i}}'(t, \boldsymbol{\Theta})$ denote the direct component impulse response and reverberate component impulse response at source direction specified by $\boldsymbol{\Theta}$, respectively. The reverberate component is made up of two parts, i.e. early and late reverberations. The reflection responses within 50-100ms are defined as early reverberations, while the responses beyond this duration are defined as late reverberation [36]. The early reverberation is believed as the main reason of the speech coloration and late reverberation can result in over-masking effect. In the model shown in (2.7), the time span of the direct sound is given by $\mathrm{T_d}$, which includes the sound propagation from the source to the listener's ear drums. Thus, we can assume that the source position-related spatial information mentioned in previous sections, e.g., head-related impulse response (HRIR), is embedded in h_{L}^{0} and h_{R}^{0} [36].

Though there are many approaches for reverberation simulation, e.g. Finite-difference time-domain method, ray-tracing and etc., the convolution based approaches have advantages on the flexibility of using HRTFs [57]. Based on the convolution model from (2.1) and (2.7), there are two types of simulation processes for reverberating binaural signals: statistical models and geometric reverberation models [36, 42]. The former uses the statistical model to simulate reverberations by assuming the reverberation is a collection of far-field sources [36]. The other type of modelling process uses reflecting surfaces (e.g., walls, floor and ceiling) to generate a geometric representation of reverberation, which embeds the directional interferences of the early reverberation. Hence, we employ the geometric reverberation model to simulate an enclosed acoustical environment for testing purposes. One of the famous geometrical methods, the image-source method [59], models the

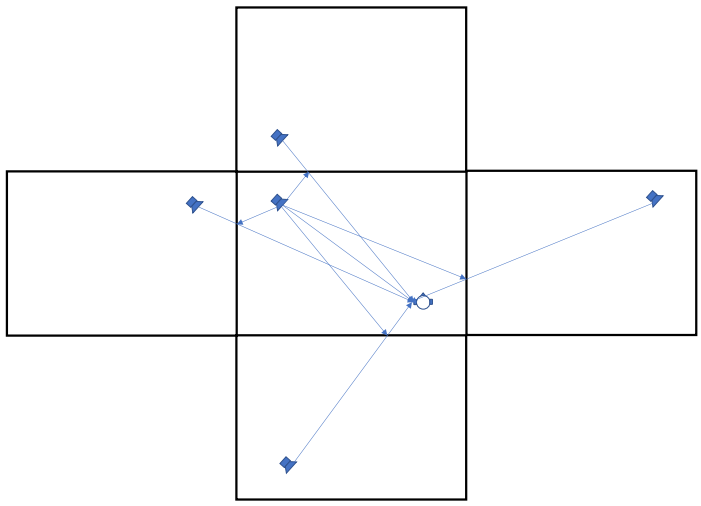


Figure 2.11: Illustration of image source method with depth of 1 in 2-D for left receiver.

reverberation as a collection of virtual sources located at the mirrored image point of the reflection surfaces. Figure 2.11 provides an example of the image-source method with an image depth of 1 and the sound propagation arrays at the left ear. The BRIR is expressed as the summation of the impulse responses convolved with

correlated HRIRs of multiple image sources, which is formulated by:

$$h_{\mathrm{i}}(t,\boldsymbol{\Theta}) = \sum_{\boldsymbol{p}=0}^{1}\sum_{\boldsymbol{r}=-\infty}^{\infty} \beta_{x^+}^{|a-p|}\beta_{x^-}^{|p|}\beta_{y^+}^{|b-q|}\beta_{y^-}^{|q|}\beta_{z^+}^{|c-j|}\beta_{z^-}^{|j|} \cdot \frac{\sigma[t-(|\boldsymbol{R_p}+\boldsymbol{R_r}|/c)]}{4\pi|\boldsymbol{R_p}+\boldsymbol{R_r}|} * \delta(\tau_{\boldsymbol{p}}) * h_{\mathrm{i}}^{0}(t,\boldsymbol{\Theta}) \tag{2.8}$$

where $\sigma(\cdot)$ indicates the Delta function, and $*$ represents the convolution operation. The β_{x^+}, β_{x^+}, β_{y^+}, β_{y^+}, β_{z^+} and β_{z^+} define the reflection coefficients of six walls in the positive and negative directions of the axes. The $\boldsymbol{R_p}$ is the sources directional vector defined on the source position (x, y, z), microphone position (x', y', z') and integer 3-vector $\boldsymbol{p} = (p, q, j)$ as:

$$\boldsymbol{R}_p = (x \pm x' + 2px', y \pm y' + 2qy', z \pm z' + 2jz')$$

The reflection image vector $\boldsymbol{R_r}$ is expressed in terms of the image order integer triplet, $\boldsymbol{r} = (b_x, b_y, b_z)$, as:

$$\boldsymbol{R}_r = 2(b_x L_x, b_y L_y, b_z L_z)$$

Therefore, we simulate the reverberant signal of a shoebox room by (2.8), with the reflection coefficients on each wall and pre-measured free-field HRIRs. In the simulation, the clean speech signal is used as source signal, which is because localising a human subject is the major use case as mentioned in the previous chapter. Besides, considering an indoor environment, the sound source is assumed to be static since the source movement speed is very slow comparing with the sound speed and normal computer processing cycle. According to (2.8), the reverberation effects of the enclosed environment become a linear combination of multiple directional signals, which result in environmental-dependent interference and distortion on the spatial features, and eventually degrade the localisation performance of the listener.

2.4.2 Convolutive Transfer Function Approximation

Equation (2.1) represents the received signal as a linear convolution in the time domain. In many audio signal processing methods, the time domain signal needs to be transferred into the frequency domain, such as the most commonly used short-time Fourier transform (STFT), which divides the signal sequences into overlapping timeframes and then transfers the signal fragments into the frequency domain using

Fourier transform. Then, (2.1) can be approximated in the T-F domain as:

$$\begin{aligned} X_{\mathrm{L},k}(f,\boldsymbol{\Theta}) &= A_{\mathrm{L}}(f,\boldsymbol{\Theta}) \cdot S_k(f) + N_{\mathrm{L},k}(f) \\ X_{\mathrm{R},k}(f,\boldsymbol{\Theta}) &= A_{\mathrm{R}}(f,\boldsymbol{\Theta}) \cdot S_k(f) + N_{\mathrm{R},k}(f) \end{aligned} \tag{2.9}$$

where k is the time frame index; f is the frequency sub-band index; and $X_{\mathrm{i},k}(f,\boldsymbol{\Theta})$,$A_{\mathrm{i}}(f,\boldsymbol{\Theta})$ $S_k(f)$ and $N_{\mathrm{i},k}(f)$ indicate the binaural signal spectra, relative transfer function, source signal spectra and additive noise spectra in the T-F domain, respectively. The approximation in (2.9) is known as the multiplicative transfer function (MTF) [60]. This approximation uses the assumption that the speech is stationary in each time frame, since we restrict the time frames to be relatively short ($< 40ms$). Normally, such approximation can be applied in a moderate reverberant environment, where the impulse response length is smaller than the time frame, and the relative transfer function is assumed to be static over time frames. However, in a reverberant room with several hundred milliseconds T_{60}, the MTF approximation would not be applicable, since the required time frames should be larger than T_{60}, while the source signal cannot be considered stationary [61].

A possible solution for this over-length problem can be found in [61, 62]. Ronen et al. recently proposed a relative transfer function (RTF) identification method for speech sources in reverberant environments. The proposed method is based on the convolutive transfer function (CTF) approximation, which is able to express the linear convolution in the time domain as a linear convolution in the STFT domain. Without the restriction of MTF approximation, the CTF approximation enables a more accurate representation of long impulse responses using short timeframes. In the CTF approximation, (2.1) can be represented as a convolution on the frame index k by:

$$\begin{aligned} \tilde{X}_{\mathrm{L},k}(f,\boldsymbol{\Theta}) &= \sum_{k'=0}^{Q_f-1} H_{\mathrm{L},k'}(f,\boldsymbol{\Theta}) \cdot S_{k-k'}(f) + N'_{\mathrm{L},k}(f) \\ \tilde{X}_{\mathrm{R},k}(f,\boldsymbol{\Theta}) &= \sum_{k'=0}^{Q_f-1} H_{\mathrm{R},k'}(f,\boldsymbol{\Theta}) \cdot S_{k-k'}(f) + N'_{\mathrm{R},k}(f) \end{aligned} \tag{2.10}$$

where Q_f is the number of causal filter coefficients that is dependent on the reverberation time at the f-th frequency bin, and the additive noise component $N'_{\mathrm{i},k}(f)$ is assumed to be individually wide-sense stationary and uncorrelated with the source signal . In the approximation above, the long RIR in the time domain has been

transferred into Q_f short spectra frames, as noted by $H_{\mathrm{i},k'}$. Particularly, the first component of $H_{\mathrm{i},k'}$ with $k' = 0$ can be considered as the STFT of the direct component h_{i}^0 in (2.7).

2.4.3 Interaural Cues Extraction

Another crucial problem is to extract the spatial features from the acoustical model in the previous section. Many efforts have been made to obtain the direct component from the received signals that are filtered by the enclosed environment [36,42,44,63–65]. Most efforts have focused on extracting the interaural transfer function (ITF), which is defined by the ratio of left and right HRTFs as:

$$ITF(\mathbf{\Theta}, f, r) = \frac{H_{\mathrm{L}}(\mathbf{\Theta}, f, r)}{H_{\mathrm{R}}(\mathbf{\Theta}, f, r)} \tag{2.11}$$

According to the above definition and (2.6), the ILD and IPD can be easily obtained by taking the magnitude component in the logarithm and the phase component of ITF. Therefore, the nature of extracting the interaural cues can be considered as estimating the ITF from received binaural signals.

In a less reverberant environment, where the length of RIR is shorter than the window frame, and the noise component is zero-mean and independent of the energy of the source signal, we can estimate the ITF by taking the expectation of the ratio between the left and right received signals [20,30]. The estimation can be expressed by:

$$I\hat{T}F(\mathbf{\Theta}, f, r) = \frac{1}{K}\sum_{k=0}^{K} \frac{X_{\mathrm{L,k}}(\mathbf{\Theta}, f, r)}{X_{\mathrm{R,k}}(\mathbf{\Theta}, f, r)} \tag{2.12}$$

This method can successfully cancel the source signal, and the remaining ratio should be dependent on HRTFs only. Keyrouz et al. proposed a cross-correlation-based localisation algorithm based on Source Cancellation Algorithm (SCA) and a pre-calculated ITF database to localise a single sound source [20, 30], which displayed remarkable performance in the presence of additive white noise. However, given that (2.12) is still based on the MTF model and STFT, the performance based on this extraction method could degrade because of the reverberation.

Cepstral Subtraction

In a closed environment, the presence of reverberation can originate gross localisation errors, as the reverberant components cause self-masking and overlap masking, and

those phonemes interfere with the spatial cue extraction from the early reflections [9]. Therefore, some signal pre-processing becomes necessary [66], and the cepstrum analysis is used to perform this task.

With a cepstral operation, the convolution of two signals in the time domain will be transformed to addition in the quefrency domain, which enables the cepstral pre-filtering. The cepstral pre-filtering has been shown to be effective in reducing the influence of early reverberations [64, 67, 68], which is based on the assumption that the minimum phase component (MPC) of the source signal in the cepstral domain varies between different frames with zero-mean, while the MPC of the RIR is stationary and can be estimated by averaging over frames. The pre-filtered signal can then be obtained by subtracting the estimated RIR MPC from the received signal cepstrum. Finally, the interaural cues can be extracted from the signal after it is transformed back to the time domain. This is detailed further in Chapter 3.

However, although the cepstral operation can decrease the adverse effect of reverberations and lead to an improvement in azimuth localisation, the contribution to elevation is limited when expanded to elevation localisation because the elevation-related features are more complex and less robust compared with the azimuth, and are more easily affected by late reverberation [36, 69].

Direct PathRelative Transfer Functions estimation

To tackle the localisation scenario that has reverberations systematically, the CTF model is applied, and an alternate ITF extraction method has been proposed [44]. Assuming that the signals from both channels in (2.10) follow the CTF approximation and can be rewritten in vector form with the causal filter coefficient Q_f as:

$$\tilde{\mathbf{x}}_{\mathrm{i},k,f} = [\tilde{X}_{\mathrm{i},k,f}, \tilde{X}_{\mathrm{i},k-1,f}, ...\tilde{X}_{\mathrm{i},k-Q_f+1,f}]^{\mathbf{T}}$$

where $(\cdot)^{\mathbf{T}}$ is the matrix transpose, Θ is omitted and f is simplified as a sub-index, then the short-term spectra of RIRs can be expressed in a similar vector form by:

$$\mathbf{h}_{\mathrm{i},k,f} = [H_{\mathrm{i},0,f}, H_{\mathrm{i},1,f}, ...H_{\mathrm{i},Q_k-1,f}]^{\mathbf{T}}$$

Now, based on (2.10), we can write the following equation as:

$$\tilde{\mathbf{x}}^{\mathbf{T}}_{\mathrm{L},k,f}\mathbf{h}_{\mathrm{R},k,f} = \tilde{\mathbf{x}}^{\mathbf{T}}_{\mathrm{R},k,f}\mathbf{h}_{\mathrm{L},k,f} \tag{2.13}$$

Then, by dividing $H_{\mathrm{R},0,f}$ on both sides, (2.13) can be reorganised as:

$$\tilde{X}_{\mathrm{L},k,f} = \mathbf{z}_{k,f}^{\mathbf{T}} \mathbf{g}_f \tag{2.14}$$

where

$$\begin{aligned}\mathbf{z}_{k,f}^{\mathbf{T}} =& [\tilde{X}_{\mathrm{R},k,f}, \tilde{X}_{\mathrm{R},k-1,f}, ...\tilde{X}_{\mathrm{R},k-Q_f+1,f}, \\ & \tilde{X}_{\mathrm{L},k-1,f}, \tilde{X}_{\mathrm{L},k-2,f}, ...\tilde{X}_{\mathrm{L},k-Q_f+1,f}]^{\mathbf{T}}\end{aligned}$$

$$\mathbf{g}_f = \left[\frac{H_{\mathrm{L},0,f}}{H_{\mathrm{R},0,f}}, ..., \frac{H_{\mathrm{L},Q_f-1,f}}{H_{\mathrm{R},0,f}}, -\frac{H_{\mathrm{R},1,f}}{H_{\mathrm{R},0,f}}, ..., -\frac{H_{\mathrm{R},Q_f-1,f}}{H_{\mathrm{R},0,f}}\right]^{\mathbf{T}}.$$

Therefore, the target now is to extract the first element of $\mathbf{g}_f$, which is defined as the direct path RTF (DP-RTF) $\hat{g}_0(f)$. By multiplying $\tilde{X}^*_{\mathrm{L},k,f}$ on both sides, (2.14) can be re-expressed in terms of auto-power spectral density (PSD) $\Phi_{\tilde{X}_{L,k,f}}$ and cross-PSD $\tilde{\mathbf{\Phi}}_{k,f}$ by:

$$\Phi_{\tilde{X}_{L,k,f}} = \tilde{\mathbf{\Phi}}_{k,f}^{\mathbf{T}} \mathbf{g}_f \tag{2.15}$$

Finally, the estimation of $\mathbf{g}_f$ can be given by the least-squares solution of (2.15) as:

$$\hat{\mathbf{g}}_f = (\tilde{\mathbf{\Phi}}_{k,f}^{\mathbf{T}} \tilde{\mathbf{\Phi}}_{k,f})^{-1} \tilde{\mathbf{\Phi}}_{k,f}^{\mathbf{T}} \Phi_{\tilde{X}_{L,k,f}} \tag{2.16}$$

The auto-PSD term $\Phi_{\tilde{X}_{L,k,f}}$ and cross-PSD term $\tilde{\mathbf{\Phi}}_{\mathbf{p},\mathbf{k}}$ can be obtained by an inter-frame spectral subtraction method proposed in [44], and the auto-STFT spectra $\Phi_{\tilde{X}_{L,k,f}}$ and cross-STFT spectra $\hat{\mathbf{\Phi}}_{\mathbf{p},\mathbf{k}}$, defined by:

$$\hat{\Phi}_{\tilde{X}_{L,k,f}} = \frac{1}{D} \sum_{d=0}^{D-1} X_{L,p-d,k} X^*_{L,p-d,k}$$

$$\begin{aligned}\hat{\mathbf{\Phi}}_{\mathbf{p},\mathbf{k}} =& [E\{X_{R,p,k} X^*_{L,p,k}\}, ..., E\{X_{R,p-Q_k+1,k} X^*_{L,p,k}\}, \\ & E\{X_{L,p-1,k} X^*_{L,p,k}\}, ..., E\{X_{L,p-Q_k+1,k} X^*_{L,p,k}\}]^{\mathbf{T}}\end{aligned}$$

Now, the ITF of the direct path to the sound source can be approximated by estimating $\hat{g}_0(f)$, and the interaural cues can subsequently be estimated. It is necessary to note that the noise component has been neglected, although an extraction method with the presence of additive noise is provided in [44]. The influence of noise will be tackled by the localisation model, which will be discussed in Chapter 5.

2.5 Binaural Localisation Algorithm

The localisation algorithm in this thesis refers to mapping a given extracted feature to the corresponding source location. One simple yet classic method to undertake this is by using the propagation model in the free-field and far-field, such as the spherical head model [56] mentioned in the previous sections. Further, when seeking to obtain a more accurate or robust localisation, or localising the sound sources in both horizontal and vertical directions, many other types of features are needed, which require an exploration of the optimisation of the localisation solution space.

The most straightforward method is to apply a grid-search [70], in which a mapping function (e.g., cross-correlation function) is used throughout the whole sound source localisation space and the output is recorded for each candidate sound source location. Then the recorded position with the maximum value of the mapping function output is treated as the solution of the localisation. In particular, this is the most commonly used mapping approach for multiple-source location estimation, and has been applied in two important examples the subspace orthogonality feature of MUSIC [21, 22, 24] and the steered-response of a delay-and-sum beamformer [71].

In addition, there are further types of mapping methods besides grid-search, which share a similar concept of training the mapping function using recorded feature data with known source locations. Consequently, in all of these methods, the propagation model is encoded in the trained mapping functions, and various methodologies are applied to obtain the mapping functions, such as the Gaussian mixture model (GMM) [41], locally-linear regression and manifold learning [32, 33, 72], and the recently well-developed neural networks [50, 73, 74].

2.5.1 Broad-band Approaches

The most commonly used approach is to estimate the time difference of arrival (TDOA) signal waves between the binaural microphones, and then the sound incidence angle can be deduced by the estimated time difference and microphone placement geometric relationship.

The most well-known time delay estimation method is the so-called generalised cross-correlation (GCC), which was proposed by Knapp and Carter [23]. The estimated TDOA $\hat{\tau}$ is obtained by searching the time delay τ, which results in the maximum cross-correlations between the left and right received signals, which can be expressed as:

$$\hat{\tau}_k = \arg\max_{\tau} \frac{1}{2\pi} \sum_f W(f) X_{\mathrm{L,k}}(f) X^*_{\mathrm{R,k}}(f) \tag{2.17}$$

where $W(f)$ is a frequency-dependent weighting function. Ideally, (2.17) could return the accurate TDOA between the microphones by detecting the most significant peak in the GCC function. However, performance could be degraded by the influence of reverberation and the background noise. Therefore, to highlight the peak of actual true time delay, the phase transform (PHAT) was proposed and became the most popular weighting function. The PHAT whitens the cross-spectrum by setting the weighting function as $W_{\mathrm{PHAT}}(f) = |X_l(f)X^*_r(f)|^{-1}$. The PHAT weighting has been proved to provide a robust localisation performance with the presence of reverberation [75,76]. Once the estimated time delay is obtained, the sound direction can be obtained by a table-lookup procedure. The mapping relationship of the actual azimuth angle and the time delay follows a monotonic relationship, and the mapping table is usually pre-measured [77].

The approaches based on TDOA are mostly used in horizontal plane localisation. To estimate the vertical position of sound sources, the spectral information and HRTF is exploited.

2.5.2 HRTF Localisation Approaches

The core concept of the HRTF-based localisation algorithms is to maximise the correlation between the left and right receivers by pinpointing a HRTF pair across all possible positions. Then, by returning the corresponding position of the identified HRTF pair, the sound source location can be localised. Given that the identification of the HRTF pair is implemented based on a stored HRTF database, the localisation space can be expanded into 3-D space by measuring the HRTFs at different azimuths and elevations [54, 55].

In this section, three HRTF-based binaural localisation algorithms are briefly introduced: the matched filtering approach, source cancellation approach and cross-convolution approach.

Matched Filtering Approach

The key idea of the matched filtering approach is to reverse the HRTF filtering effect on received signals. In an ideal free-field, the received signals modelled in (2.9) are

simplified as:

$$\begin{aligned} X_{\mathrm{L},k}(f) &= H_{\mathrm{L}}(f) \cdot S_k(f) \\ X_{\mathrm{R},k}(f) &= H_{\mathrm{R}}(f) \cdot S_k(f) \end{aligned} \tag{2.18}$$

The RTF A_{L} and A_{R}, which are the STFTs of a_{L} and a_{L}, are simplified as direct path HRTFs, $H_{\mathrm{L}}(f)$ and $H_{\mathrm{R}}(f)$, in an ideal case. The localisation algorithm seeks to filter the received signals $X_{\mathrm{L},k}(f)$ and $X_{\mathrm{R},k}(f)$ with the correct reversed HRTF pairs, $H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta})$ and $H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta})$, and then obtain the original mono-source signal, $S_k(f)$:

$$\begin{aligned} \hat{S}_{k,\mathrm{L}}(f, \boldsymbol{\Theta}) &= H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta}) \cdot X_{\mathrm{L},k}(f) \\ &= H_{\mathrm{R}}^{-1}(f, \boldsymbol{\Theta}) \cdot X_{\mathrm{R},k}(f) \\ &= \hat{S}_{k,\mathrm{R}}(f, \boldsymbol{\Theta}) \end{aligned} \tag{2.19}$$

The estimated source location $\hat{\boldsymbol{\Theta}}$ is given by maximising the cross-correlation between $\hat{S}_{k,\mathrm{L}}(f, \boldsymbol{\Theta})$ and $\hat{S}_{k,\mathrm{L}}(f, \boldsymbol{\Theta})$:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \{\hat{S}_{k,\mathrm{L}}(f, \boldsymbol{\Theta}) \otimes \hat{S}_{k,\mathrm{R}}(f, \boldsymbol{\Theta})\} \tag{2.20}$$

where $\otimes$ denotes a cross-correlation operation. However, in the match filtering approach, the inverse of HRTF might be unstable because the HRTF linear-phase component encodes the ITDs, so the unstable version must be stably approximated to contain all the direction-dependent information [17].

Source Cancellation Algorithm

As an extension of the match filtering approach, the source cancellation algorithm replaces the cross-correlation between $H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta}) \cdot X_{\mathrm{L},k}(f)$ and $H_{\mathrm{R}}^{-1}(f, \boldsymbol{\Theta}) \cdot X_{\mathrm{R},k}(f)$ by $X_{\mathrm{L},k}(f)/X_{\mathrm{R},k}(f)$ and $H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta})/H_{\mathrm{R}}^{-1}(f, \boldsymbol{\Theta})$, and (2.20) is reformed as:

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \left\{ \frac{X_{\mathrm{L},k}(f)}{X_{\mathrm{R},k}(f)} \otimes \frac{H_{\mathrm{L}}^{-1}(f, \boldsymbol{\Theta})}{H_{\mathrm{R}}^{-1}(f, \boldsymbol{\Theta})} \right\} \tag{2.21}$$

Therefore, the source cancellation algorithm can avoid the inverse calculation of HRTFs and store the ratio of HRTF pairs as the database in advance [18, 19].

Convolution-based Approach

The convolution-based approach is proposed to avoid the unstable problem caused by the inversion of HRTFs [19]. The method applies the associative property of convolution operation, and the relationship in (2.19) is reorganised as:

$$\begin{aligned}\hat{S}_{k,\mathrm{L}}(f,\boldsymbol{\Theta}) &= H_{\mathrm{R}}(f,\boldsymbol{\Theta}) \cdot X_{\mathrm{L},k}(f) \\ &= H_{\mathrm{L}}(f,\boldsymbol{\Theta}) \cdot X_{\mathrm{R},k}(f) \\ &= \hat{S}_{k,\mathrm{R}}(f,\boldsymbol{\Theta})\end{aligned} \tag{2.22}$$

and the localisation is given by the same cross-correlation operation as defined in (2.20).

2.5.3 Supervised Approaches

The table-lookup-based localisation approaches based on correlation maximisation have demonstrated success in an ideal case [21,22,24,71]. However, in realistic practice cases, the captured binaural signals are interfered with noise and reverberation, and many efforts have sought to diminish those effects and extract clean binaural spatial cues from noisy environments [36, 42, 44, 65]. In recent years, many studies have attempted to optimally deduce the source location from interfered binaural signals based on supervised-learning strategies [32, 33, 41, 50, 72–74, 78–82]. In these methods, the propagation model has been encoded in the trained mapping functions, and various methodologies have been applied to obtain the mapping functions. Compared with using a mapping function in a table-lookup approach – that is, (2.20) and (2.21) in HRTF-based approaches – a supervised-learning approach provides a probabilistic framework for multiple layers of variables, and the information from different frequency channels and different features can be analysed jointly.

Another distinct advantage of the supervised-learning approach is the environmental flexibility of localisation models, which is thanks to the multi-conditional training process. In the supervised-learning process, the training database is constructed based on prior knowledge, which refers to a collection of data pairs of binaural spatial features and the known corresponding sound location. Such data pairs are often collected or simulated in different environmental configurations, which incorporate various possible variations of spatial features [81]. Different from storing a reference database in the table-lookup approaches, the supervised-learning approach

generates a probabilistic model of the variation binaural cues based on the multi-conditional training. Thus, a more robust localisation performance across different acoustical environment can be obtained. In addition, supervised-learning-based approaches have been widely applied in the field of sound source segregation. The following content briefly reviews some classic supervised-learning approaches that have been applied in binaural localisation.

Gaussian Mixture Model Based Localisation

We first introduce the classic learning approach used in binaural localisation areathe GMM. The GMM-based localisation framework is very flexible, and can rapidly incorporate multiple-source location-dependent features. For instance, [83] proposed Model-based EM Source Separation and Localisation (MESSL) for multiple sources separation and localisation, which uses GMM to mixing multiple probabilistic single source models based on sources and time delay. In addition to tackle multi-sources problem, GMM can also be used to expand the localisation space from the horizontal domain to vertical domain by integrating elevation-related spectral features in GMM models [84–86]. In the following content, we introduce a GMM-based classifier that approximates the 2-D feature distribution of ITDs and ILDs as an example [81]. Feature distribution of ITDs and ILDs as an example [81].

In multi-conditional training, the training data can be collected from the simulated data of various acoustic configurations, which involves different source and listener positions and different reverberation and noise conditions. The training process can then be considered an approximation of frequency-dependent distributions of spatial features by a GMM classifier, whereby frequency direction-dependent diagonal GMMs can be noted as $\{\mathcal{G}_{f,1_{\boldsymbol{\Theta}}}, \mathcal{G}_{f,2_{\boldsymbol{\Theta}}}, ..., \mathcal{G}_{f,N_{\boldsymbol{\Theta}}}\}$, where N indicates the number of possible source directions in the training data. Now assuming an extracted feature vector, $\mathbf{x}_{k,f} = \{\mathrm{ITD}_{k,f}, \mathrm{ILD}_{k,f}\}$, the log-likelihood for the nth direction with a given feature vector can be calculated as:

$$\mathcal{L}(k, f, n_{\boldsymbol{\Theta}}) = \log p(\mathbf{x}_{k,f} | \mathcal{G}_{f,n_{\boldsymbol{\Theta}}}) \tag{2.23}$$

Then, a robust source location can be given by the spatial index $n_{\boldsymbol{\Theta}}$, which maximises the integration of log-likelihood on all frequencies. The frame-wise basis estimation of can be given by:

$$\hat{\boldsymbol{\Theta}}_{\mathrm{GMM}} = \arg\max_{\boldsymbol{\Theta}} \sum_{f=1}^{F} \mathcal{L}(k, f, n_{\boldsymbol{\Theta}}) \tag{2.24}$$

It should be noted that a significant advantage of the log-likelihoods accumulation is that the uncertainty of binaural cues in a single frequency channel is considered during the estimation [82], and the additional estimation selection becomes unnecessary, since it has already been embedded into the model through the training process. Therefore, the various configurations of multi-conditional training and the accumulation of log-likelihoods provide generalisation ability to the localisation model, which consequently improves the robustness of the performance.

Reverberation Localisation Model

The GMM-based approach has provided a fundamental probabilistic framework, and many supervised methods have been proposed for different applications and extensions. Woodruff and DeLiang applied the GMM model in a flexible azimuth-dependent model, which can adapt to new environments or calibrate to an unseen binaural setup [41]. In their model, they approximated the joint interaural feature observation log-likelihood at frequency channel f by:

$$\mathcal{L}(\mathrm{ITD}_f, \mathrm{ILD}_f|\theta) = \log \sum_r p_f(\mathrm{ITD}_f|r, \mathrm{ITD}_{f,\theta}) p_f(\mathrm{ILD}_f|r, \mathrm{ILD}_{f,\theta}) p_f(r)$$

where r defines the direct-to-residual ratio (DRR), which is treated as a latent variable. In such approximation, the probability space of ITDs would be multi-modal for a high-frequency channel caused by spatial aliasing, so the $P_f(\mathrm{ITD}|r, \mathrm{ITD}_\theta)$ can be obtained by,

$$p_f(\mathrm{ITD}_f|r, \mathrm{ITD}_{f,\theta}) = \sum_{k=1}^{K_f} \rho_{f,k}(r, \mathrm{ITD}_{f,\theta}) \mathcal{N}(\mathrm{ITD}_f|\mu_{f,k}(r, \mathrm{ITD}_{f,\theta}), \sigma_{f,k}(r, \mathrm{ITD}_{f,\theta}))$$

where r defines the direct-to-residual ratio, which is treated as a latent variable. In this approximation, the probability space of ITDs would be multi-modal for a high-frequency channel caused by spatial aliasing, so that the $P_f(\mathrm{ITD}|r, \mathrm{ITD}_\theta)$ can be obtained by:

Two-dimensional Locally-linear Model

The above model has justified the generalisation ability of GMMs to be applied to tackle different acoustical environments and reverberation conditions with the latent variable r. Another approach recently proposed by Deleforge *et al.* proves

that GMMs can be used to describe a locally-linear relationship between the binaural spatial cues and the source location in 3-D space [32, 33]. The authors creatively used GMMs to build probabilistic piecewise affine mapping (PPAM) between the source direction and observed binaural cues. Given that a locally smoothed linear relationship has been proved [33], we can write the following equation:

$$\mathbf{x}_{k_c,f} = \sum_{k_c=1}^{K_c} \mathrm{I}_{\{z=k_c\}}(\boldsymbol{A}_{k_c}\boldsymbol{\Theta} + \boldsymbol{b}_{k_c}) + \boldsymbol{e} \tag{2.25}$$

where k_c defines the centre index of the local linear region and K_c is the maxim number of the segregated linear area in 3-D space. $\mathbf{x}_{k_c,f}$ represents the spatial feature vector in the k_cth region at frequency f. $\mathrm{I}_{\{\cdot\}}$ is an indicator function and $\mathrm{I}_{\{\cdot\}} = 1$ when the equation in $\{\cdot\}$ is true. The linear transformation is defined by the matrix $\boldsymbol{A}_{k_c}$ and vector $\boldsymbol{b}_{k_c}$. The error term $\boldsymbol{e}$ represents the binaural features extraction error and the model reconstruction error. By assuming that the $\boldsymbol{e}$ is independent identically distributed and following Gaussian distribution with 0 mean, a probabilistic representation of (2.25) can be written as:

$$p(\mathbf{x}_{k_c,f}|\boldsymbol{\Theta}, z = k_c) = \mathcal{N}(\mathbf{x}_{k_c,f}; \boldsymbol{A}_{k_c}\boldsymbol{\Theta} + \boldsymbol{b}_{k_c}, \Sigma) \tag{2.26}$$

and the Gaussian mixture prior defined on $\boldsymbol{\Theta}$ is,

$$p(\boldsymbol{\Theta}) = \frac{1}{K_c}\sum_{k_c=1}^{K_c} \mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\Theta}_{k_c}, \Gamma_{k_c}) \tag{2.27}$$

where $\boldsymbol{\Theta}_{k_c} = \{\theta_{k_c}, \phi_{k_c}\}$ is the central location of the segregated locally-linear region. One factor to be noted is that the authors used a forwards conditional density $p(\mathbf{x}_{k_c,f}|\boldsymbol{\Theta})$, instead of the normally used inverse version, $p(\boldsymbol{\Theta}|\mathbf{x}_{k_c,f})$. With this operation, the model learns the mapping from the low-dimension source direction space to the high-dimension spatial feature vector space, which successfully avoids the over-parameterisation and makes it possible to apply the EM algorithm in practice. Hence, the observed log-likelihood of the model can be expressed by:

$$\mathcal{L}(\mathbf{x}_{k_c,f}|\boldsymbol{\Theta}) = \frac{\sum_{k_c=1}^{K_c} \mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\Theta}_{k_c}, \Gamma_{k_c})\mathcal{N}(\mathbf{x}_{k_c,f}; \boldsymbol{A}_{k_c}\boldsymbol{\Theta} + \boldsymbol{b}_{k_c}, \Sigma)}{\sum_{k_c'=1}^{K_c} \mathcal{N}(\boldsymbol{\Theta}; \boldsymbol{\Theta}_{k_c'}, \Gamma_{k_c'})} \tag{2.28}$$

Further, the model parameters $\left\{\{\Gamma_{k_c}, \boldsymbol{\Theta}_{k_c}, \boldsymbol{A}_{k_c}, \boldsymbol{b}_{k_c}\}_{k_c=1}^{K_c}, \Sigma\right\}$ can be obtained by the training process. This model proves the existence of a latent locally-linear low-

dimensional relationship between the source location and interaural spatial data, which also indicates the possibility of localising the azimuth and elevation simultaneously. This result is significant because the existence of this locally-linear relationship will de-risk and encourage more usage of other mapping models inspired by other research areas.

Other Learning-based Approaches

There exist other approaches aside from the widely used table-lookup and GMM-based methods. The most exciting is the neural network and deep learningbased approach, which is promoted by the development of the computation capability of hardware, such as graphics processing units, and the increasing interest in the human auditory cortex working mechanism [87]. The neural network and deep learning methods have achieved great success in many audio-related aspects, such as speech separation and recognition [88, 89]. It is also exciting to apply them to localisation problems.

Both neural networks and deep learning adopt a similar hierarchical structure of the neural network. In brief, the model consists of a multi-layer network, including an input layer, hidden layer (multi-layer) and output layer. Only the nodes of adjacent layers are connected, and the nodes of the same layer and cross-layer are not connected. Each layer can be considered a logistic regression model. As a development from the neural network, the deep learning or deep neural network (DNN) uses a feed-forwards training process, which allows building of a more complex network with more layers and neurons [90].

Besides using the classic binaural features discussed above [79, 87, 91], the DNN model can integrate subject behaviours as input localisation features, which is an important extension in robotic hearing. Ma et al. introduced cross-correlation with head rotation as additional input, which achieves remarkable azimuth localisation accuracy without frontback confusion [50, 51]. They also proposed a machine hearing system with top-down model-based knowledge using DNN [74], which enables the system with attention to separate the target source from the background source. Christopher et al. proposed a binaural localisation system that can drive the robot to select an action to minimise localisation actively [92]. The incorporation of transitional binaural cues and the subject activities not only provides more potential solutions for the binaural sound localisation, but also closely connects the machine hearing to other robotic perceptions in artificial intelligence.

2.6 Summary

Although many approaches of localisation mapping have been proposed, there remain many questions to be answered, such as which features are more important to the localisation, how those features relate to the specific source locations in 3-D spaces, whether feature selection improves the localisation performance, and how we can combine the mapping model and feature evaluation. Motivated by those questions, we investigate and analyse the spatial features for binaural localisation, and embed the feature evaluations in the mapping approaches to obtain a more accurate and robust localisation performance. In addition, we propose new algorithms to map the binaural features to source locations, and explore the possibility of localising a sound source with new features.

This chapter has summarised the background concepts of binaural spatial hearing. We first summarised the binaural localisation system and clarified commonly used coordinate systems for the system setup. We then reviewed the relationship between the geometric position of the sound source and several main binaural localisation cues, and introduced the HRTF to describe the interaction between the sound waves and the filtering effect of the listener's biological structure, which can be exploited for 3-D space localisation. Next, we discussed the influence of the enclosed indoor environment in the localisation, and demonstrated the simulation of geometric reverberation in an enclosed environment to be a linear combination of direct signal source and multiple image sources, which could be actively controlled for testing with complex acoustical configurations. We also reviewed the previous MTF model and the recently proposed CTF model for describing sound propagation in an enclosed environment, and presented the spatial feature extraction procedures with different models. Finally, we briefly examined the existing localisation mapping models. The spherical head model and TDOA estimation provided a solution for azimuth localisation, while the application of HRTF and other spectral cues enabled localisation on the vertical dimension. Further, the supervised-learning-based approaches can take advantage of multi-conditional training, which improves the localisation performance in the presence of noise and reverberation. Lately, the new development of the neural network framework has extended the applicable feature range and increased the possibility of integrating other subject-related perception features. An open research problem is to evaluate the feature importance of spatial cues, and combine the evaluation results in the training of localisation mapping,

which will be investigated and discussed in the remaining chapters of this thesis.

Chapter 3

Spectral Feature Processing and Composite Feature Vector

Overview: *This chapter presents a method of extracting the HRTF spectral cues using cepstral analysis for speech source localisation in the vertical direction. Binaural signals are pre-processed in the cepstral domain, so that the fine spectral structure of speech and the HRTF spectral envelope can be easily separated. We introduce: (i) a truncated cepstral transformation to extract the relevant localisation cues, and (ii) a mechanism to normalise the effects of the time-varying speech spectra. This chapter also presents a method that overcomes the limitation of elevation ambiguity in 3-D space localisation by exploiting the interaural phase and magnitude features present in the HRTF. The method introduces a new feature vector that combines these two sets of features in a non-linear fashion, and we propose a mechanism to extract this feature vector free from distortion by the speech spectra. The performance of the proposed method is evaluated and compared with a correlation-based HRTF database-matching approach and a two-step localisation technique for multiple source positions, HRTFs (individuals) and speech inputs. The results suggest that up to a 20 % improvement in localisation performance can be achieved for moderate signal-to-noise ratios.*

3.1 Introduction

The human auditory system localises a sound source by exploring the ITD/ILD and the monaural spectral cues of binaural signals received at the ears [26], which are believed to be the most important pieces of information used to determine the azimuth location of a sound source [41, 93]. The acoustic transfer function (the HRTF) encompasses these cues and has been widely adopted to design binaural source localisation systems. In general, HRTF-based localisation algorithms identify

a source location by maximising the correlation between the binaural signals [17,18] or the ITD/ILD estimates [93] in the range of possible source locations. Although this approach is known to be robust in the horizontal (azimuth) plane dominated by interaural cues [41], work on elevation localisation has been limited and challenging as a result of the diminishing interaural differences and the dominance of the spectral localisation cues [94, 95].

The perceptual experiments have shown that elevation is perceived as resonances (peaks) and cancellations (notches) of certain frequencies, which are mainly caused by scattering and diffraction of sound waves in the pinna at high frequencies (above 8 kHz) [96]. Numerous methods have been proposed to map the frequencies of these notches to the source location, from parametric models [96, 97] to finite-difference time domain models of the head [98]. However, the accuracy of this mapping is critically affected by perturbations in elevation and the noisy acoustic environment. In contrast, in a practical scenario (e.g., localising speech sources), fully using the richness of the spectral cues may not be possible because of the limited bandwidth of speech sources and additive noise. In this chapter, a pre-processing method is proposed that preserves the elevation-related cues by extracting the HRTF spectral cues through a cepstral transformation of the binaural signals that can minimise the influence of the noise conditions.

The pre-processing method in this chapter is proposed based on two speech characteristics: (i) the majority of the energy is concentrated at low frequencies, where spectral cues are negligible, and (ii) the non-stationary nature of speech. The first characteristic suggests that speech can be fairly well localised in a vertical plane because of the lower variability at frequencies above 3 to 4 kHz, and the reduced likelihood of the speech spectrum contaminating the spectral cues. The second characteristic suggests that a short-time approach (a STFT) is best suited to model the variability of the speech spectrum. In this context, the proposed cepstral processing of binaural signals retains both the fine low-frequency structure of the speech source and the HRTF magnitude information (spectral envelope) at two distinctive parts of the cepstrum. This, together with the logarithmic operation of the cepstral transform that acts as a weighting function at higher frequencies, enables a simpler separation of the two components and the use of the enlarged fluctuations in the HRTF magnitude response for localisation. Finally, the cepstral pre-filtering concept has been shown to be robust to the effects of reverberation in previous studies [63, 64], thereby suggesting that the proposed method will also be suitable for localisation in noisy conditions.

In regard to mapping the extracted features to the source direction, the computation of the correlation between the two received signals and a HRTF dataset is one of the most straightforward localisation mechanisms [17]. However, this does not consider the localisation cue distribution within the HRTF, and is typically inaccurate in noisy environments. Methods that do consider the existence and dispersion of the localisation cues generally adopt a joint [99] or two-step process [97, 100] to first estimate the azimuth direction, and the elevation afterwards. The problems with this method are: (i) multiple potential source locations that exhibit similar ITD and ILD characteristics, and (ii) the lack of a complete set of spectral cues for elevation estimation. Thus, identifying the most relevant localisation information and simultaneous estimation of both azimuth and elevation becomes essential for robust binaural source localisation. Hence, in this chapter, we propose a new feature vector for simultaneous estimation of both azimuth and elevation. First, the localisation features are characterised from the low-frequency IPD, and the mid- to high-frequency ILD/spectral cues. A new feature vector is constructed to include all these key localisation features. Second, we develop the signal processing required to extract these features from the binaural received signals. For example, the IPD is obtained from the cross-spectral density of the two ear signals and a truncated cepstral transform is used to extract the ILD and spectral cues. Finally, the optimal frequency range of the phase and magnitude features is investigated, and the overall localisation performance of the proposed feature vector-based method is compared with the simple correlation and the two-step localisation methods.

This chapter is organised as follows. Section 3.2 describes the binaural signal model in both frequency and time domains. Section 3.3 presents the cepstral domain operation, including the cepstral transformation and cepstral coefficients truncation. Following this, based on cepstral processing, a spectral feature vector is proposed and a corresponding median plane localisation is introduced in Section 3.4. Section 3.5 then explains the 3-D space localisation method based on the Euclidean distance of the composite feature vectors. The localisation performance evaluation is provided in Section 3.6 and a summary of the chapter is given in Section 3.7.

3.2 System Model

The received signal at each of the binaural receivers is modelled on a convolution of the source signal $s(t)$ and the corresponding left and right HRIRs at location

$\mathbf{\Theta}_{\mathrm{VP}} = (\theta, \phi)$, notated as $h_{\mathrm{l}}(t, \theta, \phi)$ and $h_{\mathrm{r}}(t, \theta, \phi)$, which can be represented by:

$$\begin{aligned} x_{\mathrm{l}}(t) &= s(t) * h_{\mathrm{l}}(t, \theta, \phi) + n_{\mathrm{l}}(t) \\ x_{\mathrm{r}}(t) &= s(t) * h_{\mathrm{r}}(t, \theta, \phi) + n_{\mathrm{r}}(t) \end{aligned} \tag{3.1}$$

The received signal can be interpreted in the frequency domain as:

$$\begin{aligned} X_{\mathrm{l},k}(f) &= H_{\mathrm{l}}(f, \theta, \phi) \cdot S_k(f) + N_{\mathrm{l},k}(f) \\ X_{\mathrm{r},k}(f) &= H_{\mathrm{r}}(f, \theta, \phi) \cdot S_k(f) + N_{\mathrm{r},k}(f) \end{aligned} \tag{3.2}$$

where $X_i(f, \theta, \beta)$, $H_i(f, \theta, \beta)$ and $S(f)$ represent the received signal, HRTF and source spectra at a frequency f. $N_{i,k}(f)$ represents the additive noise term and $k = 1 \ldots K$ represents the frame number. The speech signals are separated into K frames, such that the frame length is less than the stationary time duration of the signaltypically 10 to 50 ms for voiced speech [101]. In the following content, the received signal spectra are represented by spectra vectors as:

$$\begin{aligned} \mathbf{x}_{l,k} &= [X_{\mathrm{l},k}(0), ..., X_{\mathrm{l},k}(f_s/2)] \\ \mathbf{x}_{r,k} &= [X_{\mathrm{r},k}(0), ..., X_{\mathrm{r},k}(f_s/2)] \end{aligned} \tag{3.3}$$

where f_s indicates the sampling frequency. Similarly, the HRTF spectra vectors are represented by:

$$\begin{aligned} \mathbf{h}_l &= [H_{\mathrm{l}}(0), ..., H_{\mathrm{l}}(f_s/2)] \\ \mathbf{h}_r &= [H_{\mathrm{r}}(0), ..., H_{\mathrm{r}}(f_s/2)] \end{aligned} \tag{3.4}$$

In this chapter, we consider the source signals to be pure human speech; thus, the majority of the energy in the received signals at the ears is concentrated below 4 kHz (generally the formant frequencies) during voiced speech [101]. However, the pinna-caused localisation cues relevant to the elevation, especially on the median plane, are prevalent at frequencies above 3 kHz [95, 102], and the HRTF features must subsequently be extracted from the high-frequency region of the received signal spectrum.

3.3 Cepstral Domain Processing

3.3.1 Cepstral Transformation

The transformation of a signal into the cepstral domain is an inverse Fourier transform of the absolute magnitude spectrum of that signal. Thus, the cepstral spectra vector is defined as:

$$\mathbf{c}_{X_{\mathrm{i},k}} \triangleq \mathcal{F}^{-1}\left\{\log_{10}|\mathbf{x}_{\mathrm{i},k}|\right\} \tag{3.5}$$

where $\mathrm{i} \in \{\mathrm{l},\mathrm{r}\}$ indicates the left and right ear label and $\mathcal{F}^{-1}$ represents the inverse discrete Fourier transform. The cepstral transformation of (3.2) can be expressed approximately as a sum of the cepstral coefficients of the HRTF, speech and noise components, respectively, due to the logarithmic operation [63, 64] of the cepstral transform, and is defined as:

$$\mathbf{c}_{X_{\mathrm{i},k}} \triangleq \mathbf{c}_{H_{\mathrm{i}}} + \mathbf{c}_{S_k} + \mathbf{c}_{\widetilde{N}_{\mathrm{i},k}} \tag{3.6}$$

Thus, the magnitude response of the HRTF could be reconstructed by extracting the cepstral coefficients corresponding to $\mathbf{c}_{H_{\mathrm{i}}}$, and $\mathbf{c}_{\widetilde{N}_{\mathrm{i},k}}$ is the cepstral transformation of the term $(1 + N_{\mathrm{i},k}(f)/(H_{\mathrm{i}}(f) \cdot S_k(f)))$. However, in practice, the influence of the speech spectrum will distort the reconstructed HRTF, and requires a statistical normalisation of the effects of speech. Hence, to reduce this distortion, we introduce a prefilter $G(f)$, as described in Section 3.4.2. To simplify the derivation, we ignore the effect of noise ($\mathbf{c}_{N_{i,k}} \to \mathbf{0}$), but include and evaluate its effect on the localisation performance in Section 3.4 . Thus, the pre-filtered received signals can be simplified as:

$$\widetilde{X}_{i,k}(f) = G(f) \cdot H_i(f) \cdot S_k(f). \tag{3.7}$$

The corresponding cepstral domain representation becomes:

$$\mathbf{c}_{\widetilde{X}_{\mathrm{i},k}} = \mathbf{c}_G + \mathbf{c}_{H_{\mathrm{i}}} + \mathbf{c}_{S_k}, \tag{3.8}$$

where $\mathbf{g} = [\ G(0),\ \cdots,\ G(f_{\max})\ ]$ is the spectrum of the prefilter $G(f)$, and $\mathbf{c}_G \triangleq \mathcal{F}^{-1}\left\{\log_{10}|\mathbf{g}|\right\}$.

3.3.2 Cepstral Coefficients Truncation

The lower-order cepstral coefficients in (3.6) typically model the envelope of the received signal spectrum, while the higher-order coefficients model its rapid spectral fluctuations. In the case of speech sources, the higher-order coefficients are predominantly speech information [101] corresponding to the pitch and formant structure of a particular speech frame. Thus, the spectral envelope of the HRTF could be extracted by appropriately truncating the cepstral coefficients in (3.8).

Thus, we define a truncation operation $\mathcal{T}\{\cdot\}$ that retains sufficient HRTF magnitude information for source localisation. The truncated cepstral coefficients now become:

$$\widetilde{\mathbf{c}}_{\widetilde{X}_{\mathrm{i},k}} \triangleq \mathcal{T}\left\{\mathbf{c}_{\widetilde{X}_{\mathrm{i},k}}\right\} = \mathcal{T}\left\{\mathbf{c}_G\right\} + \mathcal{T}\left\{\mathbf{c}_{H_i}\right\} + \mathcal{T}\left\{\mathbf{c}_{S_k}\right\}, \tag{3.9}$$

where the truncation order is determined by the number of cepstral coefficients required to model the magnitude response of $h_i(t)$.

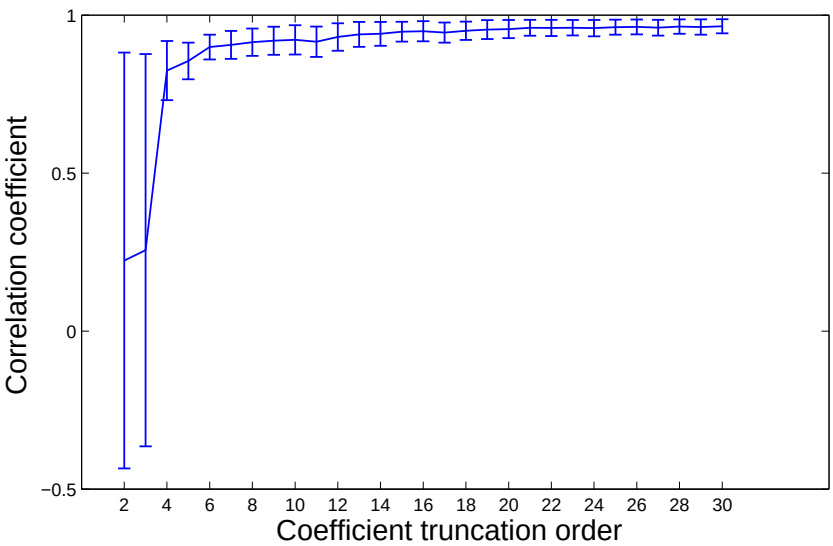


Figure 3.1: Correlation between the truncated approximation and the actual HRTF, for source locations in the median plane.

The effects of truncating the HRTF magnitude response in the cepstral domain are presented in Figure 3.1. The mean correlation and standard deviation between the actual HRTF and the truncated approximation of the HRTF in the 3.5 to 7.5 kHz frequency bandwidth are summarised for the KEMAR's median plane source locations. As expected, we observe that increasing the cepstral truncation order improves the correlation because of the inclusion of the rapid fluctuations in the HRTFs. However, our objective is to extract a smooth form of the spectral cues

in the magnitude response of the HRTFthat is, to model the general shape of the HRTF with a sufficiently small number of coefficients. We observe that the mean correlation stabilises beyond approximately the first 25 cepstral coefficients as the correlation is very close to 1 and barely unchanged. Therefore, we use a cepstral truncation order of 25 to extract the HRTF spectral envelopes of the binaural signals.

3.4 Median Plane Localisation Using Spectral Feature

We first investigate the performance of vertical plane localisation based on spectral cues only. Given that the spectral cues are significant to estimate the elevation [96, 97], in this section, the sound sources are placed on the median plane and azimuth $\alpha = 0$, where the interaural differences are minimised so that the spectral cues are entirely caused by the reflection and diffraction of pinna and head. The extraction of spectral features and a correlation-based localisation approach are included in the following content.

3.4.1 Spectral Feature Vector

In this section, the correlation between the extracted HRTF spectral envelope and the HRTFs in a pre-measured database is adopted as a metric for spectral feature-based median plane localisation. However, since the localisation cues that differentiate the source locations in the median plane are both subtle and predominantly located at higher frequencies, we first combine the extracted HRTF spectral envelopes obtained from the received binaural signals in (3.16) for the relevant frequency range, $f \in \{f_{\min}, \ \ldots, \ f_{\max}\}$. The spectral envelopes of the median plane HRTFs in the database can be expressed in a similar fashion, by applying the cepstral truncation operation $\mathcal{T}$. Thus, the reference feature vector at elevation ϕ becomes:

$$\widetilde{\mathbf{h}}(\phi) = [\widetilde{\mathbf{h}}_{\mathrm{l}}(\phi), \widetilde{\mathbf{h}}_{\mathrm{r}}(\phi)] \tag{3.10}$$

where $\widetilde{\mathbf{h}}_{\mathrm{i}}(\phi)$ defines the cepstral processed HRTF data as:

$$\widetilde{\mathbf{h}}_{\mathrm{i}}(\phi) \triangleq 10^{\mathcal{F}\{\mathcal{T}\{\mathbf{c}_{H_{\mathrm{i}}}(\phi)\}\}} = \left[\ \widetilde{H}_{\mathrm{i}}(f_{\min}, \phi), \ \cdots, \ \widetilde{H}_{\mathrm{i}}(f_{\max}, \phi) \ \right]. \tag{3.11}$$

Then, a reference vector database can be obtained by collecting $\widetilde{\mathbf{h}}(\phi)$ of all possible ϕ on the median plane.

3.4.2 Spectral Feature Extraction and Speech Normalisation

By including the effects of the cepstral truncation operation described in the previous section, (3.5) can be expressed as:

$$\widetilde{\mathbf{c}}_{\widetilde{X}_{\mathrm{i},k}} = \mathbf{c}_G + \widehat{\mathbf{c}}_{H_{\mathrm{i}}} + \mathcal{T}\left\{\mathbf{c}_{S_k}\right\}, \tag{3.12}$$

where $\widehat{\mathbf{c}}_{H_{\mathrm{i}}} \triangleq \mathcal{T}\left\{\mathbf{c}_{H_{\mathrm{i}}}\right\}$ and $\mathbf{c}_G \triangleq \mathcal{T}\left\{\mathbf{c}_G\right\}$. $\widehat{\mathbf{c}}_{H_{\mathrm{i}}}$ characterises the cepstral approximation of the HRTF magnitude responsethat is, the spectral envelopebut source localisation requires the estimation of $\widehat{\mathbf{c}}_{H_{\mathrm{i}}}$ in the presence of time-varying speech. As stated previously, we introduce the prefilter $G(f)$ to normalise the effects of the speech component $\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}$, such that:

$$\mathbf{c}_G + E\left[\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\right] = \mathbf{c}_0, \tag{3.13}$$

where $\mathbf{c}_0 = [\ c_0,\ 0,\ \cdots,\ 0\ ]$ is a constant vector and $E\left[\cdot\right]$ is the expectation operator. c_0 is an arbitrary constant, selected such that $G(f)$ normalises the distribution of speech energy across frequency, resulting in only the zero-th order cepstral coefficient. Thus, the prefilter can be expressed as:

$$\mathbf{g} = 10^{\mathcal{F}\{\mathbf{c}_0 - \mathbf{c}_S\}}, \tag{3.14}$$

where $\mathbf{c}_S \triangleq E\left[\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\right]$ is obtained empirically from the analysis of speech data obtained from multiple speakers. The design of the prefilter and the behaviour of the speech spectrum can now be exploited to extract the truncated cepstral HRTF coefficients $\widehat{\mathbf{c}}_{H_i}$, as shown below. For a particular speaker, the expectation of (3.10) can be expressed as:

$$\begin{aligned}\widehat{\mathbf{c}}_{\widetilde{X}_{\mathrm{i}}} \triangleq \frac{1}{K}\sum_{k=1}^{K}\widetilde{\mathbf{c}}_{\widetilde{X}_{\mathrm{i},k}} &= \mathbf{c}_G + \widehat{\mathbf{c}}_{H_{\mathrm{i}}} + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\} \\ &= \mathbf{c}_0 + \widehat{\mathbf{c}}_{H_{\mathrm{i}}} - \mathbf{c}_S + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\},\end{aligned} \tag{3.15}$$

where $\mathbf{c}_G = \mathbf{c}_0 - \mathbf{c}_S$ from (3.11). We exploit the property of speech, where, for sufficiently large K, the spectrum approaches a general distribution [103, 104], such that:

$$\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\} \rightarrow E\left[\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\right] = \mathbf{c}_S. \tag{3.16}$$

Thus, (3.13) becomes

$$\widehat{\mathbf{c}}_{\widetilde{X}_{\mathrm{i}}} \approx \mathbf{c}_0 + \widehat{\mathbf{c}}_{H_{\mathrm{i}}}, \tag{3.17}$$

where $\mathbf{c}_0$ is zero for all except the first elementthat is, a cepstral representation of a uniform spectrum. Thus, the estimated HRTF spectral envelope can be extracted by applying an inverse cepstral transformation to (3.14), and is given by:

$$\widehat{\mathbf{h}}_i = 10^{\mathcal{F}\left\{\widehat{\mathbf{c}}_{\widetilde{X}_{\mathrm{i}}} - \mathbf{c}_0\right\}}. \tag{3.18}$$

Then, following a similar approach as in (3.11), the extracted feature vector can be constructed by concatenate left and right features as:

$$\widehat{\mathbf{h}} = [\widehat{\mathbf{h}}_l, \widehat{\mathbf{h}}_r] \tag{3.19}$$

By comparing the extracted vector and reference vector database obtained from Section 3.4.2, the source position can be estimated.

3.4.3 Correlation-based Median Plane Localisation

Now, the source position on the median plane can be determined by the maximum cross-correlation between the extracted feature vector and the corresponding reference vector with directional parameters θ and ϕ:

$$< \widehat{\theta}, \widehat{\phi} >= \arg\max_{\theta,\phi} \left\{\widehat{\mathbf{h}} \oplus \widetilde{\mathbf{h}}(\theta, \phi) \geq \rho\right\}. \tag{3.20}$$

where $\oplus$ denotes a cross-correlation operation. Hence, for a single active sound source, the estimated source location in the median plane can be determined by evaluating (3.20) for all possible $< \theta, \phi >$ Considering that the possible noise influence and the case of no active source, an appropriate threshold is chosen, i.e., $\rho = \max\{0.95 \cdot \max\{\widehat{\mathbf{h}} \oplus \widetilde{\mathbf{h}}(\theta, \phi)\}, 0.5\}$, to identify the peak of the spectrum and the relevant source location. Section.3.7 demonstrates the detailed median plane localisation performance of the proposed method.

3.5 3-D Space Localisation Using Interaural Features

This section investigates the localisation in 3-D space based on interaural features. In this section, the extraction of interaural features is introduced and a composite feature vector is constructed based on a generic feature spectrum selection. Further, a Euclidean distancebased vector match-up method is proposed for estimating the source location.

3.5.1 Composite Interaural Feature Vector

In this section, the interaural feature vector is constructed by two interaural featuresinteraural phase and interaural magnitudeand a Euclidean vector distance-based lookup localisation mechanism is proposed for estimating source position in 3-D space.

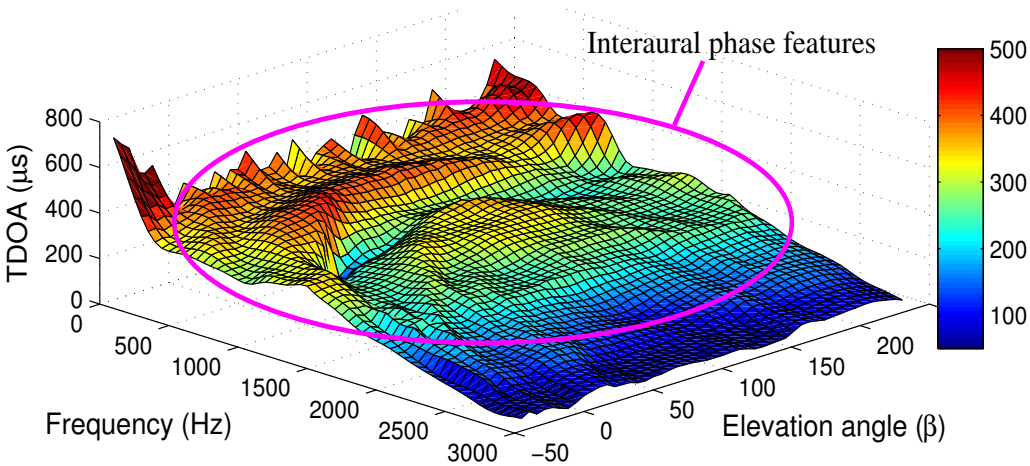


Figure 3.2: Time difference of arrival of the HRTFs with respect to the elevation φ in the sagittal plane $\vartheta = 30°$ of CIPIC 'subject_003'.

Interaural Phase Vector

The ITD that arises as a natural consequence of the spatial separation of the ears is commonly used to estimate the azimuth of an incoming sound source. Localisation techniques such as the Generalised Cross-correlation Phase Transform (GCC-PHAT) [23] method exploit this fact to estimate the broadband TDOA between the signals and estimate the source azimuth. However, this loses much of the subtle differences in phase (which are both frequency and elevation dependent, as shown in Figure

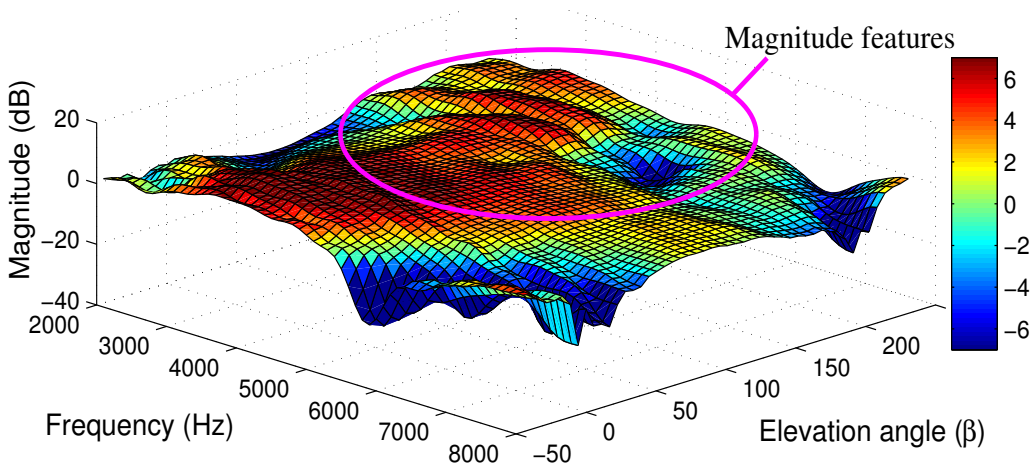


Figure 3.3: Left ear HRTF magnitude response indicating monaural magnitude features in the sagittal plane $\vartheta = 30°$ of CIPIC 'subject_003'.

3.2) introduced by the head, torso and pinna through the effects of scattering and reflections. Thus, incorporating the change in IPD with frequency, for both azimuth and elevation localisation, could lead to greater localisation accuracy. We propose that this localisation information be extracted as an interaural phase feature, and expressed as a normalised cross-PSD given by

$$\mathcal{V}^p(f,\theta,\phi) = \frac{H_r(f,\theta,\phi)H_l(f,\theta,\phi)^*}{|H_r(f,\theta,\phi)||H_l(f,\theta,\phi)^*|}, \tag{3.21}$$

where * denotes the conjugation operation. Thus, the feature vector of interaural phase information for a source located in the direction (θ,ϕ) can be expressed as :

$$\mathbf{v}^p(\theta,\phi) = \left[\ \mathcal{V}^p(f^p_{\min},\theta,\phi),\ \ \cdots,\ \ \mathcal{V}^p(f^p_{\max},\theta,\phi)\ \right], \tag{3.22}$$

where f^p indicates the selected phase spectrum frequency, and $f^p_{\min}$ and $f^p_{\max}$ represent the minimum and maximum selected phase spectrum frequency, respectively.

Interaural Magnitude Vector

The magnitude features of the HRTFs primarily comprise the ILD and monaural spectral cues (shown in Figure 4.8), and represent a location-dependent modulation of the received signal amplitude. Thus, together, the ILD and the left and right ear spectral cues (which are commonly used in elevation localisation) are interaural magnitude features that can be exploited for 3-D source localisation.

To extract these features, we adopt a modified cepstral processing method, as introduced in Section 3.4, which was previously proposed to extract the HRTF magnitude response for binaural localisation in the median plane. Here, the HRTFs

are first transformed into the cepstral domain, truncated to a finite order to remove any rapid fluctuations in the frequency domain, and finally transformed back into the frequency domain as a smoothed HRTF magnitude response. Thus, we express the magnitude feature vector of the signal received at a particular ear as:

$$\mathbf{v}_i^m(\theta,\phi) = \mathcal{C}^{-1}\Big\{\mathcal{T}\big[\ \mathcal{C}\left\{\mathbf{h}_i(\theta,\phi)\right\}\ \big]\Big\}\Big| \quad f \in [f_{\min}^m, f_{\max}^m], \tag{3.23}$$

where $\mathcal{C}$ and $\mathcal{C}^{-1}$ represent the cepstral and inverse cepstral transforms, respectively. $\mathcal{T}$ describes the cepstral truncation operation in Section 3.3 and $\mathbf{h}_{\mathrm{i}}(\theta,\phi)$ indicates the modified HRTF spectra vector as:

$$\mathbf{h}_{\mathrm{i}}(\theta,\phi) = \Big[\ H_{\mathrm{i}}(f_{\min}^m,\theta,\phi),\quad \cdots\ ,\quad H_{\mathrm{i}}(f_{\max}^m,\theta,\phi)\ \Big] \tag{3.24}$$

for a sampling rate of F_s. $f_{\min}^m$ and $f_{\max}^m$ demarcate the selected spectrum frequencies whose magnitude features are of interest to us.

Composite Feature Vector

With phase vector $\mathbf{v}^p(\Theta)$ and magnitude vector $\mathbf{v}_i^m(\Theta)$, we define a new composite feature vector as a non-linear combination of the interaural phase and interaural magnitude features described in the previous subsections. Mathematically, this feature vector can be expressed as :

$$\mathbf{v}(\theta,\phi) \triangleq \mathbf{v}^p(\theta,\phi) \odot \big\{\mathbf{v}_r^m(\theta,\phi) \oslash \mathbf{v}_l^m(\theta,\phi)\big\}, \tag{3.25}$$

where $\odot$ and $\oslash$ represent the element-wise multiplication and division of vectors, respectively. The advantage of this non-linear combination is to enlarge the differences of the feature vectors between closely spaced source positions, thus rendering it particularly suitable for 3-D localisation. It should also be noted that $\mathbf{v}^p(\theta,\phi)$ and $\mathbf{v}_i^m(\theta,\phi)$ must be of similar length; thus, one feature vector may require finer sampling in the frequency domain (in this work, a simple interpolation of $\mathbf{v}^p(\theta,\phi)$ is adopted for this purpose).

3.5.2 Interaural Feature Extraction

The received signals at the two ears in (3.2), although containing directional information, are both time variant because of the effects of speech, and are corrupted by noise. Hence, these signals must be processed further to extract the features

discussed in the previous section.

Interaural Phase Extraction

To extract the interaural phase feature vector, we define the estimated interaural phase as the mean normalised cross-power spectral densities of the two received signals in (3.2), given by:

$$\hat{\mathcal{V}}^p(f) \triangleq E\left\{\frac{X_{r,k}(f)X_{l,k}(f)^*}{|X_{r,k}(f)||X_{l,k}(f)^*|}\right\}, \tag{3.26}$$

where $E\{\cdot\}$ represents the expectation operator over time and the estimated interaural phase feature vector can be expressed as:

$$\hat{\mathbf{v}}^p = \left[\ \hat{\mathcal{V}}^p(f^p_{\min}),\ \ \cdots,\ \ \hat{\mathcal{V}}^p(f^p_{\max})\ \right]. \tag{3.27}$$

in the selected frequencies of $f \in [f^p_{\min}, f^p_{\max}]$.

Interaural Magnitude Extraction

We extract the interaural magnitude features from the received signals using a modified version of the cepstral pre-processing method introduced in Section 3.3. Through exploiting the properties of the cepstral domain signal processing, the time-averaged and cepstral-truncated signal becomes

$$\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\ \mathcal{C}\left\{\mathbf{x}_{\mathrm{i},k}\right\}\right] = \mathcal{T}\left[\ \mathcal{C}\left\{\mathbf{h}_{\mathrm{i}}(\alpha,\beta)\right\}\right] + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\ \mathcal{C}\left\{\mathbf{s}_k\right\}\right] + \mathbf{n}_{\mathrm{i}}, \tag{3.28}$$

where $\mathbf{n}_i$ represents the time averaged noise cepstrum, $\mathbf{x}_{i,k} = [X_{i,k}(f^p_{\min}), \cdots, X_{i,k}(f^p_{\max})]$ and $\mathbf{s}_k = [S_k(f^p_{\min}), \cdots, S_k(f^p_{\max})]$. Assuming sufficiently long observations and same statistical properties for $\mathbf{n}_l$ and $\mathbf{n}_r$ (i.e., $\mathbf{n}_r - \mathbf{n}_l \to 0$), the inverse cepstral transform of the difference in (3.23) between the two ears represents the magnitude feature vector. That is, we extract the feature

$$\hat{\mathbf{v}}^m \ \triangleq\ \mathcal{C}^{-1}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\ \mathcal{C}\left\{\mathbf{x}_{r,k}\right\}\right] - \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\ \mathcal{C}\left\{\mathbf{x}_{l,k}\right\}\right]\right\} \tag{3.29}$$

which is roughly equivalent to $\mathbf{v}^m_r(\alpha,\beta) \oslash \mathbf{v}^m_l(\alpha,\beta)$ in (3.25).

Based on (3.27) and (3.29), the estimated composite feature vector can be constructed by:

$$\hat{\mathbf{v}} \triangleq \hat{\mathbf{v}}^p \odot \hat{\mathbf{v}}^m. \tag{3.30}$$

Now, by pairing $\hat{\mathbf{v}}$ and reference feature vectors $\mathbf{v}(\Theta)$ from the database with all Θ with the closest vector distance, the source position Θ can be determined. Ideally, the estimated composite feature vector in (3.30) is coincident only with the composite feature vector of the true source location given by (3.25). Thus, the magnitude of the Euclidean distance between these quantities can be used to estimate the source location. We express this as a localisation error metric $\forall\Theta$, given by:

$$\mathcal{E}(\theta,\phi) = 20\log_{10}\left\|\left\{\frac{\hat{\mathbf{v}}}{\|\hat{\mathbf{v}}\|} - \frac{\mathbf{v}(\theta,\phi)}{\|\mathbf{v}(\theta,\phi)\|}\right\}\right\|, \tag{3.31}$$

the minimum of which yields the estimated source location in 3-D as

$$(\hat{\theta},\hat{\phi}) = \arg\min_{\theta,\phi}\ \{\mathcal{E}\big(\theta,\phi\big)\} \tag{3.32}$$

Therefore, the source position in 3-D space can be determined by (3.30). To obtain a more accurate and robust localisation result, a generic selection of feature spectrum frequencies is discussed in the following section.

3.5.3 Interaural Feature Selection and Composite Feature Vector

A significant issue when constructing the composite feature vector is selection of the appropriate frequency bands for feature extraction. This requires that only the most relevant localisation information is included to reduce the complexity and minimise the noise influence. Traditionally, the phase features are extracted from a relatively low-frequency range (i.e., 0 to 1.5 kHz) to calculate the ITD/IPD [78]. However, in the proposed feature vector, the phase features are used not only to estimate the azimuth, but also to localise the elevation; hence, this requires much higher frequencies to be included [96]. In the simulations, we modify the upper frequency limit of the phase feature, $f^p_{\max}$, to determine the optimal frequency range. Here, the lower frequency limit, $f^p_{\min}$, is fixed at 0, since the low-frequency phase information is essential for azimuth estimation. As for the magnitude features, it is necessary to determine the frequency band that includes the key features, subject to minimal distortion by the speech spectrum, within the speech bandwidth. Given that the

formants and a majority of the speech energy exists below 3 kHz [105], the lower frequency limit, $f^m_{\min}$, is fixed at 3 kHz and the upper limit, $f^m_{\max}$, is varied during the simulation.

3.6 Simulation

3.6.1 Median Plane Localisation

Simulation Configuration

We evaluate the performance of the localisation technique in Section 3.5.1 for median plane localisation through simulations, using MIT's HRTF measurement database of the KEMAR dummy head [55]. The simulated clean speech signals are obtained from a corpus of 34 speakers, each with 600 utterances sampled at 16 kHz, consisting of a sequence of words in a sentence. The utterances have length around 1 to 2 s. The speech data are obtained from the sample recordings used in the PASCAL CHiME Speech Separation and Recognition Challenge [106], and the binaural signals are produced by convoluting the KEMAR HRIRs with these speech signals. We apply a simple signal activity detector based on received signal level to identify the voiced signal frames, and perform the Fourier and cepstral transform operations using a 20 ms Hamming window at 10 ms intervalsthat is, a window length of 320 samples corresponding to the sampling rate of 16 kHz. The prefilter $\mathbf{g}$ is computed from the average speech cepstrum as per (3.17), and is used as a common prefilter for all 34 speakers.

The localisation performance is compared with the state-of-the-art convolution-based binaural localisation scheme described in [19], which is a superior, more noise-robust variant of the source cancellation algorithm [17,18] and the classical matched filtering approach [20]. We evaluate the performance in the 3.5 to 7.5 kHz audio bandwidth, where spectral cues are known to dominate [94], at the indicated SNRs. The effect of reverberation is not explicitly considered, but the cepstral methods are known to be robust to its effects [63, 64] because of the truncation operation's removal of the rapid spectral fluctuations. .

Figure 3.4 illustrates the comparison of the median plane localisation spectra for the convolution-based and proposed methods for a single trial at 40 dB SNR. The vertical dashed line at 10° indicates the actual source location in the median plane. Although the peak correlation for both techniques corresponds to the actual source

location, it can be observed that the convolution-based method provides a flatter source localisation spectrum, in contrast to the distinctive peak of the proposed method. Naturally, the flat spectrum increases the uncertainty of the estimated source locationthat is, the confidence interval of the estimate. For simplicity, we use a single standard deviation of the distribution of the detected source locations as a metric to quantify this uncertainty, and it is denoted by the error bars in the subsequent figures.

Figure 3.5 illustrates the average source localisation performance for a source located in the median plane at 10° intervals between -40° and 220°, in the 3.5 to 7.5 kHz audio bandwidth. 0°, 90° and 180° indicate the source locations directly in front, above and behind the KEMAR in the median plane. The results are the averages of multiple trials using different speech segments of the speakers in the CHiME speech corpus for each source location. The dashed line indicates the actual median plane source locations in the different experiment scenarios, and is the performance benchmark that the localisation algorithms should ideally follow.

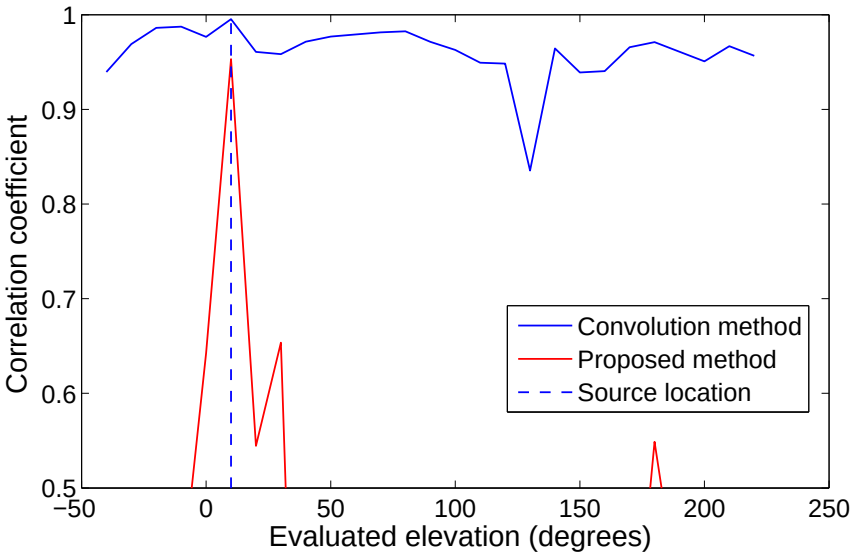
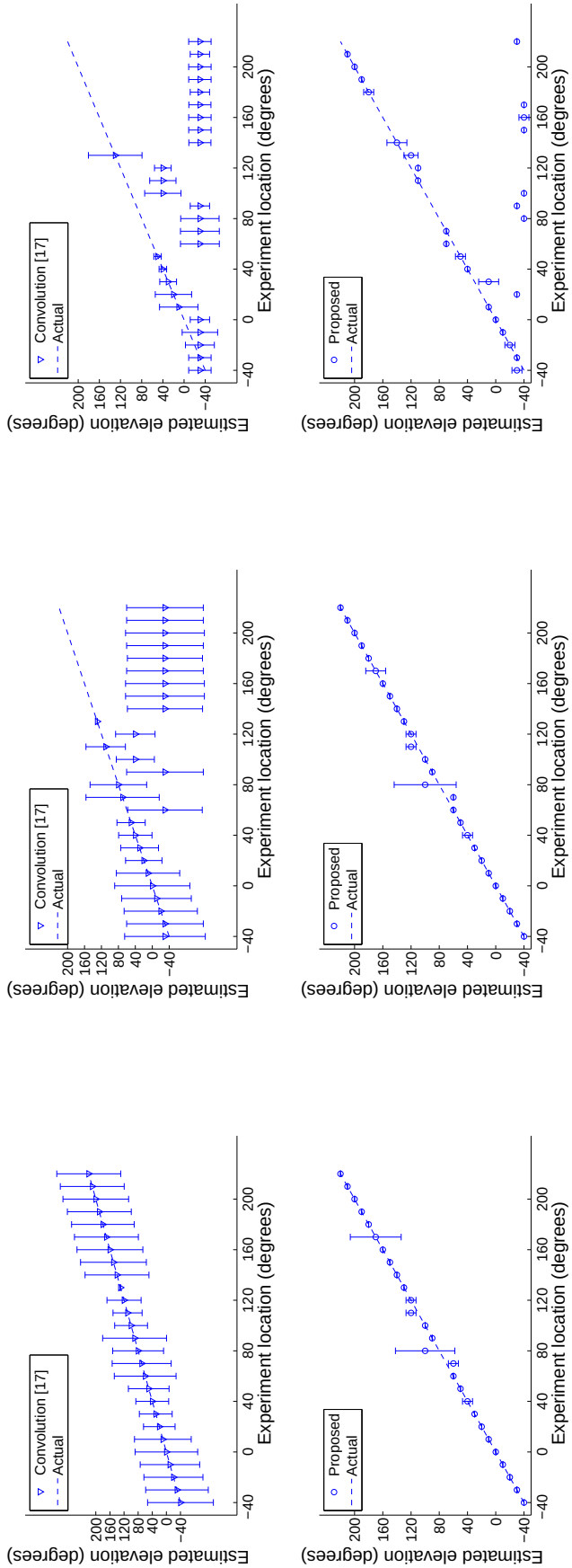


Figure 3.4: Source localization spectra of the proposed and convolution based methods for a source at 10° in the median plane.

The localisation performance results in Figure 3.5(a) show a similar localisation accuracy for both techniques, but a much greater uncertainty of the estimated location for the convolution-based method at 40 dB SNR. In general, decreasing SNR in Figure 3.5(b) and (c) results in a degradation in the performance, but the proposed method is shown to be superior and more robust to the effects of noise. Crucially, the uncertainty of the source location estimate of the proposed method

is not visibly affected. However, at 20 dB SNR in Figure 3.5(c), the performance of the proposed method diverges from the ideal benchmark in the region directly above and behind the KEMAR head. This is consistent with the known localisation capability of humans [26], and is mainly because of the lack of rich spectral cues in these regions. The greater uncertainty of the convolution-based method is primarily due to it favouring the higher energy region of the signal spectrum, and is a deficiency that is exacerbated with decreasing SNR. Overall, the greater accuracy of the proposed method at low SNR indicates a more efficient exploitation of the spectral localisation cues for binaural localisation in the median plane.

This proposed median plane localisation method has also been tested with various configurations in [107,108], which demonstrates a robust performance, especially for localisation speech signals. [107] also reports that the proposed method has more reasonable localisation results above the horizontal plane. Furthermore, the idea of removing slow changing speech component is also utilised in [108] to develop a advanced localisation method which is more robust to the source signal.



(a) SNR = 40 dB (b) SNR = 30 dB (c) SNR = 20 dB

Figure 3.5: Average single source localization performance in the 3.5–7.5 kHz audio bandwidth at (a) 40 dB, (b) 30 dB and (c) 20 dB SNR. The figures indicate the source localization spectra $P(\Theta)$ of different trials, where the source is located at the actual elevations between -40° and 220° in the median plane.

3.6.2 3-D Space Localisation with Generic Selective Feature

Simulation Configuration

The proposed binaural localisation technique in 3-D space is evaluated through simulations using the CIPIC HRTF database [54]. This approach uses the HRTFs of 45 subjects, each with 950 different locations (the azimuth angle varies from $-45°$ to $45°$ with a $5°$ interval, and the elevation varies from $-45°$ to $230.625°$ degrees with a $5.625°$ interval). The clean speech signals are obtained from the recordings used in the PASCAL CHiME Speech Separation and Recognition Challenges [106]. The database contains 34 speakers, each with 500 utterance segments sampled at 16 kHz. The received binaural signals are simulated by convoluting the HRTF with speech samples, and considering the variability of the speech spectrum, a short-time approach (a STFT) is applied. A voice activity detector identifies the voiced speech frames, while both the Fourier and cepstral transforms employ a 20 ms Hamming window with 10 ms overlap. For the 16 kHz sample rate, this implies that a window length of 320 is used throughout the simulation. The optimum feature frequency ranges are selected by comparing the localisation performances for different range combinations.

The performance of the proposed localisation technique is compared with the two-step [97, 100] and simple correlation approaches [18]. The two-step method uses ITD/ILD information to narrow down the potential source locations to a specific cone of confusion, prior to estimating the elevation using the binaural magnitude spectra. In contrast, the simple correlation approach simply computes the correlation between the received binaural signals and the complete HRTF dataset. In both cases, the best match to the HRTF dataset identifies the estimated source location. We compare the localisation performance in terms of the localisation error probabilitythat is, the likelihood of a localisation error across all 950 source locations over multiple trials (lower values imply better performance). The results indicate the mean localisation performance for all 45 subjects in the HRTF dataset, while the error bars indicate the standard deviation.

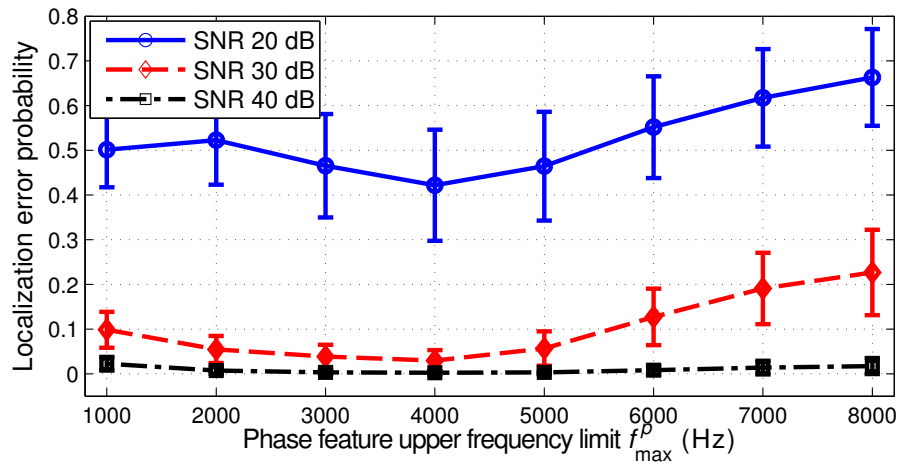
Overall 3-D Localization Performance

Aiming at selecting the optimised frequency range, a preliminary simulation is conducted. The simulation results of the localisation error rate (for all 45 subjects in the HRTF database) for feature range selection are illustrated in Figure 3.6 at

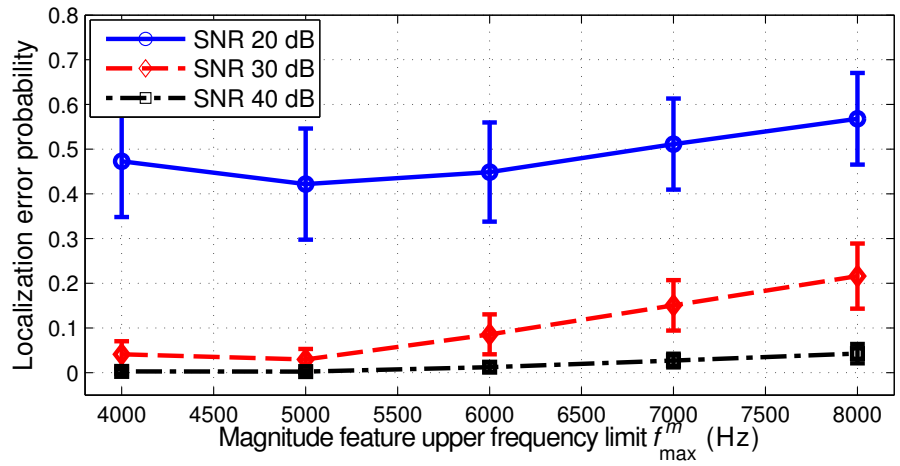
different SNRs. The results indicate that the optimum frequency bands for phase and magnitude feature extraction are $[0, 4]$ kHz and $[3, 5]$ kHza region where the localisation error probability is a minimum for all SNR. Figures 3.6(a) and (b) illustrate the effect of the upper frequency limits for the phase and magnitude features, respectively. The joint effect of both parameters is illustrated in Figure 3.6(c). They show that the interaural phase features, up to a relatively high frequency range, do actively contribute to elevation estimation, as expected intuitively from Figure 3.2. As for the magnitude feature frequency range, when the upper frequency limit is greater than 6 kHz, the received signals are more strongly affected by noise (because of the rapid decay of speech energy with frequency), which results in a degradation of the localisation performance.

Figure 3.7(a) illustrates an example spectra of the localisation error metric (described in Section 3.4) for a source located at $\Theta \equiv (20°, 16.875°)$ at 30 dB SNR. As shown, the estimated source location corresponds to the location that minimises the error metric $\mathcal{E}(\Theta)$. Further, the distinctive spike suggests that elevation ambiguity is also minimal. To evaluate the 3-D localisation performance, we use these estimates for all possible source locations and subjects, and analyse the overall localisation error probability.

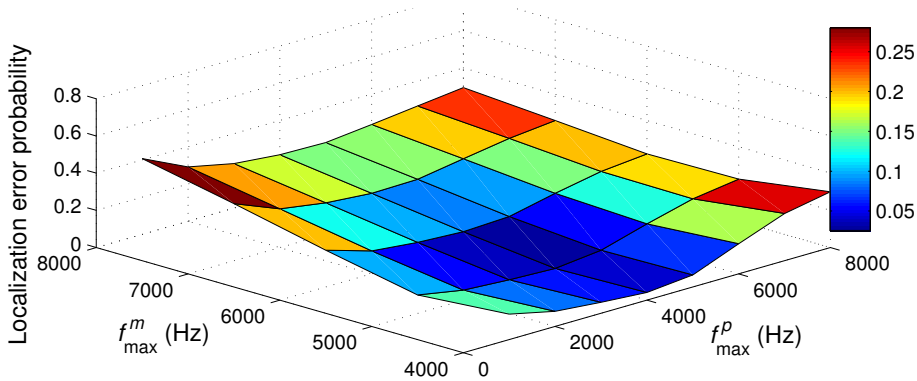
Figure 3.7(b) illustrates the overall localisation error probability and compares the performance of the proposed method and the two comparison methods. The presented data illustrate the mean performance for all 45 CIPIC subjects, where the error bar indicates the standard deviation. In general, as expected, the localisation error probability decreases with increasing SNRs. The proposed method has the best localisation performance with the lowest probability of localisation errors under all noise conditions. This suggests that the proposed method is more robust to the effects of noise. Further, the error bars of the proposed method are generally smaller than the other two methods, especially at higher SNR, which implies that the proposed method has more consistent performance for different listeners' HRTFs. In addition, comparing the proposed method and the two-step method, the crucial difference lies in the use of the phase features for elevation estimation. Here, the two-step method uses the interaural phase feature only for azimuth localisation, while the proposed method does so for joint azimuth and elevation estimation. Thus, the superior localisation performance of the proposed method proves that the phase features do indeed contain elevation localisation information, and should not be neglected in 3-D binaural localisation of speech sources.



(a) Interaural phase feature frequency range selection

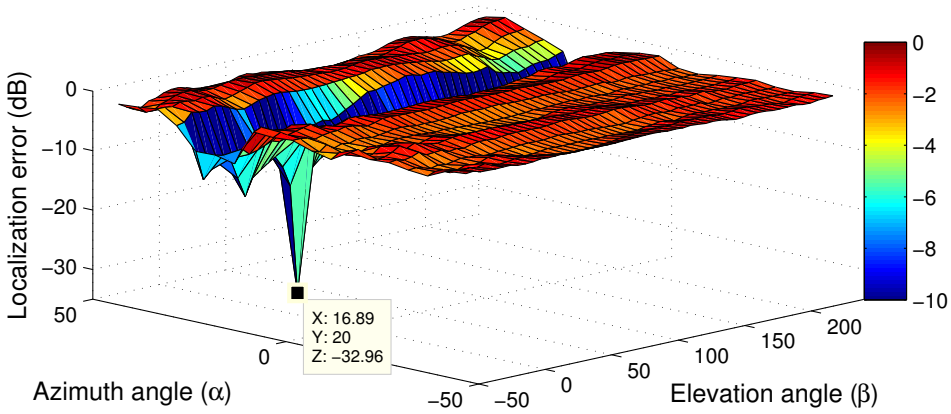


(b) Interaural magnitude feature frequency range selection

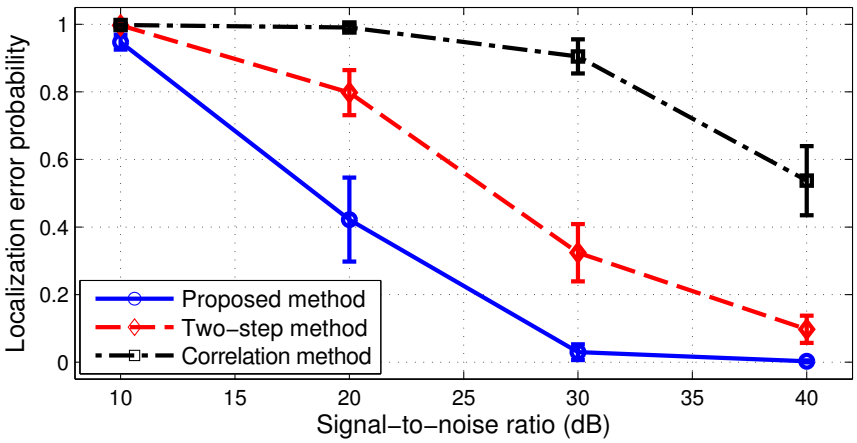


(c) Joint phase and magnitude feature frequency range selection

Figure 3.6: 3-D localization error probability for different frequency ranges of phase and magnitude features with respect to SNR. (a) Localization error probability with respect to the phase feature frequency range 0–$f^p_{\max}$ kHz and a magnitude feature range of 3–5 kHz. (b) Localization error probability with respect to the magnitude feature frequency range 3–$f^m_{\max}$ kHz and the phase feature range of [0, 4] kHz. (c) Localization error probability with respect to upper frequency limits of phase and magnitude features at 30 dB SNR.



(a) Localization error metric $\mathcal{E}$ for a source located at $\Theta \equiv (20°, 16.875°)$



(b) Overall 3-D localization performance comparison

Figure 3.7: (a) Localization error metric of the proposed method for a source located at α=20° and β=16.875°. (b) Comparison of the overall 3-D localization error probability of the proposed method, two-step method and correlation method for SNRs from 10–40 dB.

3.7 Summary and Contribution

In this chapter, we have developed a novel monaural spectral feature extraction method for the median plane localisation. The simulation results indicated that the spectral feature envelope can be extracted by cepstral processing, which is critical for successful source localisation in a vertical plane. Further, a new concept of incorporating interaural magnitude and phase features to construct a composite feature vector was proposed, and a more robust localisation performance could be obtained via the new composite feature vector.

The specific contributions of this chapter are as follows:

i This chapter proposed a method to extract the HRTF spectral envelop by truncating high order cepstral coefficients. The high order cepstral truncation would remove the formant energy of speech signal, which reduce the influence of speech signal on HRTF spectra.

ii A generic normalisation operation in cestral domain is proposed for reducing speech impacts and extracting HRTF spectra envelop. With this generic normalisation, the influence of speech spectra in spectral feature extraction could be further reduced. By Combining with cepstral domain truncation and the generic normalisation operation in the cepstral domain, a novel monaural HRTF localisation method was proposed for median plane localisation.

iii A novel composite feature vector combining the interaural features for 3-D space localisation was created and introduced. We demonstrated that this feature vector incorporated the spatial characteristics of magnitude and phase features, which could lead to more robust localisation performance.

Chapter 4

Individualised Interaural Feature Learning and Probabilistic Model

Overview: *This chapter introduces a probabilistic localisation mapping model based on the importance analysis of localisation features, especially for the vertical dimension localisation. The approach uses the Mutual Information (MI) as a metric to evaluate the most valuable frequencies of the Interaural Phase Difference (IPD) and Interaural Level Difference (ILD). Then, by using the famous RF algorithm and embedding the MI as feature selection criteria, the feature selection procedures are encoded with training of the localisation mapping. Hence, the trained mapping model is capable of using interaural features more efficiently, and, because of the multiple-tree-based model structure, the localisation model shows robust performance to noise and interference. By connecting with the Direct Path-Relative Transfer Function (DP-RTF) estimation, we propose to devise a novel localisation approach that has remarkable performance in the presence of noise and reverberation. The proposed mapping model is compared with the state-of-the-art manifold learning procedure in different acoustical configurations, and a more accurate and robust output can be observed.*

4.1 Introduction

Sound source localisation has shown its importance in various fields, such as video conferencing, virtual reality, humanoid robot interactions and hearing aids. There are many existing methods localising an acoustic source with different microphone array configurations. However, in many source localisation scenarios, because of the demand for mobility and low computational complexity, the sensor array size is limited; thus, the dual-channel localisation system has recently become popular [109–111]. Such a dual-channel acoustical sensing system can be found in the

hearing systems of mammals, with effective localisation performance. For example, accurately localising a sound source in a 3-D space over various environmental conditions is an easy task for humans, using only two ears and prior acoustical knowledge of their auditory system. The objective of binaural localisation is to mimic the human hearing mechanism by pinpointing a sound source location using a binaural microphone setup [30, 33, 41, 69].

Much behavioural and psychoacoustic evidence has confirmed that two individualised interaural cues, ITD and ILD, play an essential role in localising a sound source. Through intensive studies of the human hearing mechanism, it has become widely understood that individualised ITD and ILD are mainly determined by the filtering of the head, body and pinna, where the filtering function is the so-called HRTF. During the last two decades, many binaural localisation approaches have been developed based on interaural cues. Li and Levinson [112] adopted the Bayes rule to estimate the source position. This method uses interaural intensity difference (IID) and spectral signals to refine the initial estimate given by ITD. Viste *et al.* [93] provided an individual parametric model jointly using ILD and ITD for azimuth localisation. The model first provides a rough estimate of high standard variation based on ILD, and then refines the estimation by selecting the closest ITD. Woodruff and Wang combined ILD and ITD with a GMM, which also embedded the environmental condition as a latent parameter [41].

Further, many works have attempted to investigate the relationship between interaural cues and source position in a 3-D space. Duba summarised the relationship between interaural cues and azimuth/elevation [29]. Keyrouz *et al.* proposed a maximum cross-correlation matching method in source cancellation algorithms to pinpoint the source location in a 3-D space [18,30]. Liu *et al.* introduced a Bayesian rule-based hierarchical localisation system using IID time delay compensation and an interaural matching filter to estimate azimuth and elevation in the interaural-polar coordinate system [69, 111]. Deleforge *et al.* proved the local linear bijective mapping between interaural cues and source location on the binaural manifold, and derived a statistical localisation model using an EM algorithm [32, 33]. Weng *et al.* adopted a non-parametric tree-based learning method to retrieve the mapping between the interaural cues and source locations with fewer restrictions on its spatiotemporal characteristics and environment structure [31].

A number of the studies mentioned above have demonstrated the importance of feature spectrum selection to localisation performance, especially for elevation localisation, and various methods for feature selection have been proposed. Duba *et*

al. indicated that the IID at the frequency above 2 kHz contains the characteristics of head shadow and pinna diffraction [29], which can be exploited for elevation estimation. Algazi *et al.* retrieved the existence of elevation-dependent features at low frequency, which was attributed to head diffraction and torso reflections [54]. Deleforge *et al.* proved the elevation ambiguities of IPD at the high-frequency range (between 2 to 8 kHz) and used a feature space concatenated by full-spectrum ILD and low-frequency IPD [33].

Another challenge in sound source localisation is localising the source in the presence of noise and reverberation. There have been many attempts to obtain the interaural features in a complex environment and diminish the effects of room reverberation and background noise [44, 63, 65]. However, prior knowledge of the rooms acoustical characteristics is required for most of those methods, and, without this prior knowledge, the extracted interaural features may not be optimised and some interference would be retained. Therefore, a model with high tolerance to noise and these interferences would be expected for practical applications.

This chapter investigates the dependency between interaural features and sound source position, particularly for elevation mapping. The MI technique is used to demonstrate the existence of elevation-related dependency on both the ILD and IPD spectrum. Then, based on the results of the MI analysis, a chain rulebased probabilist model for mapping the features and source locations in a 3-D space is proposed. The model is constructed based on probability estimation trees (PETs) and RF framework. The PETs are formed by an information gainbased data partition technique. The features exploited by the model are further investigated, which justifies that the non-linear-dependent features (e.g., high-frequency IPD spectrum) can also contribute to an accurate localisation performance. Then, aiming at evaluating the accuracy and robustness of the proposed method, the localisation performances are compared with the state-of-the-art PPAM method proposed in [33] in different noise and reverberation environments without prior knowledge of the acoustical conditions.

4.2 Individualised Feature Selection Using Mutual Information

In this chapter, we focus on creating a more robust mapping between the spatial cues extracted in Chapter 3 and the source position. Considering a complex environment,

the additive noise and reverberations would distort the characteristics of both the phase and magnitude features, especially at higher frequencies (above 3 kHz) where the speech energy is comparatively lower. This is exacerbated, as most elevation localisation cues being generated by reflections and diffraction of the human body and pinna are above 3 kHz [54]. Therefore, a feature selection mechanism that maximises the direction-dependent information and minimises the effect of noise would be necessary, and a generic feature selection approach could lead to superior performance, as shown in Figure 3.6. To investigate the dependency between the features and their corresponding source positions, the Mutual Information (MI) that exists between each spatial localisation cue and the source location for particular feature spectra could be used as criteria to evaluate both the effectiveness and robustness of the feature vectors extracted for the localisation process.

In this section, the interaural phase angle and magnitude vector are denoted as:

$$\begin{aligned} \mathbf{v}^p &\triangleq [v_1^p, ..., v_L^p] \\ \mathbf{v}^m &\triangleq [v_1^m, ..., v_L^m] \end{aligned} \tag{4.1}$$

where capital letter L indicates the vector length when $f_{\min} = 0$ and $f_{\max} = f_s/2$. Here, the upper and lower limit of selected frequency for the phase feature and magnitude feature are identical, and are denoted as $f_{\min}$ and $f_{\max}$, respectively.

We then define the original feature vector $\mathbf{v} \in \mathbb{R}^{2L}$ by concatenating the phase vector $\mathbf{v}^p$ and the magnitude vector $\mathbf{v}^m$ as:

$$\mathbf{v} \triangleq [\mathbf{v}^p, \mathbf{v}^m] = [v_1, ..., v_\mu] \tag{4.2}$$

where v_μ define the element in the feature vector $\mathbf{v}$ with element index $\mu = 1, ..., 2L$.

4.2.1 Mutual Information Computation

As one of the most common measures to evaluate the dependency between variables, MI is widely used to estimate the maximally relevant feature selection that corresponds to a particular outcome [113, 114]. The MI $\mathcal{I}$ of two discrete variables, x and y, is defined as:

$$\mathcal{I}(x; y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}, \tag{4.3}$$

where $P(x)$ and $P(y)$ represent the joint probability and marginal probabilities of variable x and y, respectively. Similarly, the MI between a interaural feature (v_μ

and azimuth/elevation is given by:

$$\begin{aligned}\mathcal{I}_{\vartheta}(v_{\mu};\vartheta) &= \sum_{v_{\mu}}\sum_{\vartheta} P(v_{\mu},\vartheta)\log\frac{P(v_{\mu},\vartheta)}{P(v_{\mu})P(\vartheta)}\\ \mathcal{I}_{\varphi}(v_{\mu};\varphi) &= \sum_{v_{\mu}}\sum_{\varphi} P(v_{\mu},\varphi)\log\frac{P(v_{\mu},\varphi)}{P(v_{\mu})P(\varphi)}\end{aligned} \tag{4.4}$$

We now focus on the interaural phase and magnitude spectra feature analysis for elevation localisation, since the spectra information is the major cue to determine the elevation [58]. The dependency between spectra feature and azimuth would be skipped, because the localisation of azimuth relying on the signal arriving time and level difference between two ears is comparably easier and more robust in an interaural-polar system, as mentioned in Chapter 2.

In the elevation dependency analysis, the training dataset is denoted as $\{\mathbf{v}^{(n)},\varphi^{(n)}\}_{n=1}^{N}$, where $n = 1, ..., N$ represents the training data index. Based on consideration of computation complexity, the probability components $P(v_{\mu},\varphi)$, $P(v_{\mu})$ and $P(\varphi)$ are estimated by histogram-based probability estimation [115] as:

$$\begin{aligned}P(v_{\mu},\varphi) &= \frac{1}{N}\sum_{n=1}^{N} z(v_{\mu}^{(n)} = v_{\mu},\varphi)\\ P(v_{\mu}) &= \frac{1}{N}\sum_{n=1}^{N} z(v_{\mu}^{(n)} = v_{\mu})\\ P(\varphi) &= \frac{1}{N}\sum_{n=1}^{N} z(\varphi^{(n)} = \varphi)\end{aligned} \tag{4.5}$$

where the indicator function $z(\mathbf{Z})$ with event $\mathbf{Z}$ is defined as:

$$z(\mathbf{Z}) = \begin{cases} 1 & \mathbf{Z} \text{ is true,}\\ 0 & \text{otherwise.}\end{cases} \tag{4.6}$$

Therefore, the MI between each feature element v_{μ} from vector$\mathbf{v}$ can be obtained by (4.5). Then, by traversing all the elements in vector $\mathbf{v}$, the MI vector $\mathcal{I}(\mathbf{v};\varphi)$ is defined as:

$$\mathcal{I}(\mathbf{v};\varphi) = [\mathcal{I}(v_1;\varphi), ..., \mathcal{I}(v_{\mu};\varphi)] \tag{4.7}$$

Now in order to evaluate the dependency between elevation β and interaural cues

v_μ only and omit the impact of azimuth, the mutual information is calculated on each sagittal plane separately and the mutual information vector for each sagittal plane is denoted as $\mathcal{I}_\alpha(\mathbf{v};\varphi)$.

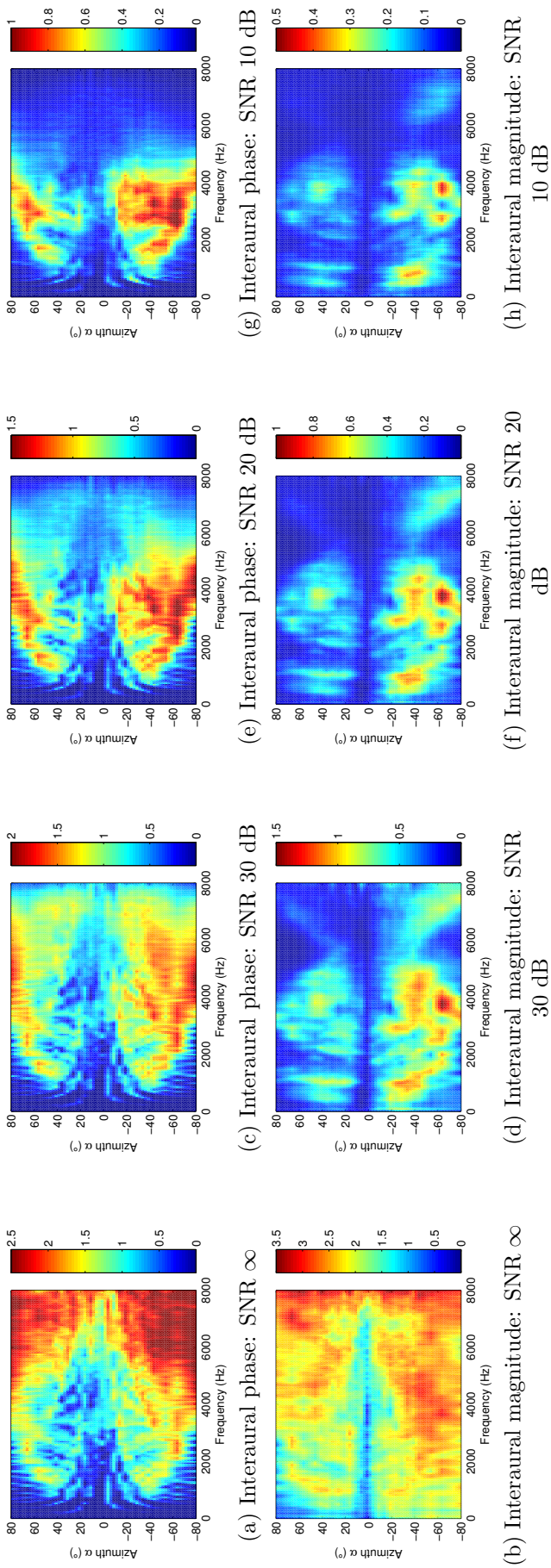
4.2.2 Analysis of Mutual Information in Interaural Cues

Figure 4.1 illustrates the variations in MI that exist between the elevation angle of the source location and the spatial cues for different azimuths and SNRs. From Figures 4.1(a) and (b), it can be observed that, in high SNR scenarios, the MI in the high-frequency range becomes dominant for elevation localisation. However, with the decreasing SNRs, the high-frequency cues no longer provide reliable localisation information, unlike the mid-frequency spatial cues. Further, the distribution (in frequency) of the most effective cues varies with different azimuths, and illustrates both the difficulty of decoupling the localisation process into separate azimuth and elevation estimation problems, as well as the challenge of localisation in the median plane [54]. Especially on the median plane, the low MI values at median plane comparing to other azimuths is because the interaural difference has smaller variation on the median plane.

Further, comparing the behaviour of the two types of spatial cues, we can observe that the importance of each changes with the SNR. For example, where no or low noise is present, the interaural magnitude cues become dominant, while, in a comparably higher noise environment, the interaural phase cues show more robustness. Collectively, these observations imply that the selection of spatial cues for the creation of a feature vector for localisation must be more nuanced than the simple selection of a fixed frequency range; thus, an adaptive noise-dependent feature selection and extraction process becomes a necessity for any noise-robust binaural localisation system. A spatial feature learning algorithm that is aware of the MI contained in each spatial cue can satisfy this requirement, and provides superior performance to the former approach, as illustrated in the following sections.

Step 1: Estimate the noise power $\mapsto \sigma^2$.
Step 2: Evaluate $\mathcal{I}_\vartheta(\mathbf{v};\varphi)$; Mutual Information of spatial cues.
  foreach φ *in the HRTF dataset* **do**
    foreach $s(t)$ *in the training speech dataset* **do**
      Compute $X_{i,k}(f_\mu)$ for a simulated noise power σ^2.
      Calculate the corresponding v_μ and $\mathbf{v}$.
      Estimate $\mathcal{I}_\vartheta(\mathbf{v};\varphi)$.
    end
  end
  Result: Obtain the set of $\mathcal{I}_\vartheta(\mathbf{v};\varphi)$ for a noise power σ^2.
Step 3: Learn the optimal combination of the spatial cues in $\mathbf{v}$.
  Calculate a mean MI $\forall\ \varphi$ from $\mathcal{I}_\vartheta(\mathbf{v};\varphi)$.
  for $l' \leftarrow 1$ **to** $2L$ **do**
    Rearrange $\mathbf{v}$ in descending order of $\mathcal{I}_\vartheta(v_\mu;\varphi)$.
    foreach $\mathbf{v}$ *derived from the training speech dataset* **do**
      Estimate the source location φ from a feature vector $\mathbf{v}$ of length l'.
      Calculate the angular localization error of φ.
    end
  end
  Result: Obtain $\mathbf{v}'$; the rearranged and selected spatial feature vector from $\mathbf{v}$, that corresponds to the minimum mean angular localization error.

Algorithm 1: Spatial feature learning for robust localization.



(a) Interaural phase: SNR ∞

(b) Interaural magnitude: SNR ∞

(c) Interaural phase: SNR 30 dB

(d) Interaural magnitude: SNR 30 dB

(e) Interaural phase: SNR 20 dB

(f) Interaural magnitude: SNR 20 dB

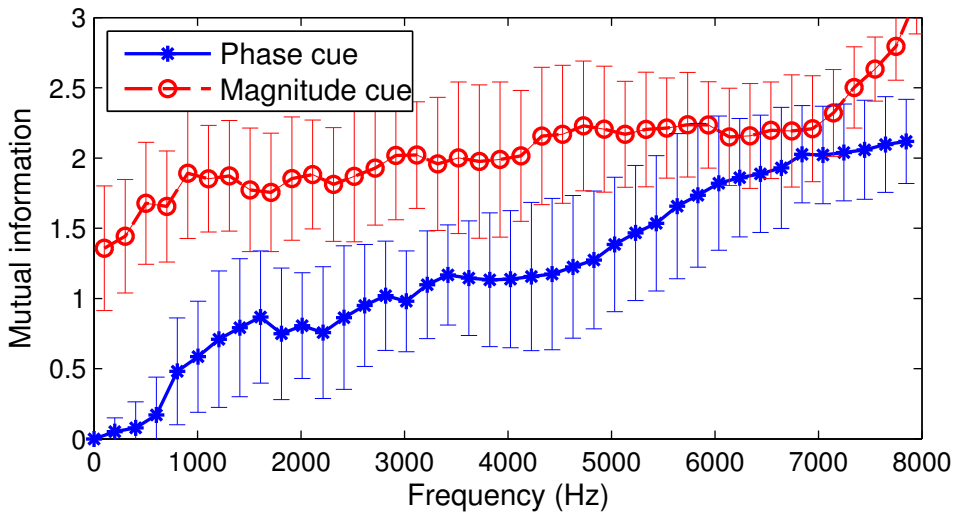
(g) Interaural phase: SNR 10 dB

(h) Interaural magnitude: SNR 10 dB

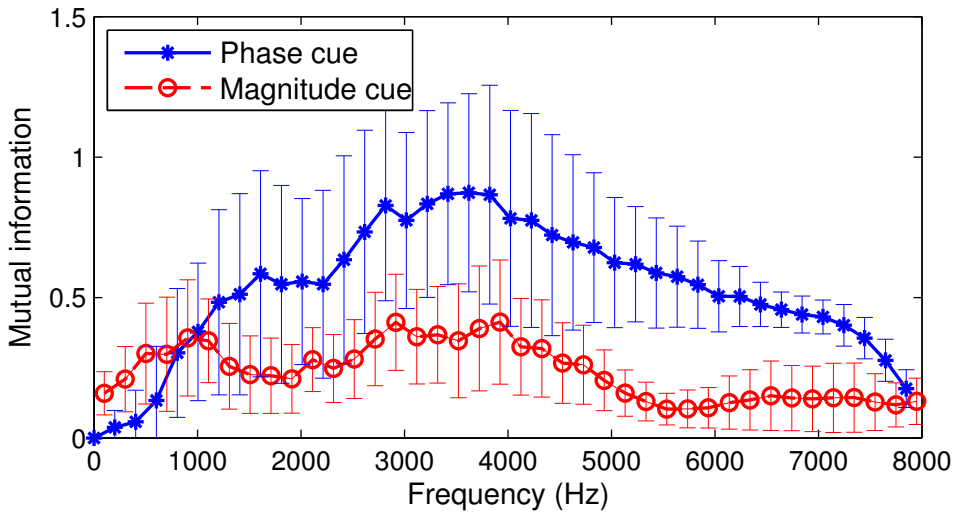
Figure 4.1: Mutual Information between the spatial cues and the elevation for a range of azimuths, frequencies and noise conditions.

4.2.3 Spatial Feature Learning and Selected Feature Vector

Algorithm 1 describes the MI-based spatial feature learning mechanism used in the remainder of this work. Given an estimated noise level, such as using the method proposed in [116], the spatial feature vector $\mathbf{v}$ and its corresponding MI in (4.7) are computed for a set of training speech signals. Figure 4.2 illustrates the variation of the MI between spatial cues and elevation angle with and without noise, and indicates the changing nature of the importance of individual spatial cues. The error bar reflects the MI variation on different azimuth planes, and the mean MI obtained for the training speech dataset is used thereafter to create a rearranged spatial feature vector $\mathbf{v}'$.



(a) SNR: ∞



(b) SNR: 20 dB

Figure 4.2: MI variation in spatial cues with respect to SNR.

To arrive at $\mathbf{v}'$, an optimal number of spatial cues to be used in the localisa-

tion process l' is computed. The average angular localisation error, obtained from the estimated source location (ϑ, φ) and its estimate $(\hat{\vartheta}, \hat{\varphi})$, is used as a metric to determine the optimal l'. This results in a set of spatial feature vectors, $\mathbf{v}'(\vartheta, \varphi)$, applicable to the specified noise level (in the case of some practical applications, it is also possible to pre-train the system for a set of known, approximate noise conditions). The spatial cues extracted from the received signals are rearranged similarly, and the resultant feature vector $\hat{\mathbf{v}}'$ and the $\mathbf{v}'(\vartheta, \varphi)$ reference features are used to localise by applying (3.29) in Chapter 3. The localisation performance will be presented in Section 4.6.

4.3 Probabilistic Localisation Model and System Design

In the previous section, a feature selection approach based on averaged MI was introduced. The average MI approach provides a selection of the most dependent spectra characteristics, while the actual feature value and most related directions are not considered. In this chapter, those two feature properties are investigated and integrated with the localisation mapping progress. With the proposed mapping operation, not only are the features exploited in advance, but the localisation robustness will also be further improved.

Aiming at localising the sound source in a real-world environment with presence of noise and interference, the localisation problem is considered probabilistic mapping between the feature vector and source location $\mathbf{\Theta}_{\text{IP}} = (\vartheta, \varphi)$, which can be expressed as:

$$\hat{\mathbf{\Theta}}_{\text{IP}} = \arg\max_{\mathbf{\Theta}_{\text{IP}}} P(\mathbf{\Theta}_{\text{IP}}|\hat{\mathbf{v}}) \tag{4.8}$$

where $P(\mathbf{\Theta}_{\text{IP}}|\hat{\mathbf{v}}_p)$ represents the probability of the sound source located at $\mathbf{\Theta}_{\text{IP}}$ with a given feature vector $\hat{\mathbf{v}}_p$. In the interaural-polar coordinate system, the generic ITDs are the same when the sources are placed on the same sagittal plane; thus, it is more applicable to estimate the azimuth ϑ first, and then estimate the elevation φ with a given ϑ. Therefore, the probability $P(\vartheta|\hat{\mathbf{v}}_p)$ can be represented based on the conditional chain rule by a multiplication of two posteriors as:

$$P(\mathbf{\Theta}_{\text{IP}}|\hat{\mathbf{v}}_p) = P(\vartheta|\hat{\mathbf{v}})P(\varphi|\vartheta, \hat{\mathbf{v}}) \tag{4.9}$$

where $P(\varphi|\hat{\mathbf{v}}_p)$ represents the probability of the source located at the sagittal plane, labelled by ϑ, and $P(\varphi|\vartheta, \hat{\mathbf{v}}_p)$ represents the probability of β with given α and $\hat{\mathbf{v}}_p$. Then, by substituting (4.9) into (4.8), the estimated source location $\hat{\mathbf{\Theta}}_{\mathrm{IP}} = \hat{f}_{\mathbf{\Theta}_{\mathrm{IP}}}(\hat{\mathbf{v}}_p)$ can be obtained. There are two obvious advantages of adopting such localisation architecture. First, in the training process, the feature characteristics caused by ϑ and φ are separated, so the localisation model can be more explainable and a more accurate localisation result can be achieved [44]. Second, in the testing process, the ϑ and φ can be estimated in parallel, and then high computation efficiency in contemporary parallel hardware can be achieved.

To estimate the $P(\alpha|\hat{\mathbf{v}}_p)$ and $P(\beta|\alpha, \hat{\mathbf{v}}_p)$ for the real-world source localisation, the estimation approaches should be able to tackle the amount of noise in the input feature vector. In the next section, the RF method is introduced to select the most important features and provide a robust estimation.

4.4 Feature Dependency Analysis and Assembled Data Partition Model

This section presents a posterior estimation method with adaptive feature selection approach. First, the dependency between the spatial parameters and feature characteristics is analysed, and then an assembled data partition model based on the feature dependency is adopted for the posterior estimations. During the model training process, the following notations are used in the later content: let $\{\mathbf{v}^{(n)}, \mathrm{y}^{(\mathrm{n})}\}_{\mathrm{n=1}}^{\mathrm{N}}$ denote the dataset of all training data pairs, where $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^{2K}$ represents the interaural feature vectors in vector space $\mathcal{V}$, and $\mathrm{y}^{(\mathrm{n})} \in \{\mathrm{y}_1, \mathrm{y}_2, ...\mathrm{y}_\mathrm{D}\}$ represents the discrete spatial label. In the training dataset, either α or β is discretised and labelled. Here, both labels of θ or φ are represented by the common symbol y for simplicity. The subscripts $n \in \{1...N\}$, $\mu \in \{1...2K\}$ and $d \in \{1...D\}$, respectively, represent the training observations index, the feature vector components and the discrete label index.

4.4.1 Data Partition and Tree-structured Model

The dependency analysis described in Section 4.2 justifies the existence of elevation-related features at full-band ILD and high-frequency IPD. To use these results and further exploit the feature characteristics, the mapping model should satisfy two essential requirements. First, the model should be able to handle both linear and

non-linear associations. Second, an adaptive feature selection should be embedded in the model training progress. Therefore, to fulfil these two requirements, a recursive tree-structured data partition techniquethe so-called decision tree methodis introduced. A tree is a hierarchical structure in which each internal node contains a specific feature condition. At each node, the training data are split into two sub-spaces by the splitting feature that maximises the MI between the selected feature and the spatial labels from sub-spaces. Notably, the MI here is also known as information gain in machine learning, which is still denoted by $\mathcal{IG}$ in latter content. Finally, each node at the terminal of the tree is defined as a leaf node, which represents a subclass of spatial labels y.

Formally, let $m \in \{1...M\}$ denote the splitting node index, and the subset of training data at mth node are defined as $\{\mathbf{v}_m^{(n)}, \mathrm{y}_\mathrm{m}^{(\mathrm{n})}\}_{\mathrm{n=1}}^{\mathrm{N_m}}$, where the subset of the feature vectors and the target labels are respectively represented by $\mathcal{V}_m \subset \mathbb{R}^{2K}$ and $\mathrm{y}_{\mathrm{d_m}}^{(\mathrm{n})} \in \{\mathrm{y}_1, \mathrm{y}_2, ...\mathrm{y}_{\mathrm{D_m}}\}$. Defining $\mathcal{V}_{m,\mu} = \{v_{m,\mu}^{(n)}\}_{n=1}^{N_m}$ as μth feature subset, the information gain with the binary splitting operation using the splitting value $v_{m,\mu}$ can be obtained by:

$$\begin{aligned}\mathcal{IG}(\mathrm{y}, \mathrm{v}_{\mathrm{m},\mu}) = &\sum_{d=1}^{D} P_m(v_{m,\mu}, \mathrm{y_d}) \log \frac{\mathrm{P_m}(\mathrm{v}_{\mathrm{m},\mu}, \mathrm{y_d})}{\mathrm{P_m}(\mathrm{v}_{\mathrm{m},\mu})\mathrm{P_m}(\mathrm{y_d})} \\ &+ \sum_{d=1}^{D} (1 - P_m(v_{m,\mu}), \mathrm{y_d}) \log \frac{1 - \mathrm{P_m}(\mathrm{v}_{\mathrm{m},\mu}, \mathrm{y_d})}{\big(1 - \mathrm{P_m}(\mathrm{v}_{\mathrm{m},\mu})\big)\mathrm{P_m}(\mathrm{y_d})}\end{aligned} \tag{4.10}$$

which can also be interpreted as subtraction of entropies by:

$$\mathcal{IG}(\mathrm{y}, \mathrm{v_{m,k}}) = \mathrm{H(y)} - \mathrm{H(y|v_{m,k})}. \tag{4.11}$$

It is more directly perceived through (4.11) that the $\mathcal{IG}$ is evaluating the decreasing uncertainty of y after splitting by $v_{m,\mu}$.

The probability components $P_m(v_{m,\mu}, \mathrm{y_d})$, $P_m(v_{m,\mu})$ and $P_m(\mathrm{y_d})$ are estimated by histogram-based probability estimation [115] as:

$$\begin{aligned}P_m(v_{m,\mu}, \mathrm{y_d}) &= \frac{1}{N_m} \sum_{n=1}^{N_m} z(v_{m,\mu}^{(n)} \leq v_{m,\mu}, \mathrm{y} = \mathrm{y_d}) \\ P_m(v_{m,\mu}) &= \frac{1}{N_m} \sum_{n=1}^{N_m} z(v_{m,\mu}^{(n)} \leq v_{m,\mu}) \\ P_m(\mathrm{y_d}) &= \frac{1}{N_m} \sum_{n=1}^{N_m} z(\mathrm{y}_{\mathrm{m,d}}^{(\mathrm{n})} = \mathrm{y_d})\end{aligned} \tag{4.12}$$

where the indicator function $z(\cdot)$ is defined the same as (4.7).

Depending on (4.10) and (4.12), one may split the dataset by selecting the feature index μ and its splitting value $v_{m,\mu}$, which maximises the value of information gain as:

$$\tilde{v}_{m,\mu} = \arg\max_{v_{m,\mu} \in \mathcal{V}_m} \mathcal{IG}(\mathrm{y}, \mathrm{v_{m,k}}) \tag{4.13}$$

where $\tilde{v}_{m,\mu}$ indicates the optimized splitting value with its feature index μ at node m.

By repeating this splitting operation until some certain criterion is reached (e.g., the maximum number of nodes M), the original dataset $\{\mathbf{v}^{(n)}, \mathrm{y}^{(\mathrm{n})}\}_{\mathrm{n=1}}^{\mathrm{N}}$ is partitioned into multiple subsets denoted as $\{\mathbf{v}_{\lambda}^{(n)}, \mathrm{y}_{\lambda}^{(\mathrm{n})}\}_{\mathrm{n=1}}^{\mathrm{N}_\lambda}$ with subset index λ. The whole training progress can also be understood as a recursive clustering process, whereby the training data with common feature characteristics described by v_k^m are clustered into the same subset λ, while the uncertainty of spatial label α is minimised, so that, in such a subset, the training data only share a few α. Further, in this recursive approach, the splitting feature index k and its corresponding value v_k^m are adaptively re-selected via (4.13) at each node, so the most spatial dependent feature characteristics are exploited to the maximum.

Then, with a test feature vector $\hat{\mathbf{v}}$, the model will first categorise it into a subset with the splitting criteria obtained during training, and then the estimated posterior $\hat{P}_\gamma(\mathrm{y_d}|\hat{\mathbf{v}})$ from a single tree γ is given by:

$$\hat{P}_\gamma(\mathrm{y_d}|\hat{\mathbf{v}}) = \frac{1}{\mathrm{N}_\lambda} \sum_{\mathrm{n=1}}^{\mathrm{N}_\lambda} \mathrm{z}(\mathrm{y}^{(\mathrm{n})} = \mathrm{y_d}|\mathrm{y}^{(\mathrm{n})} \in \mathrm{y}_\lambda)\Big|_{\hat{\lambda}=\mathcal{C}_\gamma(\hat{\mathbf{v}})} \tag{4.14}$$

where $\mathcal{C}_\lambda(\cdot)$ indicates the classification approach in the tree γ based on the splitting features, and $\hat{\lambda}$ represents the estimated subset index. To avoid over-fitting and increase the robustness of the model, a committee of multiple single trees is constructed in the next section.

4.4.2 Random Forest Bagging and Unbiased Probability Estimation

A more robust posterior estimation can be obtained by an ensemble model, and the approach to assembling multiple decision trees is known as the tree bagging technique. The idea is to attain the final posterior probability estimation by averaging the probabilities from a committee of independent identically distributed (i.d.d.)

trees. Formally, in a forest with Γ trees, the assembled posterior estimate is given by [115, 117]:

$$\hat{P}(\mathrm{y_d}|\hat{\mathbf{v}}) = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\hat{\mathrm{P}}_{\gamma}(\mathrm{y_d}|\hat{\mathbf{v}}) \tag{4.15}$$

where $\gamma \in \{1, 2, ..., \Gamma\}$ is the index of i.d.d. trees, and the i.d.d. trees are obtained by introducing randomness into their training process. Many approaches have been proposed to randomise those training processes, and the bootstrap sampling method, adopted by the RF algorithm [117], has shown its effectiveness in many applications. The independence between decision trees is achieved by introducing randomness into both the training subset and feature selection. The training datasets for each tree are obtained via bootstrap sampling, in which roughly two-thirds of the data are selected for each subset.

The averaging operation in (4.16) softly selects those more confident trees from the forest, and such selection is undertaken on the leaves of the trees. In a RF, each i.d.d. tree can be understood as a union of hyper-rectangles defined by the subsets on each leaf, and the boundaries of each hyper-rectangle are determined by the splitting features v_k^m. Although those hyper-rectangles vary with different trees, they overlap because of the similarity of feature characteristics and spatial label clustering. Thus, the operation in (4.16) selects the area with the maximum amount of overlaps of hyper-rectangles.

Finally, the source position estimation is given by:

$$< \hat{\alpha}, \hat{\beta} >= \underset{<\alpha,\beta>}{\arg\max}\, \hat{P}(\alpha|\hat{\mathbf{v}}_p)\hat{P}(\beta|\alpha, \hat{\mathbf{v}}) \tag{4.16}$$

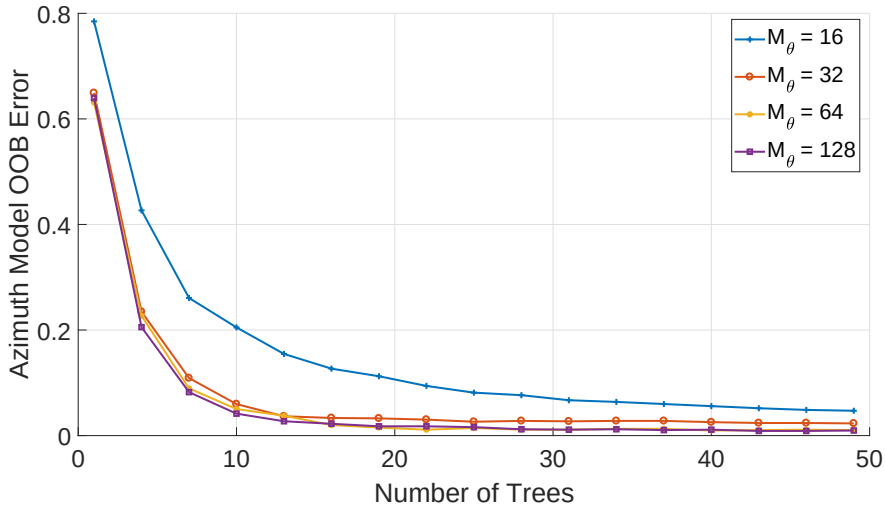
The following content presents a detailed description of model training and parameter selection, and demonstrates an interpretation of this model.

4.5 Model Training and Interpretation

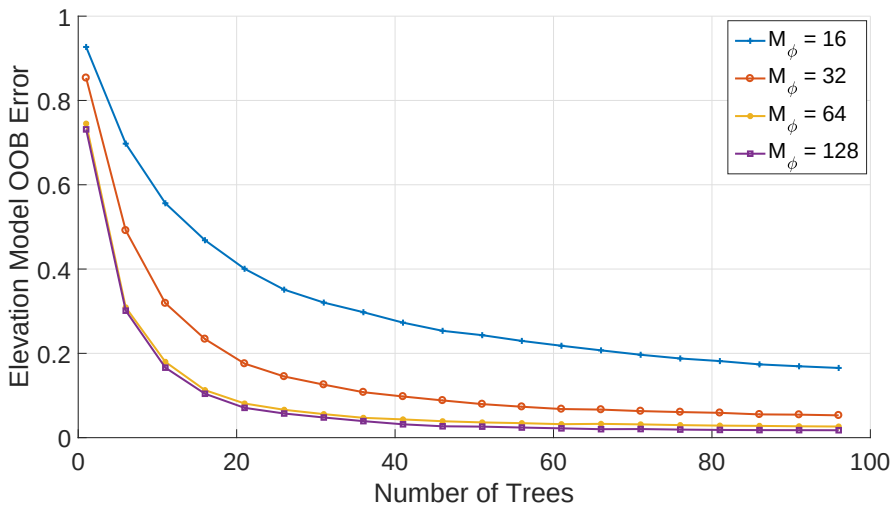
4.5.1 Model Training and Parameter Selection

A dataset of synthetic received signals is used for training purposes, and is generated by convolving the source signal and HRIR in the time domain. The Gaussian noise with about 0.5 s duration is adopted as the source signal and convolved with 003 subject HRIR from the CIPIC database [54], in which the HRIR were recorded at 25

sagittal azimuths and 50 elevations on each sagittal plane. The interaural features are extracted as described in Chapter 3, and the window length for STFT is settled as 16 ms with 8 ms shift, so there are 256 samples with a 16 kHz sampling rate. The interaural feature extracted from each frame is treated as a training sample, and 10 samples for each position are randomly selected from the dataset. Moreover, three training conditions with different SNR additive white noise are applied on the training dataset, with the aim of investigating the influence of training data quality.



(a) Azimuth Out-of-bag (OOB) error



(b) Elevation OOB error

Figure 4.3: OOB errors

In the proposed model, three parameters need to be settled during the training: the candidate feature number at each node, the splitting iteration number M and the tree number Γ. The initial value of K' can be set as the square root of the

total feature numbers in $\mathbf{v}$, according to [117], and the other two parameters for the azimuth and elevation model are determined based on an empirical approach. Fortunately, the sampling scheme of RF provides a convenient measurement to evaluate the model, which is known as out-of-bag (OOB) error. Recall that around two-thirds of training data were selected as training data after the bootstrap sampling for each tree, so that those unselected data could be used as testing data. The estimation error of the OOB data is defined as OOB error. Figure 4.3 shows the OOB errors versus Γ of the azimuth model and elevation model with different selection of M, where the splitting iteration number of the azimuth and elevation model are defined as M_θ and M_ϕ, and the tree numbers are defined as Γ_θ and Γ_ϕ, respectively. Intuitively, M should relate to the class label number and correspond to the accuracy of the model. It can be observed from Figure 4.3 that small OOB error differences are obtained after a certain number of iterationsthat is, $M_\theta = 32$ and $M_\phi = 64$which means the forests are fully developed for the classifications. As for the tree numbers, the OOB error tends to converge at $\Gamma_\theta = 30$ for the azimuth model and $\Gamma_\phi = 80$ for the elevation model. However, since a larger tree number can increase the robustness of the model, 20 extra trees are added on both models, and the final tree numbers are settled as $\Gamma_\theta = 50$ and $\Gamma_\phi = 100$.

4.5.2 Trained Model Interpretation

Figure 4.4 shows the changes in feature selection because of different training conditions, by comparing the feature usage account. The feature usage account is defined by the times of one feature has been used in the training. Given that the features with higher information gain are more likely to be selected as splitting features at each node, the usage account of a feature in the forest describes its dependency on the position estimation. It can be observed that, in the azimuth model trained with noise-free data, the more frequently used features are clustered at frequencies under 1 kHz, which correspond to the phase delays caused by the head width. However, in the model trained by a noisy dataset, the feature usage is more evenly distributed, which indicates that the model tends to use group delays between the ears for the estimation, which is because more features are necessary to a robust estimation with noisy training data. However, an opposite phenomenon is observed in the elevation model. With the decreasing of training data SNR, the more frequently used features are clustered to the frequency range around 3 to 5 kHz. This can be ascribed to the fact that more elevation information is generally drowned by the increasing

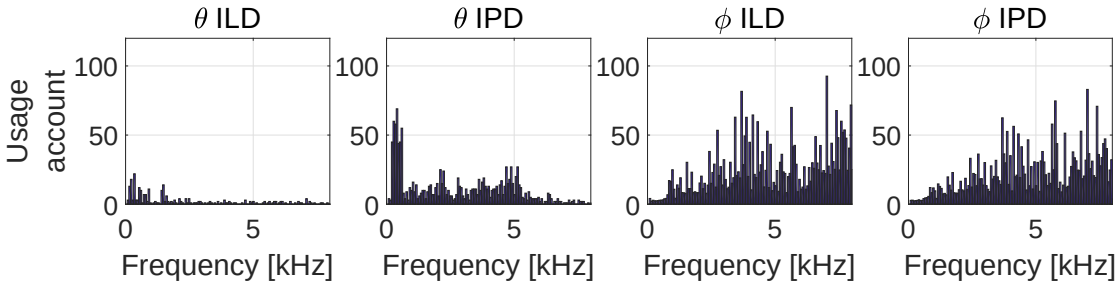
noise, which cannot provide valid localisation information. Such an adaptive selection process can only benefit the model performance when the noise is relatively low, which will diversify the feature usage and tree structures. However, when the noise is too strong and the most valuable features are degraded, the tree structures are homogenised and the robustness of the model will rapidly decrease.

Figure 4.5 demonstrates the feature characteristics captured by the splitting values from the direction $[30°, 45°]$ and $[30°, 135°]$. In each sub-plot, each column shows the additive noise SNR for a training dataset and each row shows the type of features. Each blue circle represents a left condition in a node from a single tree, which is also defined as the upper limit of $v_{m,k}$ in (4.12). Similarly, each red cross represents a right condition in a node or the lower-limit value of $v_{m,k}$. The ground truth features are plotted in black solid curves, which are obtained from pure HRTF data. Each marker, either a circle or cross, is considered a classifier to evaluate whether an extracted feature vector fits the conditions for a specific direction. By observing the distribution of markers, it again justifies that the features in frequency above 2 kHz are more valuable. The ground curves almost overlap with the dividing line between circles and crosses, which indicates that the major features have been captured by those upper and lower limits. In each row, the influence of the training condition can be compared. In the case where $SNR = \infty$, though the curve fits best, the markers are distributed closer, which means that the independence between trees is comparably low, which may lead to over-fitting on some features. Meanwhile, in $SNR = 10dB$, the markers are more dispersed and the dividing lines are blurred because of the high noise level, which will cause loss of some details of the feature characteristics, and lead to decreasing localisation accuracy.

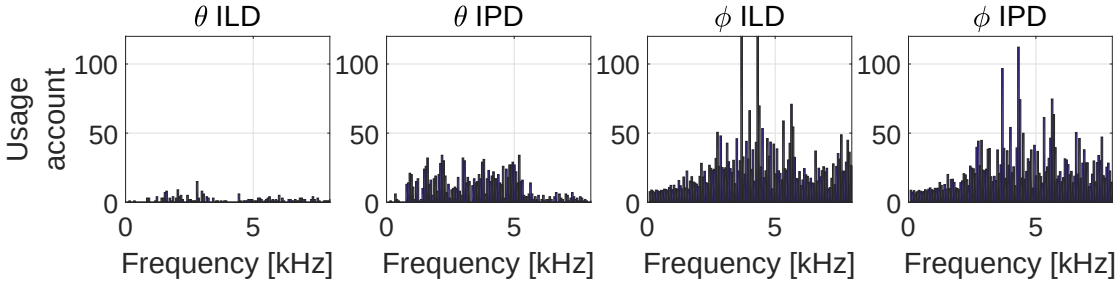
4.6 Experiments With Simulated Data

4.6.1 3-D Space Localisation with Mutual Informationbased Feature Selection

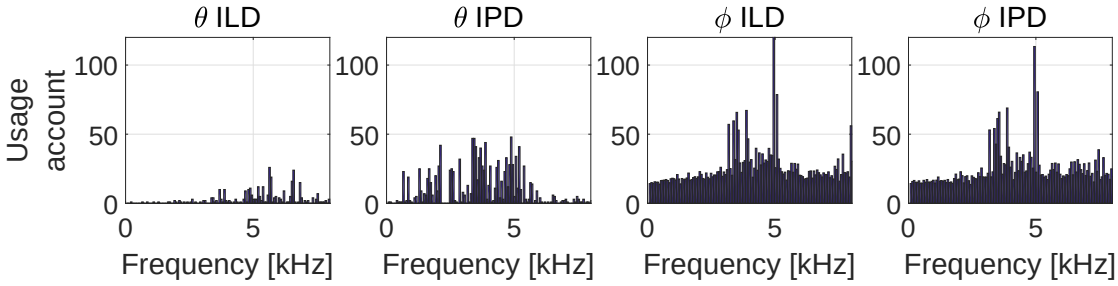
This section presents the simulation results of 3-D space localisation with MI-based feature selection. The localisation performance of the proposed method is compared with the generic composite feature-based localisation approach presented in Chapter 3 and a simple correlation-based method [17].



(a) SNR = ∞.

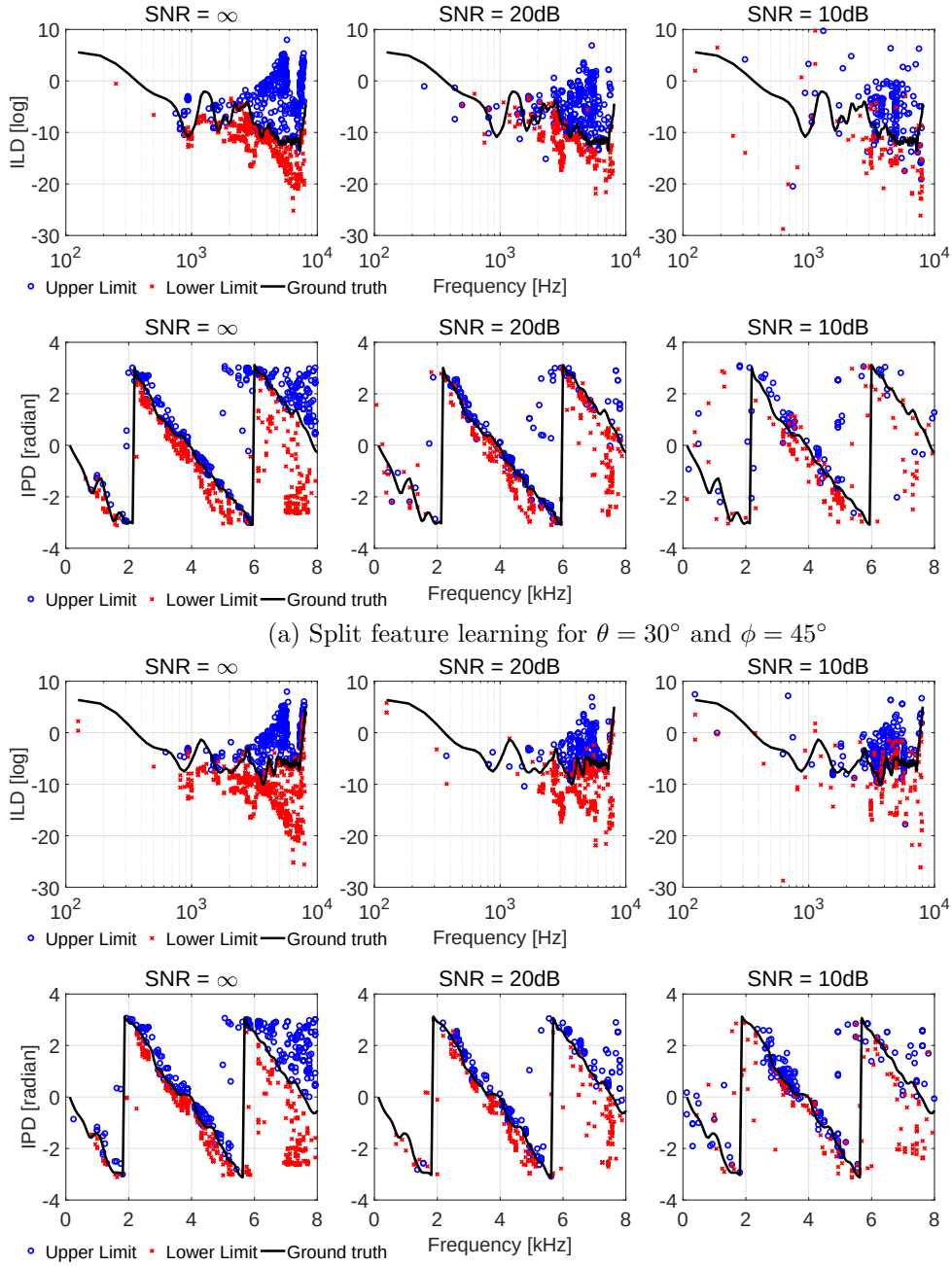


(b) SNR = 20dB.



(c) SNR = 10dB.

Figure 4.4: Feature Usage Accounts for training condition (a) SNR = ∞, (b) SNR = 20dB and SNR = 10dB. The first two columns shows the feature usage for azimuth model and the last two columns shows the average feature usage for elevation model.



(a) Split feature learning for $\theta = 30°$ and $\phi = 45°$

(b) Split feature learning for $\theta = 30°$ and $\phi = 135°$

Figure 4.5: Split feature value comparison between (a) $\theta = 30°$, $\phi = 45°$ and (b) $\theta = 30°$, $\phi = 135°$.

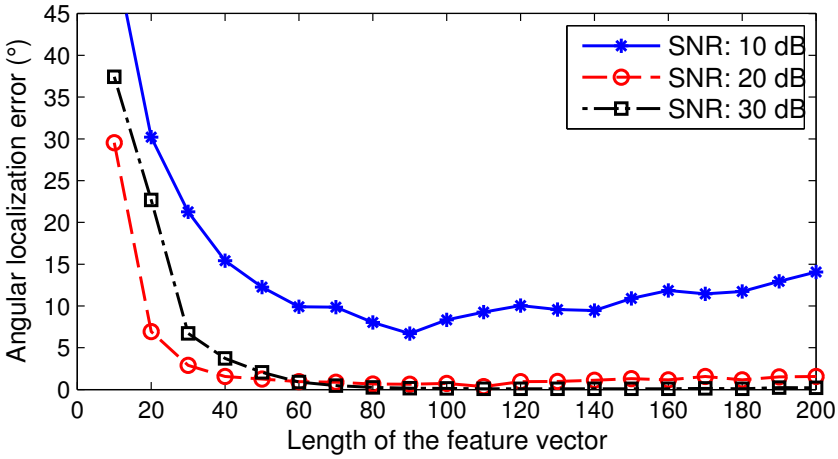


Figure 4.6: Localisation error with respect to feature vector length.

Simulation Configuration

The proposed approach is used to evaluate 950 source locations, ranging from azimuth $\alpha = -45°$ to $45°$ in $5°$ intervals and elevation $\beta = -45°$ to $230.625°$ in $5.625°$ increments, for the first 10 subjects HRTF measurements in the CIPIC database [54]. The speech samples from the PASCAL CHiME Speech Separation and Recognition Challenge [106] (34 males and females, each with 500 utterances sampled at 16 kHz) are used as inputs; 340 randomly selected utterances are used for the learning process, and a separate 200 utterances are used to evaluate the localisation performance. The binaural signals are simulated by convolving the HRTFs of different locations with the uncorrupted speech, and introducing the independent additive white Gaussian noise with three different SNRs of 10, 20 and 30 dB.

The frequency range for the generic composite feature-based approach is selected empirically, where [0,4] kHz and [3,5] kHz are the phase and magnitude feature regions for the feature-based method, while the full-band signal is used for the correlation approach. During the comparison, the mean angular error is employed as a metric to assess the localisation performance. The angular error denotes the angular distance between the estimated and actual source directions in the interaural-polar coordinate system; therefore, the estimation errors of both the azimuth and elevation (α and β) are implicitly included in the performance assessment.

Performance impact of the feature vector length

From Figures 4.1 and 4.2, it becomes apparent that the length of the feature vector $\tilde{\mathbf{v}}$ can directly influence the localisation performance. For example, a length smaller

Localization approach	**Mean angular localization error**		
	10 dB	**20 dB**	**30 dB**
Proposed learning	5.6°	0.9°	0.1°
Composite feature	24.3°	5.1°	0.9°
Cross-Correlation [17]	67.7°	58.6°	51.6°

Table 4.1: 3-D space localization performance comparison.

than the optimum will result in insufficient spatial information (especially in the case of the median plane), while a greater length could result in increased ambiguity because of the effects of noise. In both cases, the mean angular localisation error will be affected; thus, an optimum length for $\tilde{\mathbf{v}}$ that minimises this error must be computed at the noise power level observed in a particular localisation scenario. Hence, the training process described in Algorithm 1 is applied to a range of simulated speech inputs, and the optimum feature vector length and spatial cue combination is obtained dynamically based on their MI content.

Figure 4.6 illustrates the relationship between the mean angular errors and the length of the composite feature vector at different noise levels. The results are presented for three different SNRs, where the selected number of spatial cues varies from 10 to 200 in intervals of 10. The result for the 10 dB SNR case clearly illustrates the general behaviour discussed above (similar behaviour is observed at other noise levels as well), indicating an optimum feature vector length of approximately 90 elements. Note that the angular localisation error for the 30 dB scenario is larger than that for the 20 dB scenario when the feature vector length is less than 60. This suggests that a short feature vector may lead to unstable localisation performance; thus, a minimum length of the feature vector should be guaranteed.

Localisation performance

The performance of the proposed method is presented and compared with two other approaches in Table 4.1. Here, the received binaural inputs are obtained from 90 uniformly sampled source locations of the 950 locations in the HRTF dataset, and the resulting localisation error is averaged across multiple untrained speech inputs and source locations. The results indicate a significant improvement in performance over the generic composite feature-based localisation approach in Chapter 3, especially in the low SNR configurations. It was notable that the improvement predominantly stemmed from a reduction in front-to-back confusion. This suggests that the ap-

proach overcomes the lack of spectral cues located beyond the mid- to high-frequency ranges [118] that are less robust to the effects of noise. In general, the results suggest that the MI-based feature learning and rearrangement of the spatial cues in the feature vector can both improve the localisation performance and overcome the negative effect of the dynamic truncation of the feature vector to achieve greater robustness to noise.

4.6.2 3-D Space Localisation with Probabilistic Model

To evaluate the accuracy and robustness of the proposed method, localisation tests proceed with multiple testing conditions, and the localisation results are compared with the state-of-the-art PPAM method [33]. In the last part of this section, performance in different reverberant environments is also compared and presented.

Performance Measurements and Simulation Configuration

The received signals are generated by convolving the simulated BRIRs and speech utterances. The speech utterances are obtained from the PASCAL CHiME Speech Separation and Recognition Challenge database [106], which includes 34 speakers and every speaker has 500 utterances. Each testing speech utterance has around 1 s duration.

The BRIRs used in the testing are generated with CIPIC HRTF subject 003 and the Roomsim MATLAB program [119] based on the image method in four empty shoebox rooms. A rectangular room with room sizes 5 m $\times$ 5 m $\times$ 3 m is created in the simulation. The subject head is fixed at (2.5, 2.5, 1.2) m and the sound source is positioned around the subject with the same position scheme as used in the CIPIC database, so there are 1,250 source positions (25 azimuths $\times$ 50 elevations) in total during the tests. The distance between the sound source and the centre of the head is fixed as 1.5 m so that the acoustical environment can be considered a far-field environment. The strengths of reflections are manipulated by the absorption coefficients β on the walls. In the additive noise test, four sets of data are generated and tested with different SNRs. The absorption coefficients are set as $\beta = 1$, so no reflections present and the interference is entirely caused by noise. As for the reverberation test, the absorption coefficients are tuned from 0.8 to 0.2 to obtain the corresponding T_{60} varies from 100, 200, 300, 400 and 500 ms.

The localisation accuracy is evaluated by the correct rate of localisation. An estimation is considered correct if the difference between the estimated sound direction

and ground truth source position is below a certain tolerance threshold. To objectively demonstrate the 3-D position differences between the ground truth source position and the estimated direction, the angular error defined by the absolute angular difference between two directional vectors in a Cartesian coordinate system is calculated as:

$$\epsilon = \arccos \frac{\mathbf{d}_{<\alpha,\beta>} \cdot \hat{\mathbf{d}}_{<\hat{\alpha},\hat{\beta}>}}{|\mathbf{d}_{<\alpha,\beta>}||\hat{\mathbf{d}}_{<\hat{\alpha},\hat{\beta}>}|} \tag{4.17}$$

where $\mathbf{d}_{<\alpha,\beta>}$ and $\hat{\mathbf{d}}_{<\hat{\alpha},\hat{\beta}>}$ are the ground truth source direction vector and the estimation direction vector in Cartesian coordinate system, respectively. The azimuth and elevation correct rate in interaural polar system are also presented.

Localisation Performance with Different Training Environment

As discussed in 4.5.2, different training conditions also affect the quality of the model, which leads to different localisation performance. In this section, the localisation performances with different training conditions are presented and compared.

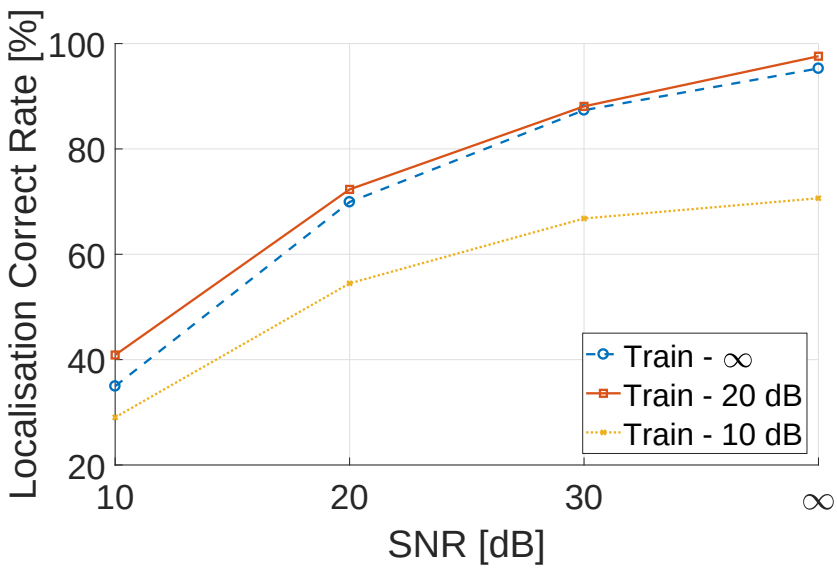


Figure 4.7: Comparing the localisation accuracy with different training conditions. The angular error tolerance is 2.5°.

In general, the localisation accuracy for all models decreases with increasing noise level. As predicted in Section 4.5.2, the model trained with moderate noise conditions performs best, and slightly outperforms the one trained with no noise data. This is because the additive zero-mean Gaussian noise increases the variance of the training data, which introduces diversity in the splitting values for all features

and forces trees to select the features that can still provide valuable information under interference. However, the localisation result of the model trained in a strong noise environment entirely degrades because the model cannot capture the details of feature characteristics from noisy data, as demonstrated in the third column of 4.5. According to this comparison result, in the following simulations, moderate noise data are used as training data for both the proposed and reference method for comparison.

Localisation Performance with Additive Noise

A localisation accuracy comparison with the presence of additive noise is shown in Figure 4.8. In general, the proposed method outperforms the state-of-the-art method with both feature vector types. It can be concluded that the RF algorithm in the proposed method is capable of constructing a finer mapping between the feature characteristics and source locations, and this mapping shows robustness to additive noise. Further, through comparing the performance with different feature vectors yet the same localisation method, it can be observed that better performance is obtained by using a full ILD and IPD spectrum in the proposed method, while a more accurate result is achieved by using the ILPD vector in PPAM. The ILPD is the composited vector that connects the low-frequency IPD and full ILD as described in [33] This different feature preference proves that the proposed method successfully exploits the non-linear cues in the high-frequency IPD spectrum, and such cues will contribute to a more accurate result.

A more detailed comparison between the azimuth and elevation estimation is provided in Table 4.2. Again, better estimation results can be obtained from the proposed method in general. In Table 4.2(a), the proposed method using full ILD/IPD spectrum vectors achieves the best result and maintains a high estimation accuracy in a severe condition. Meanwhile, the result from the proposed method using ILPD features has comparably good performance in a noise-free environment, yet degrades rapidly with decreased SNR. There are two causes of this phenomenon. First, a longer feature vector increases the diversity of available feature candidates at each splitting node; hence, the independence between decision trees increases and the forest becomes more robust. Second, the RF algorithm extracts the spatial cues at high-frequency IPD spectrum (e.g., the group delay) and those cues contribute to the estimation even in a noisy environment. In Table 4.2(b), although the elevation estimation accuracy from all methods decreases obviously with increased

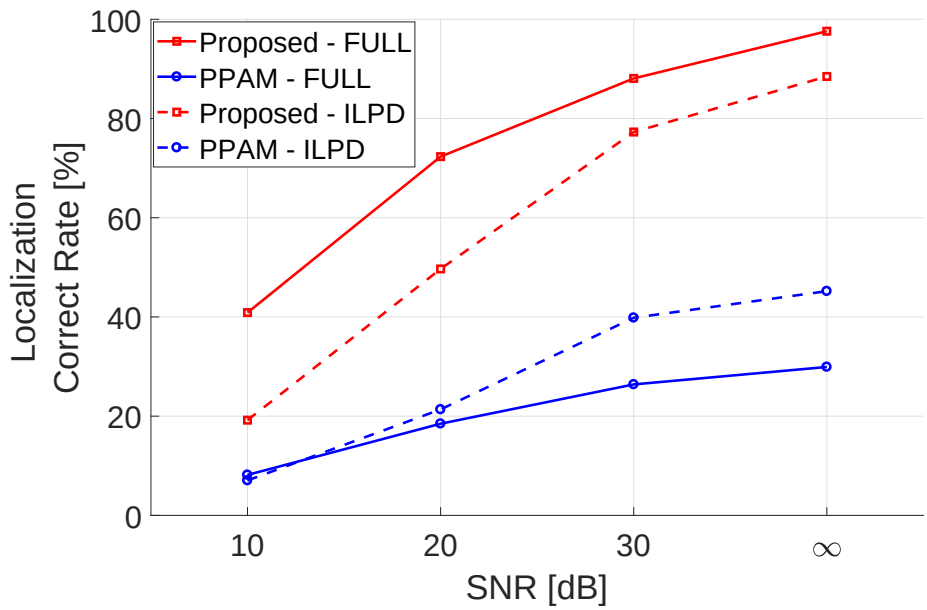


Figure 4.8: Comparing the localisation accuracy between proposed and PPAM methods using different feature vector types. The angular error tolerance is 2.5°.

noise level, the proposed method can provide a precise estimation in the high SNR scenario, compared with the results from PPAM, which indicates that elevation-related feature characteristics extracted by the RF algorithm are non-linear and more sensitive to interference.

Localisation Performance with Reverberations

To localise the sound source in a reverberant environment, the proposed method introduces the DP-RTF estimation method as a pre-processing front-end in the system. In the DP-RTF extraction, because the reverberation condition is assumed unknown, the CTF length Q_k is settled as 0.25 s and remains unchanged in all simulations. Figure 4.9 shows the compared localisation performances with multiple reverberation configurations. Although all methods are degraded in more severe environments, the proposed methods have more accurate localisation results for both feature vector types, and the best performance is achieved by using full length vectors. It can be concluded that the high-IPD features still contain source direction information and, with an appropriate application, can contribute to the source position estimation in a reverberant environment.

Similar to the previous subsection, the azimuth and elevation estimations are compared separately in Table 4.3. In the azimuth comparison, the state-of-the-art method has better performance when using the ILPD feature vector, which

Table 4.2: Azimuth and elevation estimation accuracy comparison in noisy environment

(a) Azimuth accuracy comparison

SNR	No noise		$30dB$		$20dB$		$10dB$	
Tolerance	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$
Proposed - FULL	99.44%	100.0%	98.88%	100.0%	97.20%	99.60%	94.40%	97.92%
PPAM - FULL	79.92%	91.12%	79.52%	90.72%	75.36%	87.6%	61.36%	73.84%
Proposed - ILPD	95.68%	97.12%	87.68%	90.56%	78.96%	86.00%	68.64%	79.44%
PPAM - ILPD	89.28%	96.96%	86.64%	95.68%	73.84%	89.44%	55.04%	73.84%

(b) Elevation accuracy comparison

SNR	No noise		$30dB$		$20dB$		$10dB$	
Tolerance	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$
Proposed - FULL	96.08%	99.52%	89.60%	96.72%	72.64%	83.84%	37.04%	51.20%
PPAM - FULL	28.08%	47.12%	24.48%	44.80%	19.52%	34.64%	9.20%	17.20%
Proposed - ILPD	94.40%	97.12%	76.96%	84.00%	48.72%	58.88%	17.76%	27.04%
PPAM - ILPD	44.72%	71.92%	40.72%	63.28%	24.26%	41.44%	9.84%	18.24%

Table 4.3: Azimuth and elevation estimation accuracy comparison in reverberate environment

(a) Azimuth accuracy comparison

T_{60}	$200ms$		$300ms$		$400ms$		$500ms$	
Tolerance	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$	$\leq 2.5°$	$\leq 5°$
Proposed - FULL	94.32%	97.92%	91.44%	96.72%	89.44%	95.76%	78.88%	89.12%
PPAM - FULL	81.04%	90.64%	79.84%	90.32%	77.84%	88.88%	66.35%	85.76%
Proposed - ILPD	74.24%	84.96%	72.08%	83.28%	66.32%	80.4%	53.04%	74.88%
PPAM - ILPD	86.72%	96.88%	96.40%	77.20%	77.2%	94.96%	58.00%	83.92%

(b) Elevation accuracy comparison

T_{60}	$200ms$		$300ms$		$400ms$		$500ms$	
Tolerance	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$	$\leq 2.5°$	$\leq 6°$
Proposed - FULL	75.84%	89.04%	68.48%	82.72%	55.52%	72.56%	42.64%	62.80%
PPAM - FULL	26.16%	96.88%	23.92%	40.00%	21.12%	35.52%	14.12%	24.80%
Proposed - ILPD	61.92%	42.64%	53.76%	70.32%	42.08%	61.04%	22.48%	40.24%
PPAM - ILPD	37.36%	61.60%	34.08%	55.44%	28.96%	46.80%	15.68%	28.00%

indicates that the high-IPD features are crucial to the proposed method to tackle the interference caused by reverberation. As for the elevation localisation, a considerably

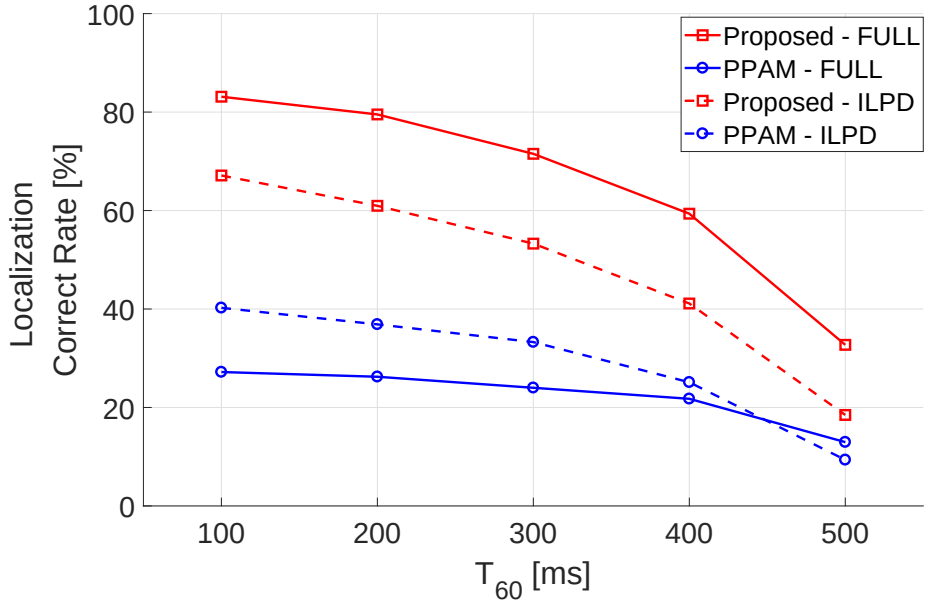


Figure 4.9: Comparing the localisation accuracy between proposed method and PPAM with different T_{60}. The angular error tolerance is 2.5°.

better elevation performance can be obtained by using the proposed methods with both vector types, which proves that mapping of elevation exploited by the RF algorithm can work well against the reverberant interference.

4.7 Conclusion

This chapter has proposed a novel elevation localisation model for binaural localisation based on the RF algorithm. The algorithm operates on the interaural featuressuch as ILD and IPD in the spectral domain. During the training process, the RF algorithm successfully selected the most reliable feature of a spectrum by using information gain, and constructed the elevation-dependent mappings simultaneously based on its selections for each sagittal plane. In the testing, the multi-tree structure showed robustness to additive noise and flexibility to an unfamiliar acoustical environment. Further, a hierarchical localisation system was also proposed for 3-D space localisation, which achieved outstanding localisation performance in the simulation.

The specific contributions of this chapter are as follows:

i Applying MI-based feature selection for interaural features, which proved that localisation performance can be improved with selected features.

ii Providing an importance analysis for interaural features.

iii Adopting a tree-structured model to exploit the feature characteristics for binaural localisation.

iv Proposing a probabilistic mapping model for 3-D binaural localisation, and investigating the tree-based model in binaural localisation.

iv Proposing a robust binaural localisation model for an unfamiliar environment, including additive noise and reverberations.

Chapter 5

Active Binaural Localisation with Head Rotation

Overview: *This chapter explores the probability of using simple ITD and the active movement of listeners' heads to localise single and multiple sound sources in 3-D space. We first present and discuss the ITD behaviour with three different types of head movement: pitch, roll and yaw. It has been shown that rotation around the vertical axis reduces the most direction ambiguity in 3-D space. We adopt the Gaussian process regression model as the mapping model to implement this active localisation model using ITD only. The proposed model shows high tolerance to additive noise and great robustness with reverberation. Further, we combine the rotating ITD feature and gammatone frequency cepstral coefficients (GFCC) and propose a new localisation method that is capable of separating and localising multiple speech sources.*

5.1 Introduction

In Chapter 4, we proposed a method to localise a sound source in a complex environment with the static subject depending on the HRTF-related interaural cues and several pre-processing procedures. We concluded that pinpointing the speaker in a 3-D space is highly related to the spectral characteristics of the high-frequency domain, especially for elevation localisation. However, those characteristics must be summarised by a high-dimension feature vector and a large dataset would be necessary for model training . Moreover, those features could be easily corrupted by interference in a practical environment, such as reverberation and multiple sources. Therefore, a more robust method with the capability of localising numerous sources is desired, and, in this chapter, we propose an active binaural localisation mechanism with head rotation to tackle the challenges mentioned above.

Instead of exploiting the static feature characteristics, the proposed model aims to apply the dynamic cues caused by subject motion, which is inspired by observing the nature of human spatial hearing. Human listeners always adjust their head postures when trying to localise the sound sources in a severe environment, and it is widely accepted that such slight head movement can efficiently resolve ambiguities in sound source localisation [45–47]. Wallach [45] proved that head movements can provide information for both azimuth and elevation estimation, and asserted that movement without displacement of the interaural axis would not contribute to the localisation, which has been proven and confirmed by Perrett [46] and Wightman [47]. In addition, Kneip [120] and [121] Thurlow *et al.* summarised the head rotation into three basic types as *rotation* or *yaw* around the vertical axis, *tip* or *pitch* around the interaural axis and *pivot* or *roll* around the horizontal baseline.

Recently, a few methods attempted to incorporate head movement in binaural localisation. Zhong et al. [48] combined head turn motion data and extended the Kalman filter to reduce the localisation error of azimuth and elevation. Braasch et al. [49] averaged cross-correlation patterns across different head orientations to resolve frontback confusion in anechoic conditions, and Ma et al. [50, 51] adopted head rotation to resolve frontback confusion with various acoustic configurations. However, most existing methods either require a recording while the head is rotating [48] or need a high-dimension feature vector for training [50,51], which may result in complexity of the system during practical use. In this chapter, we describe a concise rotation model with low requirements on received signals for 3-D space localisation in a complex environment. The model only uses the ITD as the spatial cue and only requires the received signal at two static states, before and after rotation. In other words, the system does not record signals while the head is rotating. In addition, we propose a new source separation method based on the rotation model to tackle the multiple-speakers scenario.

This chapter is organised as follows. First, we summarise the relationships between the ITD and three types of head movement [120,121] and evaluate the contributions of the movement patterns in Section 4.2. Next, in Section 5.3, we develop a Gaussian process regression-based model for two-state head rotation localisation. We then endow the model with the ability to localise multiple sources, and propose an online source separation method by jointly using ITD cues and the GFCC of the signals in Section 5.4. Finally, the localisation performance of the proposed model is evaluated based on a simulated dataset with various noise and reverberation levels.

5.2 Fundamentals of Head Movement

5.2.1 Interaural Cues and Spherical Head Model

In this chapter, the vertical-polar (VP) coordinate system is adopted purposing for convenience , since the change of azimuth θ and elevation ϕ in such a system can directly transform the head rotation of *yaw* angle and *pitch* angle. The interaural cues for a spherical head model have previously been reviewed [122]. As reviewed in Figure 5.1(a), the head is approximated as a rigid sphere, with radius a, and the two microphones are placed diametrically across the spherical head on the x-axis. The single point source is located at an infinite distance with an incident angle α, so the sound wave can be considered a plane wave when arriving at the microphones. The incidence angle α is defined by the angle between the vector from the origin to the source position and the positive direction of the x-axis on the surface determined by the point source and microphones (e.g., plane SOR). With an ideal spherical head [122, 123], the generalised ITD can be evaluated by:

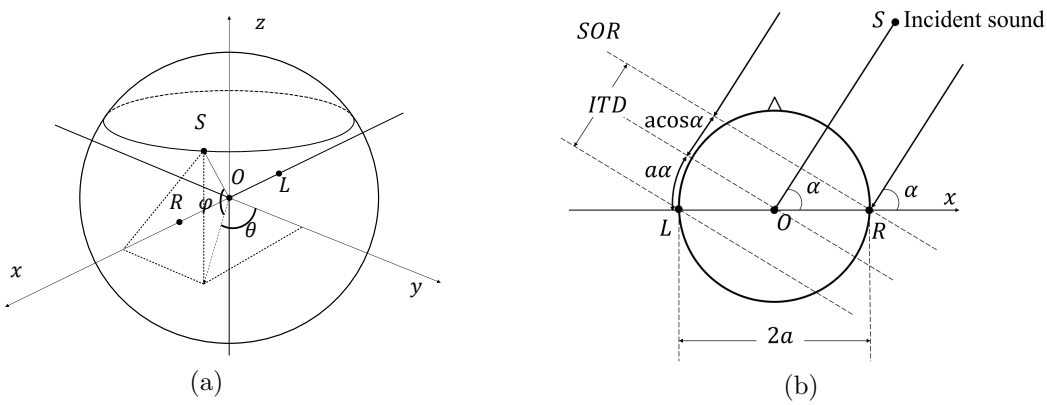


Figure 5.1: Vertical Polar System and far-field sound wave propagation with consideration of head scattering

$$\tau(\alpha) = \frac{a}{c}(\cos\alpha + \alpha) \tag{5.1}$$

Here, the positive and negative ITD value indicates the source located within the right and left semi-sphere, respectively.

The interaural cues in the frequency domain (e.g., the ILD and IPD) are defined

as in the previous chapters:

$$\begin{aligned} ILD(\alpha, f) &= 20\log_{10}\left|\frac{H_l(\alpha, f)}{H_r(\alpha, f)}\right| \\ IPD(\alpha, f) &= \angle\frac{H_l(\alpha, f)}{H_r(\alpha, f)} \end{aligned} \tag{5.2}$$

where $H_l(\alpha, f)$ and $H_r(\alpha, f)$ indicates the left and right HRTF, respectively, which can be approximated by [57]:

$$\begin{aligned} H_l(\alpha, f) &= -\frac{1}{(ka)^2}\sum_{l=0}^{\infty}\frac{(2l+1)j^{l+1}(-1)^l P_l(\cos\alpha)}{dh_l(ka)/d(ka)} \\ H_r(\alpha, f) &= -\frac{1}{(ka)^2}\sum_{l=0}^{\infty}\frac{(2l+1)j^{l+1} P_l(\cos\alpha)}{dh_l(ka)/d(ka)} \end{aligned} \tag{5.3}$$

where $k = 2\pi f/c$ is the wave number, $P_l(\cdot)$ denotes the Legendre polynomial of degree l and $h_l(ka)$ denotes the lth order spherical Hankel function of the second kind.

In (5.1) and (5.2), the only directional parameter for the function of interaural cues is the incidence angle α; thus, the difference in interaural cues is actually caused by the changing of α based on the different source position (θ, ϕ), and the relationship between α and (θ, ϕ) can be found as:

$$\alpha = \cos^{-1}(\cos\phi\sin\theta) \tag{5.4}$$

Now the alternation of (θ, ϕ) resulting from head movements is first transformed to the changing of α, and the corresponding variation in interaural cues can subsequently be discovered. In fact, the nature of active sound source localisation by moving the head orientation is to change α and its corresponding interaural cues, and then, based on the variation of interaural cues, to evaluate the initial sound source direction.

5.2.2 Source Direction Representation in Cartesian Coordinates

Given that the VP coordinate system is always referenced to the subject, a representation of head posture and source location with consistency before and after head movements is necessary. To simplify the derivation process, such representa-

tion is given by a matrix transformation in the Cartesian coordinate system. The transformation from VP coordinates to the Cartesian coordinates is defined as:

$$\begin{cases} \mathrm{x} = \cos(\phi)\sin(\theta) \\ \mathrm{y} = \cos(\phi)\cos(\theta) \\ \mathrm{z} = \sin(\phi) \end{cases} \tag{5.5}$$

and the transformation form Cartesian coordinates to VP coordinates is defined as:

$$\begin{cases} \theta = \sin^{-1}\left(\dfrac{\mathrm{x}}{\sqrt{\mathrm{x}^2+\mathrm{y}^2}}\right) \\ \phi = \sin^{-1}\left(\dfrac{\mathrm{z}}{\sqrt{\mathrm{x}^2+\mathrm{y}^2+\mathrm{z}^2}}\right) \end{cases} \tag{5.6}$$

and the incident angle can be represented as:

$$\alpha = \cos^{-1}\left(\frac{\mathrm{x}}{\sqrt{\mathrm{x}^2+\mathrm{y}^2+\mathrm{z}^2}}\right) \tag{5.7}$$

Now we define the source location vector before head movement as $\boldsymbol{S} = \{\mathrm{x}, \mathrm{y}, \mathrm{z}\}^{\mathrm{T}}$ and the source location vector after head movement as $\boldsymbol{S}_\eta = \{x_\eta, y_\eta, z_\eta\}^T$, where η defines the head movement types. The head movement can be represented as:

$$\boldsymbol{S}_\eta = \boldsymbol{R}_\eta \boldsymbol{S} \tag{5.8}$$

where $\boldsymbol{R}_\eta$ is a 3×3 metric that defines the rotation metrics. The following content analyses the different head movement given by different $\boldsymbol{R}_\eta$, and discusses the related transformation of $\boldsymbol{S}$ and $\boldsymbol{S}_\eta$.

5.2.3 Head Movement Representation and Rotated Interaural Time Delay

In this study, only the pure rotating movements are analysed, which means that one coordinate remains invariant during the head rotating. Thus, there are three types of head movement considered [120, 121]. Correspondingly, let $\eta = [\mathrm{x}, \mathrm{y}, \mathrm{z}]$ define the head rotating around the x-axis, y-axis and z-axis, which can also be re-expressed as pitch, roll and yaw, respectively. The rotation angle is denoted as Δ_η, and the rotating direction is set as anti-clockwise from the view of positive direction of the corresponding axis.

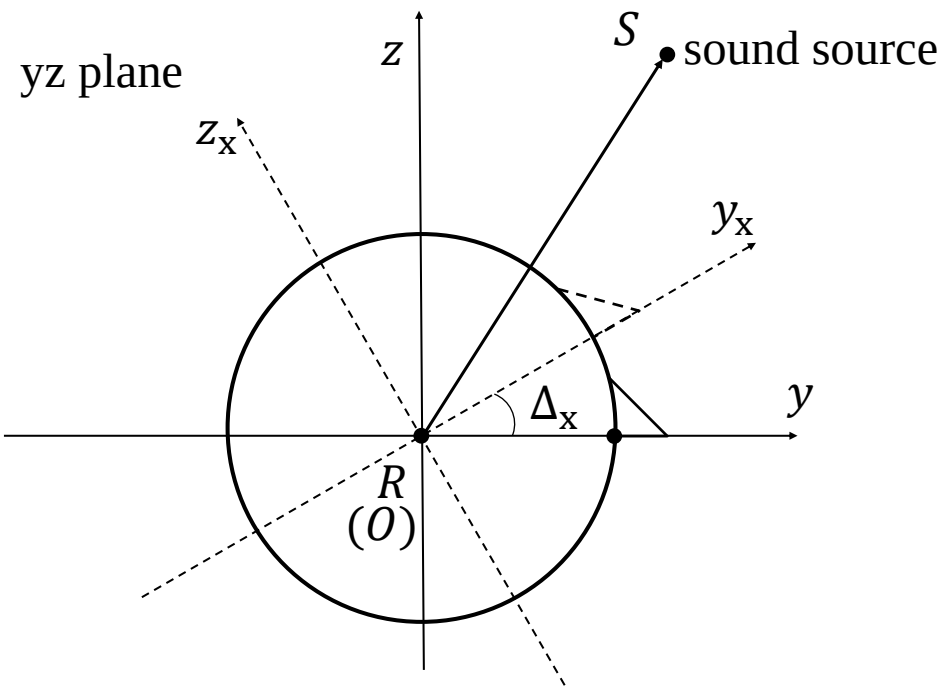


Figure 5.2: Lateral view for head rotation around x-axis

Pitch: Rotation around x-Axis

First, the rotation around the x-axis is considered. Figure 5.2 illustrates this type of rotation on the yz plane. The rotation metric $\boldsymbol{R}_{\mathrm{x}}$ can be found as:

$$\boldsymbol{R}_{\mathrm{x}} = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & \cos\Delta_{\mathrm{x}} & \sin\Delta_{\mathrm{x}} \\ 0 & -\sin\Delta_{\mathrm{x}} & \cos\Delta_{\mathrm{x}} \end{Bmatrix} \tag{5.9}$$

The rotated coordinates for vector $\boldsymbol{S}_{\mathrm{x}}$ can be calculated as:

$$\boldsymbol{S}_{\mathrm{x}} = \begin{Bmatrix} x_{\mathrm{x}} \\ y_{\mathrm{x}} \\ z_{\mathrm{x}} \end{Bmatrix} = \begin{Bmatrix} x \\ y\cos\Delta_{\mathrm{x}} + z\sin\Delta_{\mathrm{x}} \\ -y\sin\Delta_{\mathrm{x}} + z\cos\Delta_{\mathrm{x}} \end{Bmatrix} \tag{5.10}$$

By substitute (5.10) into (5.7), we can obtain that,

$$\alpha_{\mathrm{x}} = \alpha$$

This indicates that the rotation around the x-axis would not lead to any variation in the incident angle; thus, no extra information can be generated by such head move-

ment for the spherical model [45, 120]. Although a pitching movement will result in some minor variation in ILD and IPD because of the changes in pinna filtering for a humanoid subject, a more significant spatial cue would be expected. Therefore, this type of head movement is not an efficient method for source localisation.

Roll: Rotation around y-axis

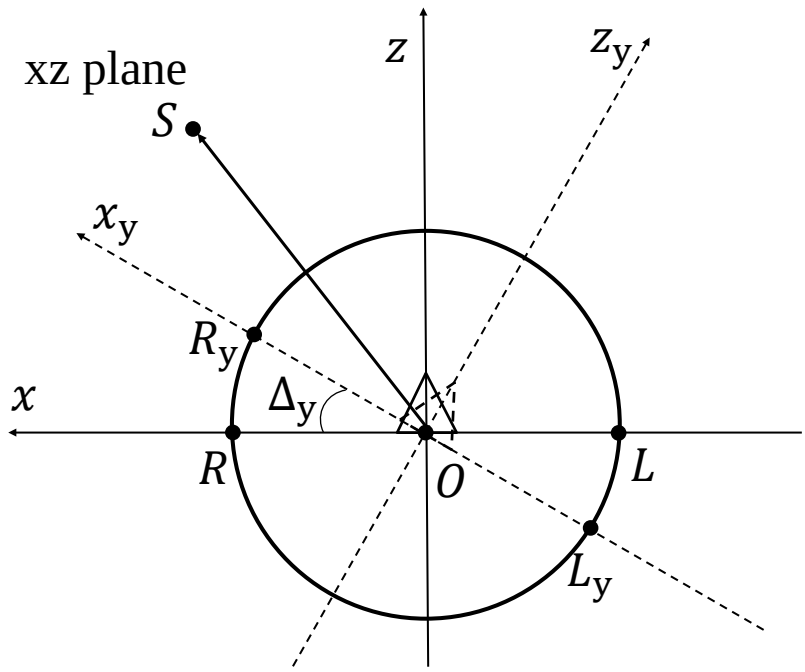


Figure 5.3: Front view for head rotation around y-axis

We then analyse the rotation around the x-axis. Figure 5.5 shows the rolling of the head on the xz plane. The rotation metric $\boldsymbol{R}_{\mathrm{y}}$ can be found as:

$$\boldsymbol{R}_{\mathrm{y}} = \begin{Bmatrix} \cos\Delta_{\mathrm{y}} & 0 & \sin\Delta_{\mathrm{y}} \\ 0 & 1 & 0 \\ -\sin\Delta_{\mathrm{y}} & 0 & \cos\Delta_{\mathrm{y}} \end{Bmatrix} \tag{5.11}$$

The alternate coordinates are obtained by:

$$\boldsymbol{S}_{\mathrm{y}} = \begin{Bmatrix} x_{\mathrm{y}} \\ y_{\mathrm{y}} \\ z_{\mathrm{y}} \end{Bmatrix} = \begin{Bmatrix} x\cos\Delta_{\mathrm{y}} + z\sin\Delta_{\mathrm{y}} \\ y \\ -x\sin\Delta_{\mathrm{y}} + z\cos\Delta_{\mathrm{y}} \end{Bmatrix} \tag{5.12}$$

By substituting (5.11) into (5.7), the incident angle after rotation is:

$$\alpha_{\mathrm{y}} = \cos^{-1}\left(\frac{x\cos\Delta_{\mathrm{y}} + z\sin\Delta_{\mathrm{y}}}{\sqrt{x^2+y^2+z^2}}\right) \tag{5.13}$$

and by replacing x, y and z with θ and ϕ, the following relation can be obtained:

$$\boxed{\begin{aligned}\alpha_{\mathrm{y}} &= \cos^{-1}(\cos\alpha\cos\Delta_{\mathrm{y}} + \sin\phi\sin\Delta_{\mathrm{y}})\\ &= \cos^{-1}(\cos\phi\sin\theta\cos\Delta_{\mathrm{y}} + \sin\phi\sin\Delta_{\mathrm{y}})\end{aligned}} \tag{5.14}$$

From (5.14), it can be observed that the incident angle is modified by the rotation $\boldsymbol{S}_{\mathrm{y}}$, and its value also depends on the initial source position (θ, ϕ). In an equation system of (5.14) and (5.7) with known α, α_{y} and Δ_{y}, the value of ϕ and absolute value of θ are solvable. This indicates that the rotation around the y-axis does provide the solution for the elevation localisation, yet still suffers from frontback confusion, since the sign of θ cannot be determined in the range $[0, \pi]$ and the range $[\pi, 2\pi]$.

Figure 5.4 illustrates the variation of ITD value with increasing Δ_{y}. In each plot, the ITD is defined as positive when the sound wave comes from the left side of the subject and the positive value of Δ indicates that the rotation angle is identical to Figure 5.5. The left column presents the rotation ITD when the source is located in front of the subject, and the right column displays the scenario with the source placed behind the subject. By focusing on one column only, e.g. left column, the elevation can be deduced from the derivation of ITD. However, comparing left and right column, the ITD variation are identical, which indicates that the head rotation around y-axis cannot provides information to distinguish frontal and rear sources. Therefore, this type of rotation still suffers from frontback confusion; thus, further spatial cues are needed to identify the exact location of the source [120]. Therefore, the rotation around the y-axis is also inapplicable for ITD-only 3-D space localisation.

Yaw: Rotation around z-axis

At last, the rotation around z-axis is considered. Fig.5.5 illustrates the binaural microphone system placement with rotation around z-axis. The rotation metric $\boldsymbol{R}_{\mathrm{z}}$ is obtained as:

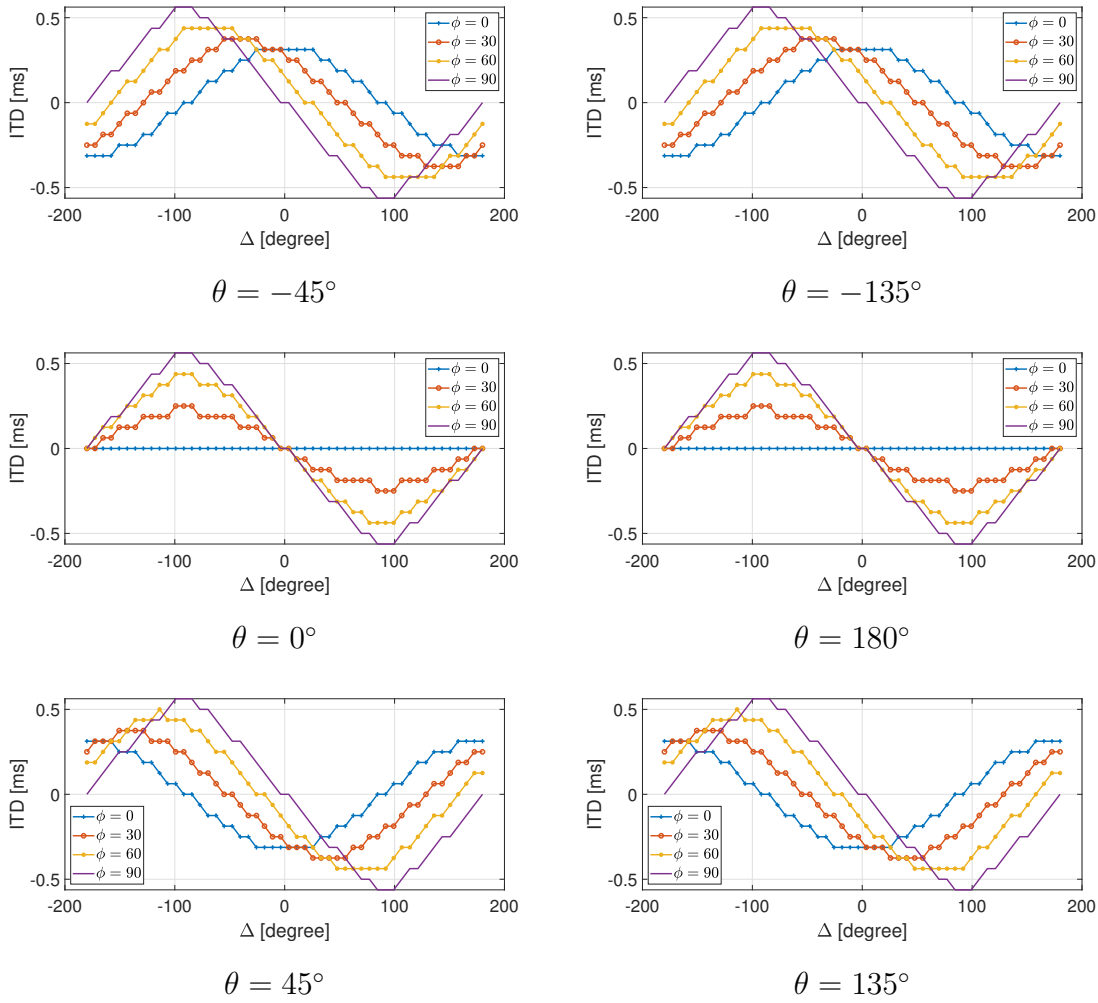


Figure 5.4: Theoretical ITD with rotation varying angle Δ around y-axis. The sound source is located at the upper hemisphere with $\theta \in [-135°, -45°, 0°, 45°, 135°, 180°]$ and $\phi \in [90°, 60°, 30°, 0°]$ in the VP system. The left column indicates the ITD variation when the sound source is located in front of the subject, and the right column indicates the ITD variation when the sound source is located at the back of the subject. The first, second and third rows display the ITD when the sound source is located at the left, middle and right side of the subject, respectively.

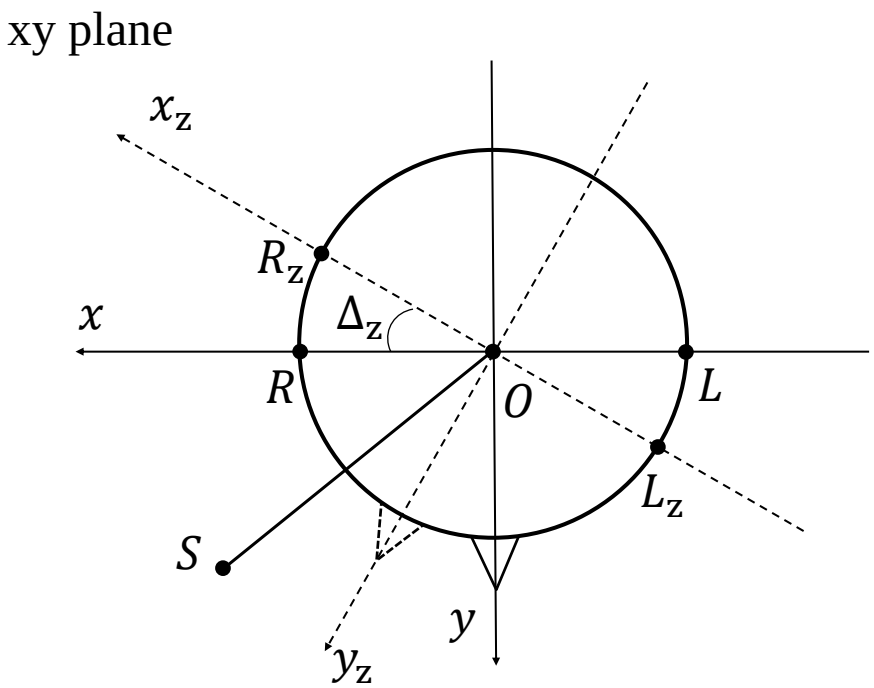


Figure 5.5: Top view for head rotation around z-axis

$$\boldsymbol{R}_{\mathrm{y}} = \begin{Bmatrix} \cos\Delta_{\mathrm{y}} & \sin\Delta_{\mathrm{y}} & 0 \\ -\sin\Delta_{\mathrm{y}} & \cos\Delta_{\mathrm{y}} & 0 \\ 0 & 0 & 1 \end{Bmatrix} \tag{5.15}$$

and the corresponded rotated coordinates is,

$$\boldsymbol{S}_{\mathrm{z}} = \begin{Bmatrix} x_{\mathrm{z}} \\ y_{\mathrm{z}} \\ z_{\mathrm{z}} \end{Bmatrix} = \begin{Bmatrix} x\cos\Delta_{\mathrm{z}} + y\sin\Delta_{\mathrm{z}} \\ -x\sin\Delta_{\mathrm{z}} + y\cos\Delta_{\mathrm{z}} \\ 0 \end{Bmatrix} \tag{5.16}$$

Consequently, the incident angle after rotation is calculated by:

$$\alpha_{\mathrm{z}} = \cos^{-1}\left(\frac{x\cos\Delta_{\mathrm{z}} + y\sin\Delta_{\mathrm{z}}}{\sqrt{x^2+y^2+z^2}}\right) \tag{5.17}$$

and a more elegant expression of α_{z} can be found as,

$$\boxed{\begin{aligned} \alpha_{\mathrm{z}} &= \cos^{-1}(\cos\alpha\cos\Delta_{\mathrm{z}} + \cos\phi\cos\theta\sin\Delta_{\mathrm{z}}) \\ &= \cos^{-1}(\cos\phi\sin(\theta+\Delta_{\mathrm{z}})) \end{aligned}} \tag{5.18}$$

Similar to the rotation around the y-axis, by solving the equation system of (5.18) and (5.7), the value of θ and absolute value of ϕ can be resolved. In this type of movement, the azimuth of the initial source position can be determined, while the sign of the elevation is uncertain, which means that the yawing cannot solve the updown ambiguity. Further, by comparing (5.18) and (5.7), such rotation can be considered as a shift in θ only, which will result in a more directive transformation between α and α_{z}.

Figure 5.6 illustrates the theoretical generalised ITD of a far-field sound source placed on the upper hemisphere. It can be observed that all source positions are distinguishable with a rotation around the z-axis. The curve of ITD variation with head rotation is trigonometric-like in all circumstances. More specifically, the azimuth can be estimated based on the initial phase of the curve, and the elevation can be estimated based on the amplitude of the curve. The source close to the horizontal plane leads to a larger amplitude; however, the source located at the top of the subject minimises the variation of ITD. Further, by comparing the left and right column in Figure 5.6, the frontback confusion can be easily resolved by observing the derivative of the ITD with rotation. With the current definition, the frontal sources have a negative derivation of ITD curve at $\Delta = 0$, while, in contrast, the sources from the back have a positive derivation. Therefore, based on the observations from Figure 5.6, for an arbitrary source on the upper hemisphere, the source location can be deducted with a known rotating angle around the z-axis.

In summary, for the three types of single rotation discussed in this section, the rolling and yawing can provide valuable information for localising a sound source in 3-D space, although each method suffers from a type of ambiguity. The rotation around the z-axis is a more effective method for the following reasons. First, the transformation between the incident angles before and after rotation is simpler and more direct, with the rotation angle only applying on one directional parameter, θ. Second, the updown ambiguity would result less inconvenient in practical applications, compared with the frontback ambiguity. Given that the subject's body will block a large area in a 3-D sphere and the source target that interests human listeners is usually located on the horizontal plane and upper hemi-sphere, the ambiguity of up and down is acceptable. Besides, the floor and shoulder reflection can also be helpful to resolve up-down reflection, though they are not included in this chapter. Finally, considering human head mobility, head yawing has much higher flexibility than does head rolling, which means that a larger range of rotation angle Δ can be adopted in practice. Therefore, the design of the localisation model will be based

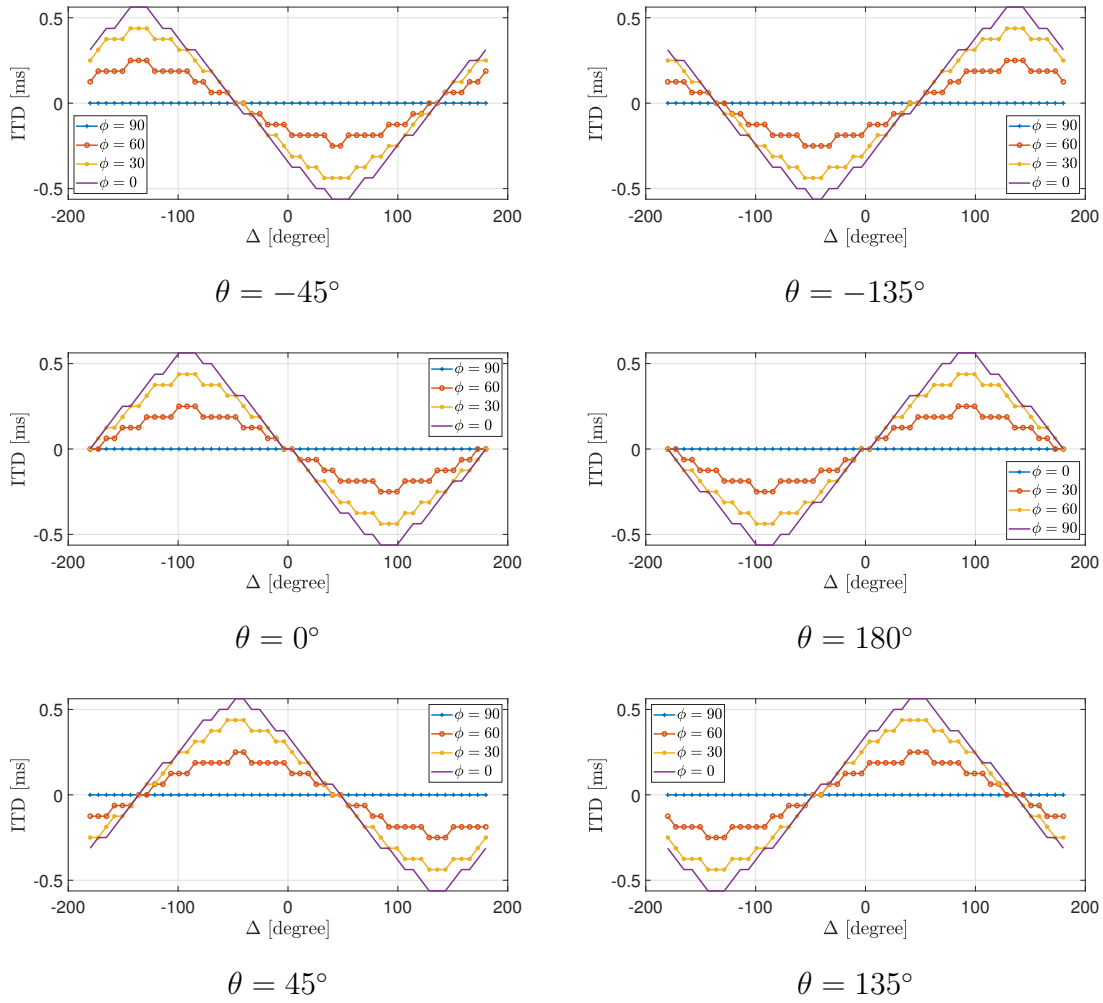


Figure 5.6: Theoretical ITD with rotation varying angle Δ around z-axis. The sound source is located at the upper hemisphere with $\theta \in [-135°, -45°, 0°, 45°, 135°, 180°]$ and $\phi \in [90°, 60°, 30°, 0°]$ in the VP system. The left column indicates the ITD variation when the sound source is located in front of the subject and the right column indicates the ITD variation when the sound source is located at the back of the subject. The first, second and third rows display the ITD when the sound source is located at the left, middle and right side of the subject, respectively.

on the yawing movement in the following content.

5.3 Binaural Localisation with Head Rotation

Based on the discussion in Section 5.2, a binaural localisation system using head rotation cues can be proposed. The system design is presented in Figure 5.7. The received signals before and after rotation are first transformed to the T-F domain by the STFT, and then the ITDs are estimated based on those received signals. If there is only one speaker, a 2-D rotation ITD vector $\vec{\boldsymbol{\tau}}$ will be formed; otherwise, a speaker-separation module will be activated and a rotation ITD metric $\hat{\boldsymbol{\tau}}$ will be generated. Finally, a pre-trained rotation model provides the final estimation of source locations based on $\vec{\boldsymbol{\tau}}$ or $\hat{\boldsymbol{\tau}}$. In this section, a single-speaker scenario is considered and the estimation of ITD and training of the rotation model are discussed. The multiple-source scenario and speaker-separation module will be expanded in Section 5.4.

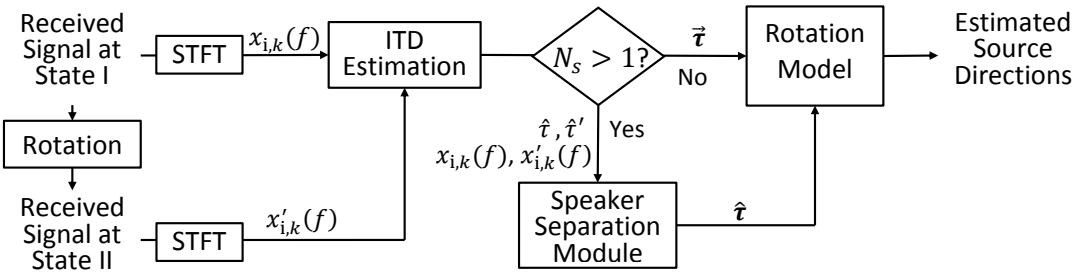


Figure 5.7: Binaural localisation system with head rotation

5.3.1 Interaural Time Difference Estimation

Recall that, in the content introduced in Chapter 2, we adopted the GCC-PHAT method [23] to estimate ITD before and after the head rotation at a corresponding timeframe of k and k', denoted as $\hat{\tau}_k$ and $\hat{\tau}'_{k'}$, respectively. The estimation of $\hat{\tau}_k$ and $\hat{\tau}'_{k'}$ is given by:

$$\begin{aligned} \hat{\tau}_k &= \arg\max_{\tau} \mathcal{F}^{-1}\left(\frac{x_{r,k}(f)x^*_{l,k}(f)}{|x_{r,k}(f)x^*_{l,k}(f)|}\right) \\ \hat{\tau}'_{k'} &= \arg\max_{\tau} \mathcal{F}^{-1}\left(\frac{x'_{r,k'}(f){x'_{l,k'}}^*(f)}{|x'_{r,k'}(f){x'_{l,k'}}^*(f)|}\right) \end{aligned} \tag{5.19}$$

In a single-speaker scenario, the speaker position is assumed static, and the voiced frames should have remarkable directional features while noise frames do not have such features. Therefore, a voting strategy is adopted to obtain the estimated ITDs for the received signals:

$$\begin{aligned} \bar{\tau} &= \arg\max_{\tau} \sum_{k=1}^{K} z(\hat{\tau}_k = \tau) \\ \bar{\tau}' &= \arg\max_{\tau} \sum_{k=1}^{K} z(\hat{\tau}'_{k'} = \tau) \end{aligned} \tag{5.20}$$

where $z(\cdot)$ is the indicator function. Then the rotation ITD vector is formed by concatenating $\bar{\tau}$ and $\bar{\tau}'$ as:

$$\vec{\boldsymbol{\tau}} = [\bar{\tau}, \bar{\tau}'] \tag{5.21}$$

In the following content, a probabilistic model based on $\vec{\boldsymbol{\tau}}$ is proposed.

5.3.2 Two-state Head Rotation Model

We assume that the horizontal and vertical location of a sound source are independent; thus, the probability of source location with given $\vec{\boldsymbol{\tau}}$ can be expressed by a multiplication of the probabilities of azimuth and elevation as:

$$P(\theta, \phi \mid \vec{\boldsymbol{\tau}}) = P(\theta|\vec{\boldsymbol{\tau}})P(\phi|\vec{\boldsymbol{\tau}}) \tag{5.22}$$

Therefore, the models of θ and ϕ can be trained separately from the training dataset. Here, considering the training set $\{\boldsymbol{\tau}^{(n)}, \mathrm{y}^{(n)}\}_{n=1}^{N}$, where $\boldsymbol{\tau} = [1, \vec{\boldsymbol{\tau}}]^{\mathrm{T}}$ and y denote either θ or ϕ. From the observations in Figure 5.6, the ITDs with head rotation can provide discrimination for both θ and ϕ. We use an unknown latent function $f(\boldsymbol{\tau})$ to describe this relationship as:

$$\mathrm{y} = f(\boldsymbol{\tau}) + \varepsilon \tag{5.23}$$

where ε indicates the noise term subject to the zero-mean Gaussian distribution with variance σ^2, denoted as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We then use the Gaussian process regression model and training data to estimate $f(\boldsymbol{\tau})$. In the Gaussian process model, the unknown latent function $f(\boldsymbol{\tau})$ is treated as a collection of random variables whose subset $\{f(\boldsymbol{\tau}^{(1)}), ..., f(\boldsymbol{\tau}^{(N)})\}^{\mathrm{T}}$ with inputs $\boldsymbol{T} = \{\boldsymbol{\tau}^{(1)}, ..., \boldsymbol{\tau}^{(N)}\}^{\mathrm{T}}$ is multivariate

Gaussian denoted as:

$$\{f(\boldsymbol{\tau}^{(1)}), ..., f(\boldsymbol{\tau}^{(N)})\}^{\mathrm{T}} \sim \mathcal{N}(\mu(\boldsymbol{\tau}), \mathcal{K}(\boldsymbol{T}, \boldsymbol{T})) \tag{5.24}$$

where $\mu(\boldsymbol{\tau}) \in \mathbb{R}^{N\times 3}$ defines a prior mean, which can be set as zero purposing without losing generality. $\mathcal{K}(\boldsymbol{T}, \boldsymbol{T}) \in \mathbb{R}^{N\times N}$ indicates the covariance matrix of pairwise covariance function evaluations between inputs $\boldsymbol{\tau} \in \boldsymbol{T}$ as:

$$\mathcal{K}(\boldsymbol{T}, \boldsymbol{T}) = \begin{bmatrix} \kappa(\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(1)}) & \kappa(\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(2)}) & \cdots & \kappa(\boldsymbol{\tau}^{(1)}, \boldsymbol{\tau}^{(n)}) \\ \kappa(\boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(1)}) & \kappa(\boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(2)}) & \cdots & \kappa(\boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(n)}) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\boldsymbol{\tau}^{(n)}, \boldsymbol{\tau}^{(1)}) & \kappa(\boldsymbol{\tau}^{(n)}, \boldsymbol{\tau}^{(2)}) & \cdots & \kappa(\boldsymbol{\tau}^{(n)}, \boldsymbol{\tau}^{(n)}) \end{bmatrix} \tag{5.25}$$

in which $\kappa(\cdot)$ indicates the covariance function or the kernel function. Therefore, the training process can obtain the covariance matrix in (5.21) through the training data and (5.22).

As for the testing dataset $\boldsymbol{T}^{\dagger} = \{\boldsymbol{\tau}^{\dagger(1)}, ..., \boldsymbol{\tau}^{\dagger(N^{\dagger})}\}^{\mathrm{T}}$, tagged by $(\cdot)^{\dagger}$, their corresponding output $\{f(\boldsymbol{\tau}^{\dagger(1)}), ..., f(\boldsymbol{\tau}^{\dagger(N^{\dagger})})\}^{\mathrm{T}}$ is still a multivariate Gaussian as:

$$\{f(\boldsymbol{\tau}^{\dagger(1)}), ..., f(\boldsymbol{\tau}^{\dagger(N^{\dagger})})\}^{\mathrm{T}} \sim \mathcal{N}(\bar{\mathrm{y}}^{\dagger}, \bar{\Sigma}^{\dagger}) \tag{5.26}$$

where the posterior mean $\bar{\mathrm{y}}^{\dagger}$ and covariance functions $\bar{\Sigma}^{\dagger}$ are given by

$$\begin{cases} \bar{\mathrm{y}}^{\dagger} = (\mathcal{K}^{\dagger})^{\mathrm{T}} \hat{\mathcal{K}}^{-1} \mathbf{y} \\ \bar{\Sigma}^{\dagger} = \mathcal{K}^{\dagger\dagger} - (\mathcal{K}^{\dagger})^{\mathrm{T}} \hat{\mathcal{K}}^{-1} \mathcal{K}^{\dagger} \end{cases} \tag{5.27}$$

in which $\hat{\mathcal{K}} = \mathcal{K}(\boldsymbol{T}, \boldsymbol{T}) + \sigma^2 \boldsymbol{I}$ captures the smoothness of y. $\mathcal{K}^{\dagger} = \mathcal{K}(\boldsymbol{T}, \boldsymbol{T}^{\dagger})$ and $\mathcal{K}^{\dagger\dagger} = \mathcal{K}(\boldsymbol{T}^{\dagger}, \boldsymbol{T}^{\dagger})$ are pairwise covariance matrices between training and test data and covariance matrix of test data, respectively. Therefore, with distribution $\mathcal{N}(\bar{\mathrm{y}}^{\dagger}, \bar{\Sigma}^{\dagger})$, the probability density function of a test input can be obtained by:

$$p(\mathrm{y}|\boldsymbol{\tau}^{\dagger}) = \mathcal{N}(\mathrm{y}; \bar{\mathrm{y}}^{\dagger}, \bar{\Sigma}^{\dagger}) \tag{5.28}$$

then by replacing y with θ and ϕ and substituting (5.25) in to (5.19), the source position can therefore be given by

$$(\hat{\theta}, \hat{\phi}) = \arg\max_{\theta,\phi} p(\theta|\boldsymbol{\tau}^{\dagger}) p(\phi|\boldsymbol{\tau}^{\dagger}) \tag{5.29}$$

Note that, in (5.23) and (5.24), the head rotation model has the potential to localise multiple sources when $N_s > 1$. In the next section, the head rotation model is applied for the multiple-source localisation task.

5.4 Multiple Speech Sources Localisation

This section investigates localising multiple speech sources with the proposed head rotation model. The following assumptions are made to describe a multiple-source scenario. We assume that the sources at arbitrary static locations with a known source number N_s on the upper hemisphere are activated simultaneously, and the speech signals are independent, whereas the speakers' voice characteristics should be distinguishable and the W-disjoint orthogonal assumption, which assume that only one speaker is active at one position, can be applied [124]. Intuitively, according to Figure 5.6, it can be observed that if two sources are located different azimuth but same elevation plane, it can be distinguished by ITD, while for different elevation position can be separated by different ITD variation in the head movement.

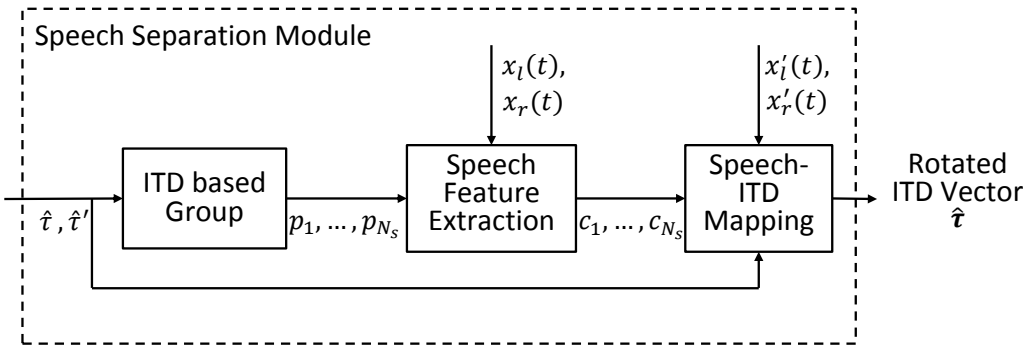


Figure 5.8: Speech Separation Module

Aimed at separating multiple speakers and estimating their positions, the speaker-separation post-processing module is introduced as shown in Figure 5.8. The general idea is to group the T-F frames of received signals based on the known source number N_s and the estimated time delays $\hat{\tau}$ and $\hat{\tau}'$. Then the speech features from T-F frame groups are summarised to match up the corresponding time delays before and after the rotation . The final estimation of source directions is given by the proposed rotation model.

5.4.1 ITD Difference-based Grouping

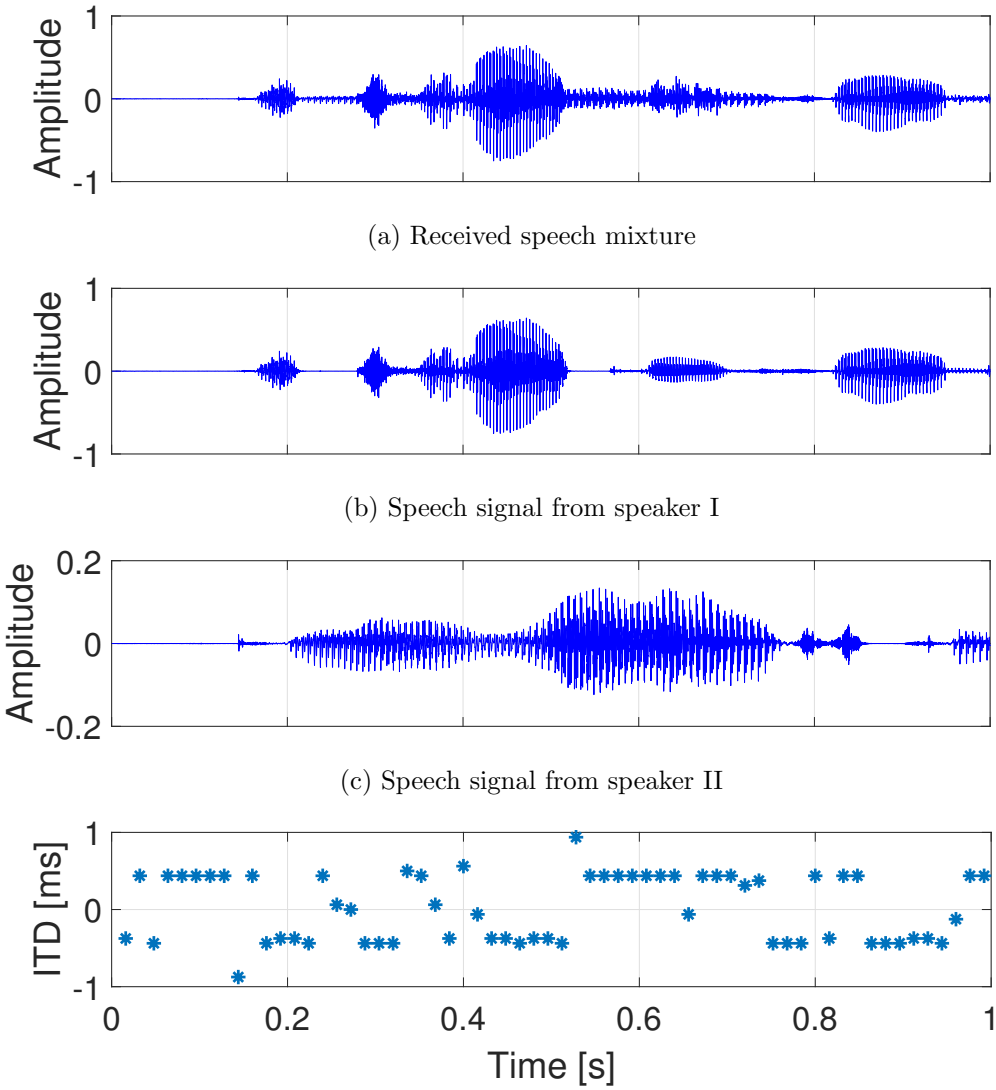
In the W-disjoint orthogonal assumption, one source has dominating energy in some T-F frames and the spatial cues extracted from these T-F frames correspond to the dominant source. Based on this assumption and motivated by [125], the extracted ITD $\hat{\tau}_k$ and $\hat{\tau}'_{k'}$ from (5.19) are used to select and segregate those single-speaker-dominated T-F frames. Figure 5.9(a) displays an example of a two-speaker mixture received by the left ear and the corresponding short-time ITD. Figure 5.9(b) and Figure 5.9(c) present the clean speech signals from speaker I and speaker II, respectively. It can be seen that the individual speech signals exhibit regions with large amplitude over some consecutive frames, which results in a consistent value in short-time ITD. For instance, speech I has dominant energy from ≈ 0.4 s to ≈ 0.5 s, and speech II is more significant from ≈ 0.55 s to ≈ 0.75 s. Meanwhile, the ITD behaviours match those changing dominant signals with a continuous negative value at the frames within 0.4 to 0.5 s, and a positive value from 0.55 s to 0.75 s. Moreover, the ITD behaves unstably and discontinuously at the overlapping area (e.g., 0.3 to 0.4 s). As a result of the variations of the speech signal, in those regions where the speech signals have similar strength, the ratios in (5.19) are not stable and consistency in ITD is not observed. Therefore, we can use the ITD consistency to detect the individual signal-dominated region from the mixed signals and separate the sources via the ITD value.

The consistency of $\hat{\tau}_k$ is checked by a difference function that is defined as:

$$u(\hat{\tau}_k) = \begin{cases} 1 & \hat{\tau}_k - \hat{\tau}_{k-1} = \hat{\tau}_{k+1} - \hat{\tau}_k, \\ 0 & \text{otherwise.} \end{cases} \tag{5.30}$$

Here, we denote $\hat{u}_\tau = u(\hat{\tau}_k)$ as the consistency indicator of ITD samples. Thus, the ITD samples with consistency are selected by $\hat{\tau}_{\tilde{k}} = \hat{\tau}_k|_{\hat{u}_\tau=1}$, where the selected frame index is denoted as $\tilde{k} \in \{1, ..., K\}$. Therefore, the detection and ITD estimation of current active speakers can be jointly obtained with $\tilde{k}$.

Next, we apply the k-nearest neighbours (KNN) algorithm on $\hat{\tau}_{\tilde{k}}$ to obtain the candidate ITD categories and their corresponding classified frame groups. Given that the sources number is known as N_s, the class number in the KNN method is settled as N_s and the classified frame groups are labelled with $\ell = \{1, ..., N_s\}$. Thus, the averaged ITD in each group is denoted as $\bar{\tau}_\ell = \{\bar{\tau}_1, ..., \bar{\tau}_{N_s}\}$ and, for each frame $\tilde{k}$, the corresponding label is obtained as:



(a) Received speech mixture

(b) Speech signal from speaker I

(c) Speech signal from speaker II

(d) ITD of received mixture signal

Figure 5.9: An example of simulated two-speech mixture with short-time ITD analysis. The two speakers are placed at $(-40°, 163.125°)$ and $(45°, 95.625°)$. (a) The wave form for the mixed received signal from the left ear. (b) The clean speech from speaker I. (c) The clean speech from speaker II. (d) Short-time ITD analysis for the speech mixture.

$$\ell_{\tilde{k}} = \arg\min_{\ell} \|\hat{\tau}_{\tilde{k}} - \bar{\tau}_{\ell}\| \tag{5.31}$$

Therefore, each T-F frame $x_{\mathrm{i},\tilde{k}}$ is labelled with $\ell_{\tilde{k}}$, which can be represented by a data pair as $\{x_{\mathrm{i},\tilde{k}}, \ell_{\tilde{k}}\}$. Let $\mathcal{X}_{\mathrm{i}}^{\ell} = \{x_{\mathrm{i},\tilde{k}}, \ell_{\tilde{k}}\}_{\ell_{\tilde{k}}=\ell}$ denote the set of received signal frames with the label ℓ and the frame index $\tilde{k} = \{1, ..., \tilde{K}^{\ell}\}$. Similarly, this grouping approach would be applied on the received signals after rotation. The ITD categories after rotation are denoted as $\bar{\tau}'_{\ell'} = \{\bar{\tau}'_1, ..., \bar{\tau}'_{N_s}\}$ and the corresponding grouped

datasets are denoted as $\mathcal{X}_{\mathrm{i}}^{\ell'} = \{x'_{\mathrm{i},\tilde{k}'}, \ell_{\tilde{k}'}\}_{\ell_{\tilde{k}'}=\ell'}$. Now, to obtain the rotation ITD vector $\boldsymbol{\tau}$ for each speaker correctly, the order of estimated ITDs before and after the rotation must be obtained from the same speaker. This is considered a permutation problem, which is resolved by summarising and pairing the speaker features from $\{\mathcal{X}_{\mathrm{i}}^{\ell}\}_{\ell=1}^{N_s}$ and $\{\mathcal{X}_{\mathrm{i}}^{\ell'}\}_{\ell'=1}^{N_s}$ as follows.

5.4.2 Permutation Based on Speaker Features Matching

To form the correct rotation vector $\boldsymbol{\tau}$, the estimated ITDs before and after rotation need to be classified by the speaker identity and matched up, and the speaker-related monaural features are essential for such classification. Inspired by [126], the GFCC are employed as the discriminative feature. The GFCC have been shown to be a good feature in many evaluations [127–129] and have been widely adopted in many speech separation methods [130–133].

In the speaker feature extraction module shown in Figure 5.8, we use gammatone filterbank as a peripheral processing, and the bandwidths of the gammatone filter bank are set as equivalent rectangular bandwidths. The impulse response of a filter is defined as [126]:

$$g(f_c, t) = t^{(q-1)} e^{-2\pi b(f_c)t} \cos(2\pi f_c t) \tag{5.32}$$

where subscript $c = \{1, ..., C\}$ denotes the channel index of the filter and f_c denotes the central frequency of the filter channel. We use 32 channels in total for each receiver. The filter bandwidth is represented by $b(f_c)$ and the filter order q is set as 4. With this filter bank, the windowed received signal is decomposed into the time-frequency domain. The filtered signal frames are denoted as $\bar{x}_{\mathrm{i}}(k, c)$, and the GFCC feature of the i receiver is obtained by being compressed with a cosine transform as [126]:

$$\mathbf{c}_{\mathrm{i}}(k, j) = \sqrt{\frac{2}{C}} \sum_{c=0}^{C-1} \{\frac{1}{3} \log(\bar{x}_{\mathrm{i}}(k, c)) \cos[\frac{j\pi}{2C}(2c+1)]\} \tag{5.33}$$

where $C = 64$ determines the number of frequency channels and j indicates the coefficient index, which ranges from 0 to 35 based on the suggestion from [134].

To match up the estimated ITD before and after rotation, the GFCC features are extracted and averaged from each ITD class ℓ, and the corresponding GFCC feature vector is defined as:

$$\bar{\mathbf{c}}_{\mathrm{i},j}^{\ell} = \frac{1}{\tilde{K}^{\ell}} \sum_{\tilde{k}=1}^{\tilde{K}^{\ell}} \mathbf{c}_{\mathrm{i}}(\tilde{k}, j) \tag{5.34}$$

Similarly, the average extracted GFCC feature after rotation can be obtained, which is denoted by $\bar{\mathbf{c}}^{\ell'}_{\mathrm{i},j}$. Then, with the extracted $\bar{\mathbf{c}}^{\ell}_{\mathrm{i},j}$ and $\bar{\mathbf{c}}^{\ell'}_{\mathrm{i},j}$, the estimated ITDs $\hat{\tau}_\ell$ and $\hat{\tau}'_{\ell'}$ can be matched. The matched ITD label for ℓ after rotation is obtained by:

$$\hat{\ell} = \arg\min_{\ell'} \sum_{\boldsymbol{i}=\boldsymbol{l},\boldsymbol{r}} \|\bar{\mathbf{c}}^{\ell'}_{\mathrm{i},j} - \bar{\mathbf{c}}^{\ell}_{\mathrm{i},j}\| \tag{5.35}$$

Then the estimated ITD after rotation can be matched to $\hat{\tau}_\ell$ by label index $\hat{\ell}$, and the 2-D rotation feature vector with label ℓ can be formed as:

$$\vec{\boldsymbol{\tau}}_\ell = \left[\hat{\tau}_\ell, \hat{\tau}'_{\hat{\ell}}\right] \tag{5.36}$$

and the rotation feature metric which contain N_s sources is obtained by:

$$\hat{\boldsymbol{\tau}} = \left[\vec{\boldsymbol{\tau}}^{\mathrm{T}}_1, \vec{\boldsymbol{\tau}}^{\mathrm{T}}_2, ..., \vec{\boldsymbol{\tau}}^{\mathrm{T}}_{N_s}\right]^{\mathrm{T}}_{N_S \times 2} \tag{5.37}$$

where $\hat{\boldsymbol{\tau}}$ is a $N_S \times 2$ matrix where each row represents a rotation ITD vector for each active source. Then, by substituting $\hat{\boldsymbol{\tau}}$ into the rotation model, as shown in (5.29), the locations of N_s sources can be obtained.

5.5 Simulation Result

In this section, we test the proposed head rotation localisation model in various configurations with simulated data. The performances of localising a single model in a complex environment are compared with the previous proposed model in Section 4, and the ability to pinpoint multiple sources is also exploited.

5.5.1 Simulation Data and Performance Metrics

Similar to previous chapters, the dataset for training and testing models was generated by convolving simulated BRIR and source signals. The BRIR data were generated with a Roomsim simulator [119] embedded with subject_003 HRIR data [54]. As mentioned above, the simulated sound sources were placed on the upper hemisphere with ϑ varying from $-80°$ to $80°$, and φ varying from $0°$ to $180°$ in the interaural-polar coordinates. The simulated room has dimensions of 5 m $\times$ 5 m $\times$ 3 m, and the subject is placed in the middle of the room with position (2.5 m $\times$ 2.5 m $\times$ 1.5 m). The source signal for training is 1 s length white noise, and for testing is randomly selected from the speech database of the PASCAL CHiME Speech

Separation and Recognition Challenge [106]. The sampling rate for both sources is 16 kHz.

The simulated T_{60} varies from 100 to 600 ms with a fixed reflection ratio from 0.2 to 0.7 for all six walls. The additive white noises with SNR range from 30 to 10 dB are also added to the received signals. In simulating the rotation, each original source signal is segregated into two parts. The first half source signal is convolved with initial directional BRIR and the second half is convolved with alternated BRIR after rotation. Aiming at explore the best localisation performance with human subject mobility limits, the rotation angle Δ is fixed as 45° and the discrete position in the CIPIC database closest to the rotated direction are used as an approximation of rotated BRIR direction. The time consumption of rotating is neglected because it is assumed that the system would not record the signals during the rotating.

We employ two performance metrics for all configurations. We use the average absolute angular error to describe the preciseness of the localisation, which is defined by the angular difference between estimated direction vector and the ground truth direction vector, as defined in the previous chapter. By using such a metric, the confusion of transforming between interaural-polar coordinates and VP coordinates can be bypassed. Another metric used to evaluate the accuracy of the results is the correct rate. A sound source is considered correctly localised if the absolute angular difference between the estimated position and ground truth position is less than a certain threshold. Then the percentage between incidences that are correctly localised and the total number of test incidences are defined as the correct rate. The selection of correctness threshold is based on the ITD estimation resolution, which is constrained by the sampling frequency. With a sampling frequency of 16 kHz, the resolution of ITD in Figure 5.4 and Figure 5.6 is 0.0625 ms, which corresponds to a difference of about 10° of α based on (5.1). Therefore, the preciseness of the ITD model is 10° and the correctness tolerance threshold is set as $\pm 10°$.

5.5.2 Single-source Localisation

We first test the proposed method with the single-source dataset used in Chapter 4. Table 5.1 shows the average error on a different noise reverberation level. The results suggest that the proposed method is insensitive to non-directional noise. In a low-reverberant environment (e.g., $T_{60} < 400$ ms), the average errors have small variation with varying SNR, which proves that the ITD provides a robust spatial cue for localisation and the rotation model can successfully resolve the cone-of-

confusion problem by exploiting the rotation-caused ITD differences. However, the error rises with increasing reverberation, which indicates that the ITD-only-based rotation model may be misdirected by the directional interference, since the system introduced in Section 5.3 did not equip any de-reverberation procedure. In particular, when $T_{60} = 500$ ms, the localisation error is even higher at a high SNR scenario. This is because, in a high-reverberant and low-noise environment, the reflected signal waves have similar consistency as the direct path signal, while the speaker feature-based separation procedures have little effect on those reverberations, since they all come from the same speaker. Introducing the de-reverberant pre-processing could be a possible resolution to tackling such a high-reverberant environment.

T_{60}	**SNR**		
	10 dB	**20 dB**	**30 dB**
No Reverberation	5.27°	5.17°	5.12°
200 ms	5.4°	5.3°	5.2°
300 ms	5.6°	5.8°	5.5°
400 ms	6.1°	5.8°	5.8°
500 ms	7.6°	8.8°	15.1°

Table 5.1: Single source localisation error.

The performance evaluations based on a correct rate with different noise and reverberation levels are presented in Figure 5.10 and Figure 5.11, respectively. Figure 5.10 shows that the ITD rotation model has remarkable accuracy and robustness to additive noise. The high correctness level proves that the rotation could provide sufficient spatial information for the sound source localising on the upper hemisphere, and, thanks to the insensitivity of ITD to the noise, the localisation performances hardly change even in a severe noisy configuration.

In Figure 5.11, although the performances of both methods are affected by the increasing reverberation level, the rotation model outperforms the RF method. Given that the rotation is dynamically introducing more spatial information, especially for solving the cone-of-confusion problem, the rotation model can provide a more accurate elevation estimation. Table 5.2 demonstrates the averaged elevation estimation error in the interaural-polar coordinates, which is actually the localisation performance on a confusion cone. It is clear that the rotation model always generates a smaller error, which indicates that the difference of ITD before and after rotation can offer a more reliable cue for the elevation ambiguity.

In summary, the proposed rotation model can give a robust direction estimation for the single active source in the presence of noise, and the cone-of-confusion

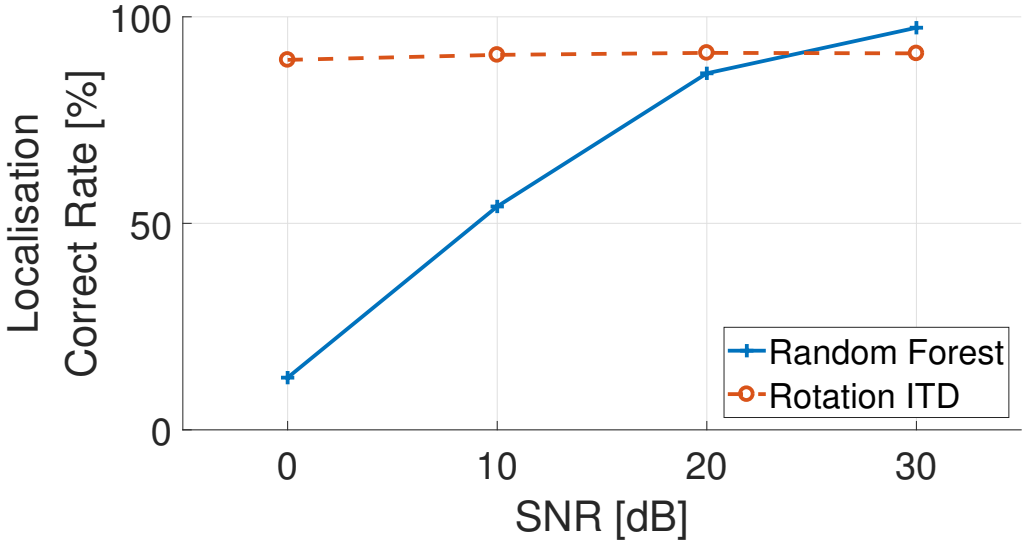


Figure 5.10: Localisation correct rate comparison with different noise levels in an anechoic chamber with ±10° threshold.

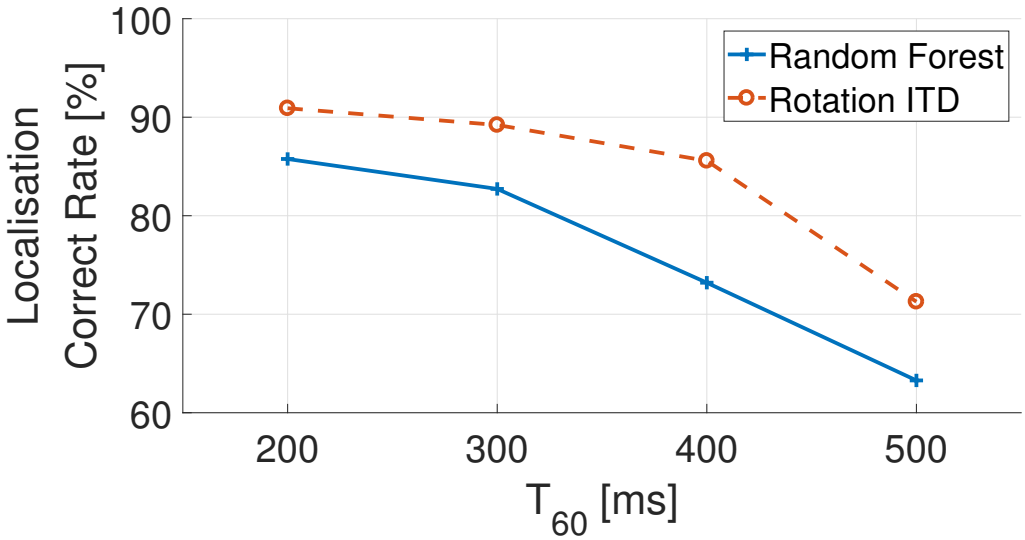


Figure 5.11: Localisation correct rate comparison with different reverberation levels with ±10° threshold. The noise level is fixed as 30 dB SNR.

Methods	T_{60}			
	200 ms	**300 ms**	**400 ms**	**500 ms**
Random Forest	12.2°	14.5°	22.0°	26.0°
Rotation ITD	7.1°	8.0°	9.6°	18.4°

Table 5.2: Average angular error comparison between Random Forest model and proposed rotation ITD model.

problem is efficiently resolved by exploiting the ITD information with a two-state rotation.

5.5.3 Multiple Sources Localisation

We then investigate the ability of the rotation model to localise multiple sources. More specifically, a dataset of a two-source signal mixture is tested with multiple noise configurations. The received signal mixture is simulated by linearly adding two pieces of single-speaker test data in the time domain, and the source locations are randomly selected from the single-source database without overlapping. For each noise level, 500 mixtures are generated and tested.

As for the performance evaluation, the localisation results are still evaluated by the average angular error and correct rate, while each signal mixture is considered as two localisation incidents. Therefore, the average angular error and the correct rate are calculated on 1,000 incidents. To characterise those unsuccessful localisations, another metricnamed the one-source correct rateis introduced to describe the results in which at least one source position is successfully localised in each mixture.

Metrics	*SNR*				
	∞	**30 dB**	**20 dB**	**10 dB**	**0 dB**
Average Angular Error	11.9°	12.2°	13.0°	19.0°	30.0°
Correct Rate	79.60%	81.60%	79.90%	71.40%	56.00%
One-source Correct Rate	93.60%	94.40%	96.60%	95.60%	90.00%

Table 5.3: Localisation performance of ITD rotation model for two-sources localisation.

The performances shown in Table 5.3 suggest that the proposed method is capable of successfully localising most sources in a scenario with two active speakers. A notable degrading of results can only be observed in a very noisy environment, in which case a noise-reduction method would be expected. These results justify that the proposed speech separation module can efficiently separate the source signals, even without prior knowledge of the speakers. Further, the one-source correct rate indicates that, in most cases, at least one source can be correctly detected and localised, which proves that the separation module is capable of exploiting the surviving spatial information from a strong noise effect.

5.6 Conclusion

This chapter has investigated the effect of head rotation in binaural localisation. It has discussed three types of head rotation and their consequential effects on the ITD, and indicated that the ITD variation caused by a single determined rotation around the interaural axis can provide sufficient spatial information for both azimuth and elevation localisation. We then proposed a statistical model using the ITD cues with the rotation for the 3-D space binaural localisation. Further, we explored the capability to localise multiple sources, and proposed a multi-source separation method using ITD difference and speech features. The proposed method was tested with simulated data in various environment configurations, and the rotation model showed remarkable robustness and accuracy of localisation in all testing configurations. Finally, the model was tested with a multi-source scenario, and the model demonstrated capability to separate and localise the speakers with a pair of mixed received signals.

The specific contributions of this chapter are as follows:

i Investigating the effect of yawing, pitching and rolling on the ITD in binaural localisation with a spherical model, which proved that yawing movement can efficiently resolve frontback confusion and cone of confusion. This enabled extension of the existing horizontal plane localisation methods to a 3-D space localisation.

ii Proposing a Gaussian process regression-based model with outstanding vertical plane localisation ability based on active rotation. This model is very accessible in practice because it only requires estimating ITDs at two ITDs when the head is stationary before and after rotation. With such a system design, the existing ITD estimation methods can be incorporated with the proposed rotation model with minor modification.

iii Proposing an online speech separation method based on the two-state ITD differences and the speech feature characteristics, which proved that the ITD can be applied as a reliable speech detector in a multiple-source environment.

iv Using simulation evaluations to demonstrate the remarkable robustness of 3-D space localisation in the presence of noise and reverberation without any pre-processing procedure. This suggests that the dynamic the spatial cues

provided by the head rotation or microphone movement should be considered and applied in the real-world localisation tasks.

Chapter 6

Localisation Experiments in Practical Environments

Overview: *This chapter explores the probability of using simple ITD and the active movement of listeners' heads to localise single and multiple sound sources in 3-D space. We first present and discuss the ITD behaviour with three different types of head movement: pitch, roll and yaw. It has been shown that the rotation around the vertical axis reduces most of the direction ambiguity in 3-D space. We adopt the Gaussian process regression model as the mapping model to implement this active localisation model using ITD only. The proposed model shows high tolerance to additive noise and great robustness with reverberation. Further, we combine the rotating ITD feature and GFCC and propose a new localisation method with the capability of separating and localising multiple speech sources.*

6.1 Introduction

As presented in Chapter 4 and Chapter 5, the proposed methods can localise sound sources accurately and robustly with simulated data. However, testing the proposed methods in practical environments remains necessary because the simulated data are generated based on a mathematical model with several assumptions, which might be difficult to achieve in a practical environment. Therefore, in this chapter, the proposed localisation methods are tested with data recorded in a real indoor environment with the existence of reverberation and background noise.

The experimental evaluation of the localisation performance is conducted in an enclosed laboratory at the Australian National University. A GRAS KEMAR manikin Type 45BA human-like simulator is used to collect the binaural testing data, which is performed by playing a source signal via loudspeakers fixed at dif-

ferent directional positions. Two Type 40AG polarised pressure microphones are placed at the entrance of the ear canal of the dummy head. To test the robustness and practicality of the proposed method, the testing dataset is collected in a real laboratory environment, while the training data are synthesised as described in the previous chapters based on the open HRTF database (i.e., CIPIC database) [54]. Hence, the testing environment is completely unfamiliar to the localisation model, and no prior knowledge about the room acoustical characteristic is learnt during the training process.

The remainder of this chapter is structured as follows. In Section 6.2, we present the details of our experimental equipment and explain the experiment procedures. Section 6.3 shows the preparation of the testing data, including the tested locations and calibration of the raw signals. Finally, the localisation performance is demonstrated in Section 6.4, in which both passive and active systems are tested, and the possible errors of the results are discussed.

6.2 Experiment Facility and Room Configurations

The experiment system can be generally separated into two independent subsystems: the audio playback system and the binaural signal recording system. The general structure and connection relationship of the systems is shown in Figure 6.1, and the setup of the experiment is shown in Figure 6.2. In this system, the source signal is played by the audio playback system via spatially placed loudspeakers. The emitted signal then propagates through the space and is captured by the recording system. The recorded signals are pre-processed and delivered to another computer for the localisation. The later content includes the more detailed configurations of the two systems.

In the audio playback system, 22 loudspeakers are placed on a regular dodecahedron. The loudspeakers are fixed on the middle of the edges and face the centre of the dodecahedron, as shown in Figure 6.2. Figure 6.3 demonstrates the ground truth loudspeaker array placement positions and their corresponding labels in the Cartesian coordinate system. All loudspeakers are driven by the Dante audio network system, as shown in Figure 6.4(a), which is controlled by a computer. As for the recording system, the signals are captured by two Type 40AG polarised pressure microphones planted in the ear canals of the dummy head. The captured signals

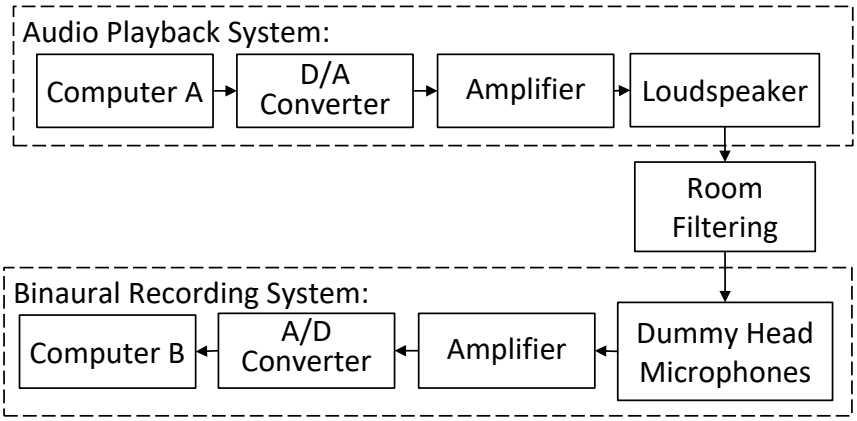


Figure 6.1: The experiment outline: The audio playback system and binaural recording system

are then converted via the U-PHORIA UMC202HD audio interface and transferred to another computer via a USB connection.

In the experiment, we first switch on the recording system, and then play the source signal with one of 22 loudspeakers with selected positions. The signals captured by the recording system are converted and delivered to the localisation systems, and the estimated source location can be obtained. In such a system setup, the generation of directional audio signal and the process of localisation are physically isolated, so the performance of the localisation entirely relies on the recorded signals, and no prior knowledge about the source signal and environmental acoustic configurations can be obtained by the localisation system. Therefore, the robustness and flexibility of the proposed methods can be examined.

6.3 Testing Positions and Microphone Data Preprocessing

As aforementioned, the source signal is played by an array of 20 loudspeakers, so the ground truth source positions are identical to the locations of the loudspeakers. Figure 6.3 and Table 6.1 illustrate the ground truth positions of loudspeakers represented in the interaural-polar coordinate system with the label indices. The testing loudspeaker positions are not presented in the training positions. Thus, a correct estimation should be the nearest location label to the ground truth. Notably, there are seven loudspeakers placed beyond the training region as shown in Figure 6.2

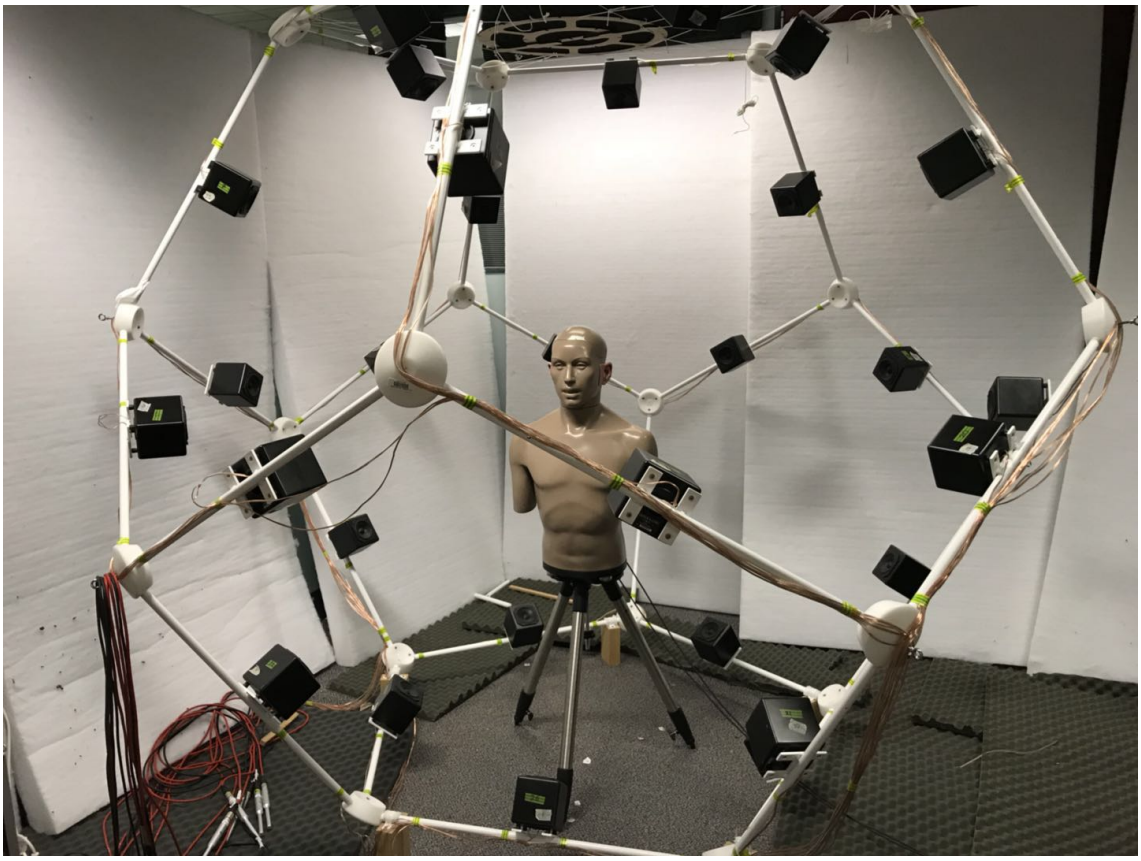


Figure 6.2: The hardware set-up. The loudspeakers are positioned on the middle of edges of a dodecahedron frame, and the dummy head simulator with two microphones are placed in the centre of the speaker arrays.

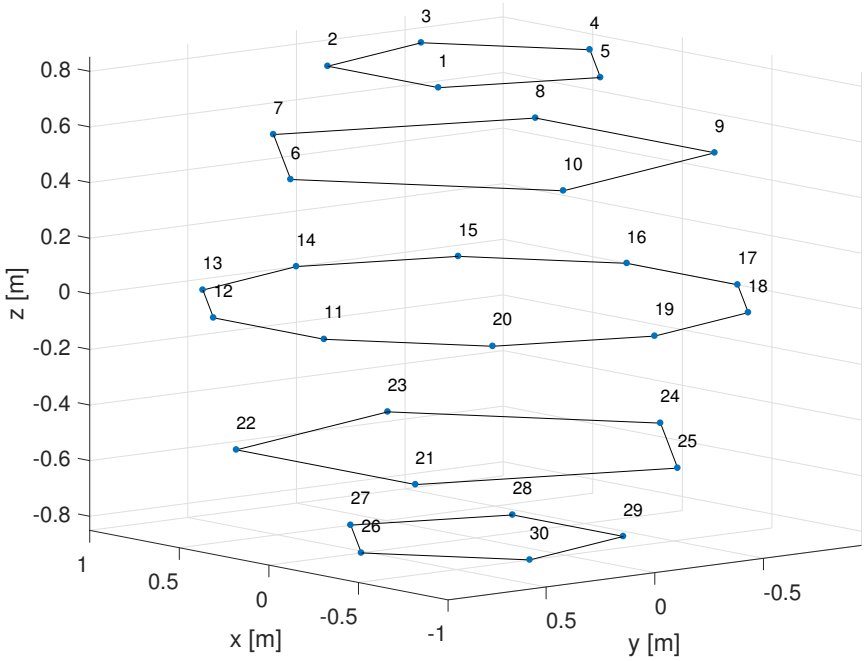


Figure 6.3: The ground truth position of sound sources and their corresponding labels

and marked from 23 to 30 in Figure 6.3. Those loudspeakers are muted during the following tests, since the localisation of these speakers is meaningless.

One issue that must be noted is that the impulses of the two microphone channels cannot be identical in the real world, and small differences lead to inaccuracy in localisation. Therefore, in the binaural recording system, the raw signals captured by the binaural microphones are equalised before passing to the localisation model, which aims at minimising the amplitude difference between channels. The equalisation is performed by multiplying a generic compensation coefficient ζ_l on the left-ear channel, and this coefficient is pre-measured in two steps. First, one 2 s white noise signal is played as the source signal by each loudspeaker, and the captured signals are notated as $x_{1,l,k}(t,\Theta)$ and $x_{1,r,k}(t,\Theta)$. Second, the connection of two channels is switched and the same source signals are played by each speaker again. The signals captured by the switched channel are notated as $x_{2,l,k}(t,\Theta)$ and $x_{2,r,k}(t,\Theta)$. Ideally, if the two channels are identical, we would have $x_{1,l,k}(t,\Theta) = x_{2,r,k}(t,\Theta)$ and $x_{1,r,k}(t,\Theta) = x_{2,l,k}(t,\Theta)$. Therefore, the estimated left-ear compensation coefficient $\hat{\zeta}_l$ can be estimated by averaging all frames and positions as:

$$\hat{\zeta}_l = \sum_{k,\Theta} \left[\frac{x_{2,r,k}(t,\Theta)}{x_{1,l,k}(t,\Theta)} + \frac{x_{1,r,k}(t,\Theta)}{x_{2,l,k}(t,\Theta)} \right] \tag{6.1}$$

(a) Dante audio network system

(b) U-PHORIA UMC202HD audio interface

Figure 6.4: System hardware

and hence, the binaural signals for localisation are calibrated by,

$$\begin{aligned} x_l(t,\Theta) &= \hat{\zeta}_l \cdot x_{1,l,k}(t,\Theta) \\ x_l(r,\Theta) &= x_{1,r,k}(t,\Theta) \end{aligned} \tag{6.2}$$

Given that the calibration is applied on the left channel for all source positions in the time domain, it is spatial and frequency independent, and does not introduce external information to estimate the sound source position.

6.4 Experiment Result

The experiment includes both passive localisation and active localisation. Aimed at testing the generalisation ability of the proposed models, especially the localisation performance in a unfamiliar environment, the models are trained with simulated data, as presented Chapter 4 and Chapter 5. The white noise signal is used as additive noise in the training data and the SNR is fixed as 20 dB. Given that the model is trained with simulated data, the testing room can be considered an unfamiliar environment, and the model can only use the learnt localisation cues of the dummy head.

6.4.1 Passive Localisation

The localisation performance of the passive model introduced in Chapter 4 is demonstrated in Table 6.1. The localisation tests are repeated 10 times, and, in each test, the random selected speech utterances (around 1 to 2 s per utterance) are played as the source signal through the first 22 speakers. The results justify that the model can effectively localise the source positions. The azimuth localisation is very accurate, since the model returns the nearest azimuth class to the ground truth position. Although the accuracy of elevation localisation is degraded compared with the simulation results in the previous chapters, the absolute angular error remains around 10 degrees, which corresponds to around 17 cm differences with a 1 m source distance. Several factors may cause the decreasing of localisation accuracy. One obvious cause is the difference of HRTF between the CIPIC database and the used dummy head. Those differences could result from assembling the dummy head in practice, which would affect the features in a high-frequency region. This issue could be resolved by replacing the training data with measured HRTFs of the subject being used. In addition, the reflections from objects in the laboratory may interfere with the

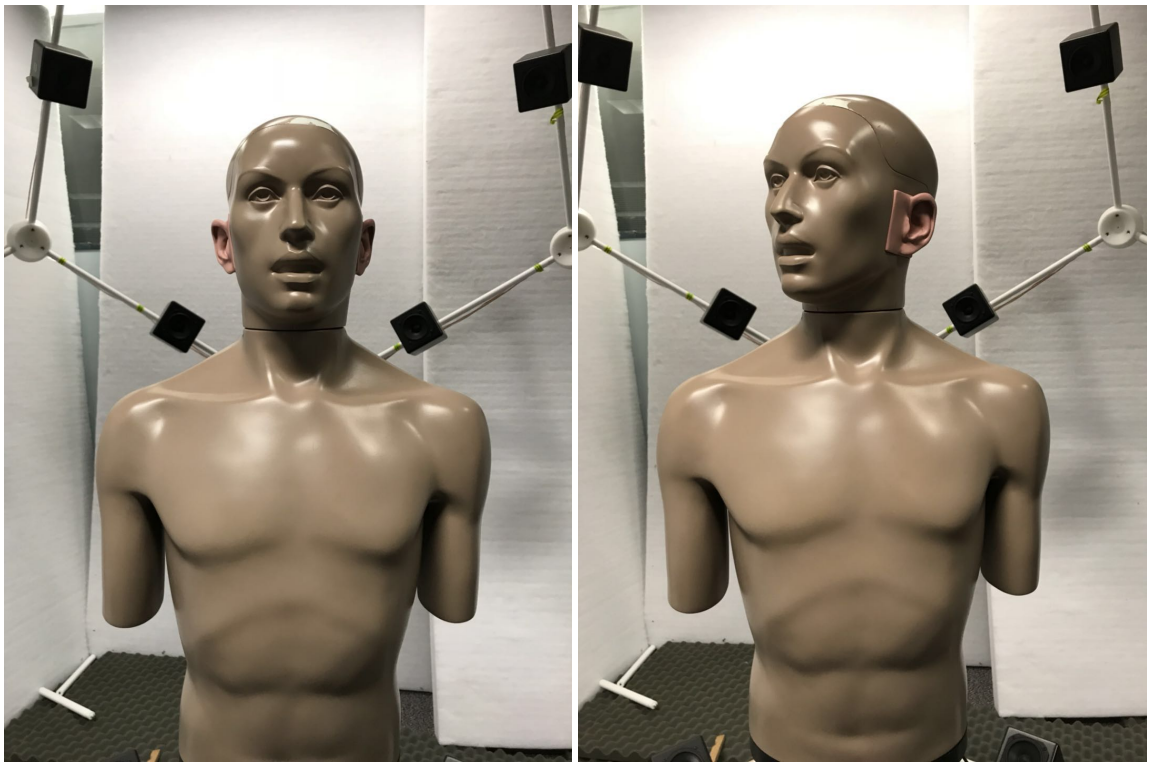
localisation, since there are no de-reverberation operations during testing.

Loudspeaker No.	True Azimuth	True Elevation	Estimated Azimuth	Estimated Elevation	Estimated Error
1	**-18.0°**	**63.4°**	**-22.0°**	**81.7°**	**35.3°**
2	**18.0°**	**63.4°**	**20.0°**	**67.5°**	**4.3°**
3	**30.0°**	**100.8°**	**30.0°**	**71.4°**	**31.5°**
4	**0.0°**	**121.7°**	**-5.0°**	**123.8°**	**2.0°**
5	**-30.0°**	**100.8°**	**-30.0°**	**84.4°**	**14.0°**
6	**0.0°**	**31.7°**	**0.0°**	**33.8°**	**2.0°**
7	**54.0°**	**63.4°**	**55.0°**	**60.2°**	**3.6°**
8	**30.0°**	**142.6°**	**30.0°**	**140.6°**	**1.7°**
9	**-30.0°**	**142.6°**	**-30.0°**	**135.6°**	**6.1°**
10	**-54.0°**	**63.4°**	**-55.0°**	**59.1°**	**2.8°**
11	**-18.0°**	**0.0°**	**-20.0°**	**0.0°**	**2.0°**
12	**18.0°**	**0.0°**	**19.5°**	**10.7°**	**11.9°**
13	**54.0°**	**0.0°**	**55.0°**	**0.0°**	**1.5°**
14	**90.0°**	**0.0°**	**80.0°**	**-39.9°**	**10.0°**
15	**54.0°**	**180.0°**	**55.0°**	**194.1°**	**9.6°**
16	**18.0°**	**180.0°**	**20.0°**	**177.8°**	**3.5°**
17	**-18.0°**	**180.0°**	**-18.0°**	**157.5°**	**24.4°**
18	**-54.0°**	**180.0°**	**-55.0°**	**182.8°**	**2.2°**
19	**-90.0°**	**0.0°**	**-80.0°**	**-39.9°**	**10.0°**
20	**-54.0°**	**0.0°**	**-55.0°**	**1.7°**	**2.7°**
21	**-30.0°**	**-37.4°**	**-30.0°**	**-38.8°**	**1.9°**
22	**30.0°**	**-37.4°**	**30.0°**	**-33.8°**	**3.1°**

Table 6.1: Localisation performance of the passive model. The ground truth loudspeaker locations are represented in $\mathbf{\Theta}_{\mathrm{IP}}$.

6.4.2 Active Localisation

The performance of the active model with head rotation is also tested with the same experiment configurations. The rotation is implemented by manually rotating the subject's head only as shown in Figure 6.5. The experiment reflects one of the advantages of the proposed rotation model using ITD only again. Obviously, the reflection pathways from the subject body to the ears, as well as the corresponding interaural features, have been changed with the rotation of the listener's head, as shown in Figure 6.5, while the ITDs remain unchanged, since the head diameter is static. Therefore, the models trained based on ITDs extracted from the HRTFs without head rotation are still valid.



State I: head posture before rotation. State II: head rotated with $\Delta = 45°$

Figure 6.5: Two States of head rotation model.

The active localisation results are listed in Table 6.2. According to the results, no frontback azimuth or elevation confusion is observed using the proposed active method, which is thanks to the frontback asymmetry ITD behaviour while rotating the head.

The posted results are compared with the baseline localisation method, as mentioned in Chapter 4. Table 6.3 compares the localisation performances between the proposed passive and active method with the PPAM model, which can be considered a passive localisation method [33]. The configurations of the three methods are set the same as described in previous chapters, and the results are compared through the average angular error and the correctness percentage with different tolerance error.

In the comparison, the proposed passive method has a five degree improvement in the average angular error, which means that the proposed method can provide a more accurate localisation. As for the correct localisation percentage, most localisation errors are smaller than five degrees, and less than 10% tests have errors that greater than 20 degrees . This suggests that the proposed method also has a more precise localisation result than does the PPAM method. In general, the proposed

Loudspeaker No.	True Azimuth	True Elevation	Estimated Azimuth	Estimated Elevation	Estimated Error
1	**-18.0°**	**63.4°**	**-16.9°**	**55.7°**	**7.4°**
2	**18.0°**	**63.4°**	**15.5°**	**65.2°**	**3.0°**
3	**30.0°**	**100.8°**	**27.9°**	**97.1°**	**3.8°**
4	**0.0°**	**121.7°**	**2.8°**	**138.8°**	**15.3°**
5	**-30.0°**	**100.8°**	**-35.1°**	**99.0°**	**5.3°**
6	**0.0°**	**31.7°**	**2.6°**	**34.9°**	**4.1°**
7	**54.0°**	**63.4°**	**55.7°**	**45.3°**	**10.6°**
8	**30.0°**	**142.6°**	**27.7°**	**132.0°**	**9.6°**
9	**-30.0°**	**142.6°**	**-30.9°**	**143.2°**	**0.5°**
10	**-54.0°**	**63.4°**	**-54.8°**	**63.7°**	**0.8°**
11	**-18.0°**	**0.0°**	**-17.7°**	**11.6°**	**11.2°**
12	**18.0°**	**0.0°**	**15.0°**	**1.1°**	**3.1°**
13	**54.0°**	**0.0°**	**55.1°**	**4.3°**	**2.8°**
14	**90.0°**	**0.0°**	**81.0°**	**36.4°**	**9.0°**
15	**54.0°**	**180.0°**	**54.4°**	**167.9°**	**7.1°**
16	**18.0°**	**180.0°**	**15.2°**	**171.9°**	**8.3°**
17	**-18.0°**	**180.0°**	**-16.0°**	**148.4°**	**30.3°**
18	**-54.0°**	**180.0°**	**-56.9°**	**172.6°**	**5.1°**
19	**-90.0°**	**0.0°**	**-80.0°**	**58.0°**	**10.0°**
20	**-54.0°**	**0.0°**	**-53.7°**	**26.9°**	**15.8°**

Table 6.2: Localisation performance of the active model.

Methods	**Angular Error**	$\leq 5°$	$\leq 10°$	$\leq 15°$
PPAM	$13.97 \pm 10.80°$	17.73%	48.64%	84.09%
Proposed - Passive	$8.46 \pm 12.67°$	68.64%	80.00%	93.10%
Proposed - Active	$10.84 \pm 5.50°$	45.45%	77.27%	90.91%

Table 6.3: Passive localisation performance comparison.

model provides better passive localisation performance in a practical laboratory environment. As for the active localisation, the performance is slightly degraded compared with the proposed passive method, but a notable improvement can be observed on the standard deviation of the angular error. This indicates that the active head movement can effectively resolve frontback confusion and obviously increase the preciseness of the localisation.

6.4.3 Multiple-source Localisation

Finally, the multiple-source localisation is tested with the above experiment setup. For each testing signal, two randomly selected utterances are played by two different random speakers from number 1 to 20, and 200 signals are tested. The localisation results evaluated in correctness percentage are shown in Table 6.4.

Metrics	**1-source**	**2-source**
$\leq 5°$	43.60%	26.25%
$\leq 10°$	72.40%	65.62%
$\leq 15°$	75.94%	69.40%

Table 6.4: Localisation performance of two-sources localisation.

From Table 6.4, it can be observed that the proposed method successfully separates and localises most test signals during the experiments. However, there is a degrading, compared with the simulation results, which could be caused by multiple factors. For example, in the laboratory, part of the noise is not entirely uncorrelated white noise, as assumed in the simulation (e.g., the air conditioner), which could result in a bias in the estimation of the sound source direction.

6.5 Conclusion

This chapter has demonstrated the design of the binaural testing system and presented practical testing of the proposed localisation methods in a laboratory environ-

ment. Three listing scenarios have been tested, which included passive localisation, active localisation and multiple-source localisation. The testing results proved the effectiveness of the proposed methods in the real-world environment, and indicated remarkable localisation performance.

The specific contributions of this chapter are as follows:

i Testing the proposed localisation method in a practical environment, which proved that the proposed method can effectively localise the sound source without prior knowledge of the environment.

ii Testing the head rotation model in a practical scenario without de-reverberation processing, which also proved that the head movement can significantly promote the localisation performance in an indoor environment.

iii Verifying the compatibility of the proposed methods with existing hardware equipment, and designing a binaural testing system with an array of 30 loudspeakers and a KEMAR dummy head.

Chapter 7

Conclusion and Further Research Directions

This chapter states the general conclusions drawn from this thesis, as well as possible future research arising from this work. A summary of the research contributions can be found at the end of the previous chapters and is not repeated here.

7.1 Conclusion

This thesis was concerned with understanding binaural sound source localisation and attaining insight into the spatial features that are exploited by human listeners. Motivated by the growing interest in exploiting human auditory localisation information and building humanoid binaural hearing models, this thesis considered two open problems: (i) evaluating and retrieving the most valuable spatial-related interaural features, and (ii) creating a mapping model between those features and the sound source locations.

By introducing the existing binaural localisation methods, we identified the problem of binaural spatial information extraction to exploit the interaural features from a dual-channel system, and selected the most valuable spectral ranges for different types of features. The most significant conclusion was that spectral-based selection of interaural features can improve localisation performance in the presence of noise, especially for localisation in the vertical directions, which is based on machine learning methods to evaluate the dependency between features and the source positions. It was also shown that the significance of localisation cues depends not only on frequencies, but also on the corresponding sound source positions, and such observation provides a new possible means to construct binaural localisation models. The proposed localisation model integrates the feature selection procedure in the training process, which provides more means to evaluate the importance of binaural

features corresponding to the localisation performance.

The remainder of this thesis investigated the influence of listeners active movement and the related dynamic features. We summarised those movements into three typical styles, and devised the corresponding interaural feature behaviour based on a spherical listener model. The analysis demonstrated that the rotation on the horizontal plane provided the most distinct features for localisation in the 3-D auditory scenes. Further, such movement transferred the vertical difference of the incident sound direction into a variety of interaural cues, which provided a more robust way to estimate the vertical location of sound sources, and enabled the model to distinguish multiple sources during a period of time. Hence, the model was capable of estimating multiple sound sources based only on the interaural signal arrival time differences before and after rotation. Overall, this thesis has shown that selecting a particular range of features can improve the sound source localisation performance in 3-D spaces.

The preferred feature ranges are dependent on both frequency and source direction. A novel localisation model that emphasised those features' characteristics during training was presented, and the corresponding localisations were performed on a KEMAR dummy head in a real laboratory environment. In addition, the active head rotation movement of listeners could provide simple yet robust dynamic cues for 3-D space localisation, which enables the listener to separate multiple sources and estimate their directions easily.

7.2 Further Research Direction

A number of problems that can be explored following on the basic concepts that have been proposed in this thesis, which gives rise to a series of possible future research directions. Listed below is a subset of these problems related to the binaural localisation.

Generic Binaural Localisation Model

One avenue of future research is to generate a generic binaural localisation model. The localisation models proposed in this thesis were customised based on a single subject, which only reflected the specific localisation performance for one set of HRTF. In future work, the valuable features between different subjects could be studied, and a generic localisation model applicable to most human-like binaural

dummy head setups could be proposed. Such a generic model would have great value in building a human-like auditory robot.

Active Localisation Based on Complex and Arbitrary Movement

In Chapter 5, we investigated three types of head movement and their corresponding variation of ITDs. It was shown that the rotation movement can provide the most robust and directive cues for localisation. However, human head movement can be a combination of those three types of movement, so the best posture corresponding to the HRTF for localisation can be studied.

Correlation between the Localisation Model and Human Localisation Performance

The localisation models proposed in this thesis were mainly based on the HRTF from the CIPIC datasets, which were collected from 45 human subjects. In future work, the relationship of localisation performance between the proposed model and real humans could be studied. Such a comparison could result in a better understanding of spatial feature usage from the perspective of psychoacoustics, which would be beneficial to design a new generation of hearing aids and testing devices.

Simulated Human Hearing Model

The current localisation model used HRTF data collected from the human subject; however, these data were not necessarily reflecting actual human spatial hearing behaviour . A more human-liked localisation model is expected , which would not only localise the sources, but also reappear the errors and mistakes as humans do. Such a model would make a huge contribution to the development of virtual audio and spatial audio device testing.

Bibliography

[1] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in video conferencing," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr 1997, vol. 1, pp. 187–190 vol.1.

[2] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, Aug. 2000, vol. 2, pp. II909–II912.

[3] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robots: Sound localization and separation," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, Dec 1997.

[4] R. A. Brooks, C. Breazeal, M. Marjanovi, B. Scassellati, and M. M. Williamson, "The cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, Berlin, Heidelberg, 1999, pp. 52–87, Springer Berlin Heidelberg.

[5] D. R. Begault and L. J. Trejo, *3-D sound for virtual reality and multimedia*, Academic Press Inc. Cambridge, Cambridge, 2000.

[6] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc*, vol. 49, no. 10, pp. 904–916, 2001.

[7] I. J. Tashev and H. Gamper, "Head-related transfer function personalization for the needs of spatial audio in mixed and virtual reality," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3536–3536, 2017.

[8] B. G. Bregman, *Auditory scene analysis: the perceptual organization of sound*, The MIT Press, Cambridge, MA, 1990.

[9] D.L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[10] Lord Rayleigh O.M. Pres. R.S., "On our perception of sound direction," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.

[11] L. A. Jeffress, "A place theory of sound localization.," 1948.

[12] I.J. Hirsh, "The influence of interaural phase on interaural summation and inhibition," *The Journal of the Acoustical Society of America*, vol. 20, no. 4, pp. 536–544, 1948.

[13] J. C. R. Licklider, "The influence of interaural phase relations upon the masking of speech by white noise," *The Journal of the Acoustical Society of America*, vol. 20, no. 2, pp. 150–159, 1948.

[14] R. Lyon, "A computational model of binaural localization and separation," in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr 1983, vol. 8, pp. 1148–1151.

[15] M. Bodden, "Modeling Human Sound Source Localization and the Cocktail-Party-Effect," *Acta Acustica 1*, vol. 1, pp. 43–55, 1993.

[16] J. Blauer, *Spatial Hearing*, MIT Press, Cambridge, MA, England, 1997.

[17] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-d localization based on hrtfs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006, vol. 5, pp. 341–344.

[18] F. Keyrouz and K. Diepold, "an enhanced binaural 3d sound localization algorithm," in *Signal Processing and Information Technology*, Vancouver, BC, Canada, 2006, pp. 663–665.

[19] M. Usman, F. Keyrouz, and K. Diepold, "Real time humanoid sound source localization and tracking in a highly reverberant environment," in *Proc. 9th International Conference on Signal Processing (ICSP)*, Beijing, China, 2008, pp. 2661–2664.

[20] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3D localization based on HRTFs," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Darmstadt, Germany, 2006.

[21] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[22] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, Aug 1985.

[23] C. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[24] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1207–1217, 2014.

[25] Jack Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, 1974.

[26] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, no. 1, pp. 135–159, Feb. 1991.

[27] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: effect of pinnae cavity occlusion," *Journal of the Acoustical Society of America*, vol. 53, no. 2, pp. 400, 1973.

[28] V. R. Algazi, C. Avendano, and R. O. Duba, "Elevation localization and head-related transfer function analysis at low frequencies.," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, Mar. 2001.

[29] R. O. Duda, "Elevation dependence of the interaural transfer function.," *Binaural and spatial hearing in real and virtual environments.*, pp. 49–75, 1997.

[30] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2098–2107, Sep. 2014.

[31] J. Weng and K. Y. Guentchev, "Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning.," *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 310–323, 2001.

[32] A. Deleforge and R. Horaud, "2d sound-source localization on the binaural manifold.," in *IEEE International Workshop on Machine Learning for Signal Processing*, Santander, 2012, pp. 1–6.

[33] A. Deleforge and F. Forbes, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 1440003, Feb. 2015.

[34] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2915–2929, Jan. 2005.

[35] J. Woodruff and D. L. Wang, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[36] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues.," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[37] H. WALLACH, E. B. NEWMAN, and M. R. ROSENZWEIG, "The precedence effect in sound localization," *The American journal of psychology*, vol. 62, no. 3, pp. 315–336, 1949.

[38] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc*, vol. 20, no. 2, pp. 146–159, 1972.

[39] J. Huang, N. Ohnishi, and N. Sugie, "A biomimetic system for localization and separation of multiple sound sources," *IEEE Transactions on Instrumentation and Measurement*, vol. 44, no. 3, pp. 733–738, Jun 1995.

[40] J. Huang, N. Ohnishi, X. Guo, and N. Sugie, "Echo avoidance in a computational model of the precedence effect," *Speech Communication*, vol. 27, no. 3, pp. 223 – 233, 1999.

[41] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[42] K Lebart, J M Boucher, and P N Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. United With Acust.*, vol. 87, pp. 359–366, 2001.

[43] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence.," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[44] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2171–2186, Nov. 2016.

[45] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940.

[46] S. Perrett and N. Noble, "The effect of head rotations on vertical plane sound localization," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2325–2332, 1997.

[47] F. L. Wightman, "Resolution of frontback ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 105, pp. 2841–2853, 1999.

[48] X. Zhong, L. Sun, and W. Yost, "Active binaural localization of multiple sound sources," *Robotics and Autonomous Systems*, vol. 85, pp. 83–92, 2016.

[49] J. Braasch, S. Clapp, A. Parks, T. Pastore, and N. Xiang, "binaural model that analyses acoustic spaces and stereophonic reproduction systems by utilizing head rotations," *The Technology of Binaural Listening*, p. 201223, 2013.

[50] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[51] N. Ma, T. May, , T. Wierstorf, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2699–2703.

[52] E. Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[53] E. C. Cherry, "Resolution of frontback ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, vol. 25, pp. 554559, 1953.

[54] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Paltz, NY, USA, Oct. 2001, pp. 99–102.

[55] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," Tech. Rep., MIT Media Lab Perceptual Computing, 1994.

[56] R. s. Woodworth and H. Schlosberg, *Experimental psychology (Second Edition)*, Holt, Rinehart & Winston Inc., New York, USA, 1954.

[57] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*, J. Ross Publishing, Plantation, FL, 2013.

[58] O. R. Duda, "Elevation dependence of the interaural transfer function," *Binaural and spatial hearing in real and virtual environment*, vol. ch.3, pp. 49–75, 1997.

[59] B. A. Jont and A. B. David, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[60] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[61] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, May 2009.

[62] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[63] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments.," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 99–102, Feb.

[64] A. Stphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[65] Westermann A., J. M. Buchholz, and Dau T., "Binaural dereverberation based on interaural coherence histograms.," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2767–2777, May 2013.

[66] C. M. Zannini, R. Parisi, and A. Uncini, "Binaural sound source localization in the presence of reverberation," in *2011 17th International Conference on Digital Signal Processing (DSP)*, July 2011, pp. 1–6.

[67] R. Parisi, R. Gazzetta, and E. D. Di Claudio, "Prefiltering approaches for time delay estimation in reverberant environments," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 3, pp. III–2997–III–3000.

[68] R. Nagase, K. Oishi, and T. Furukawa, "A cepstrum prefiltering approach for doa estimation of speech signal in reverberant environments," in *21st European Signal Processing Conference (EUSIPCO 2013)*, Sept 2013, pp. 1–5.

[69] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter.," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4771–4783, Sept. 2015.

[70] Caleb Rascon and Ivan Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184 – 210, 2017.

[71] J.C. Chen, K. Yao, and R. E. Hudson, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *EURASIP J. Adv. Signal Process.*, vol. 4, pp. 359–370, Dec 2003.

[72] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, April 2015.

[73] Nikhil Deshpande and Jonas Braasch, "A correlation-based binaural localization model with reflection identification using neural networks," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1814–1814, 2018.

[74] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, Nov 2018.

[75] T. Gustafsson, B. D. Rao, and M. Trivedi, "Analysis of time-delay estimation in reverberant environments," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 2, pp. II–2097–II–2100.

[76] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 2565–2568.

[77] Kalle J. Palomki, Guy J. Brown, and DeLiang Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 4, pp. 361 – 378, 2004, Special Issue on the Recognition and Organization of Real-World Sound.

[78] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Trans. Syst., Man, Cybern., Part B*, vol. 36, no. 5, pp. 982–994, Oct. 2006.

[79] M. S. Datum, F. Palmieri, and A. Moiseff, "An artificial neural network for sound localization using binaural cues," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 372–383, 1996.

[80] J. Hu and W. Liu, "Location classification of nonstationary sound sources using binaural room distribution patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 682–692, May 2009.

[81] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, Jan 2011.

[82] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 463–479, 2006.

[83] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[84] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1323–1331, 1993.

[85] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002.

[86] R. Baumgartner, P. Majdak, and B. Laback, *Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications*, pp. 93–119, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[87] Hugh O'Dwyer, Enda Bates, and Francis M. Boland, "A machine learning approach to detecting sound-source elevation in adverse environments," in *Audio Engineering Society Convention 144*, May 2018.

[88] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[89] Tobias May, "Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 2, pp. 406–414, Feb. 2018.

[90] Yoshua Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[91] C. Schauer, T. Zahn, P. Paschke, and H. . Gross, "Binaural sound localization in an artificial neural network," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, June 2000, vol. 2, pp. II865–II868 vol.2.

[92] C. Schymura, J. D. Rios Grajales, and D. Kolossa, "Monte carlo exploration for active binaural localization," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 491–495.

[93] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[94] J. Blauert, "Sound localization in the median plane," *Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

[95] B. Rakerd, W. M. Hartmann, and T. L. McCaskey, "Identification and localization of sound sources in the median sagittal plane," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2812–2820, 1999.

[96] K. Iida, M. Itoh, A. Itagaki, and M Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, Aug. 2007.

[97] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots - building audio-motor maps based on the HRTF," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006, pp. 1170–1176.

[98] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *Journal of the Audio Engineering Society*, vol. 132, no. 6, pp. 3832–3841, 2012.

[99] K. D. Martin, "Estimating azimuth and elevation from interaural differences," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 1995, pp. 96–99.

[100] H. Liu and J. Zhang, "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *Proc. 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1424–1428.

[101] P. Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, Cambridge, 2009.

[102] J. Blauert, "Sound localization in the median plane," *Acustica*, vol. 22, pp. 205–213, 1969/70.

[103] G. R. Doddington, "Speaker recognition - identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[104] Cox R. M. and Moore J. N., "Composite speech spectrum for hearing aid gain prescriptions," *J. Speech Hear Res.*, vol. 31, no. 1, pp. 102–107, 1988.

[105] D. S. Deepawale and R. Bachu, "Energy estimation between adjacent formant frequencies to identify speaker's gender," in *Proc. 5th International Conference on Information Technology: New Generations (ITNG)*, Las Vegas, NV, USA, Apr. 2008, pp. 772–776.

[106] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech Language*, vol. 27, no. 3, pp. 621–633, May 2013.

[107] B. R. Hammond and P. J. B. Jackson, "Robust full-sphere binaural sound source localization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 86–90.

[108] B. R. Hammond and P. J. B. Jackson, "Robust full-sphere binaural sound source localization using interaural and spectral cues," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 421–425.

[109] H. Liu, Z. Fu, and X. Li, "A two-layer probabilistic model based on time-delay compensation for binaural sound localization.," in *IEEE International Conference on Robotics and Automation*, Karlsruhe, 2013, pp. 2705–2712.

[110] H. Liu and J. Zhang, "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 1424–1428.

[111] H. Liu, J. Zhang, and Z. Fu, "A new hierarchical binaural sound source localization method based on interaural matching filter.," in *IEEE International Conference on Robotics and Automation*, Hong Kong, 2014, pp. 1598–1605.

[112] D. Li and Levinson. S. E., "A bayes-rule based hierarchical system for binaural sound source localization," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. 521–524, Apr. 2003.

[113] P. Hanchuan, L. Fuhui, and D. Chris, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 27, no. 8, pp. 1226 – 1238, Aug. 2005.

[114] L. T. Vinh, S. Lee, Y. T. Part, and B.J. Auriol, "A novel feature selection mehtod based on normalized mutual information," *Springer Appl. Intell.*, vol. 37, pp. 100 – 120, Jul. 2012.

[115] W. Fan, E. Greengrass, J. McCloskey, P. s. Yu, and K. Drammey, "Effective estimation of posterior probabilities: explaining the accuracy of randomized decision tree approaches," *IEEE International Conference on Data Mining (ICDM'05)*, p. 8, Nov. 2005.

[116] A.H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 521–533, May 2009.

[117] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[118] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am*, vol. 118, no. 1, pp. 364–374, Jul. 2005.

[119] D. Campbell, "The roomsim user guide (v3.3)," 2004.

[120] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3108–3119, 2008.

[121] W. R. Wightman, J. W. Mangels, and P. S. Runge, "Head movement during sound localization," *The Journal of the Acoustical Society of America*, vol. 42, pp. 489–493, 1967.

[122] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, pp. 3048–3058, Jul. 1998.

[123] R. O. Duda, C. Avendano, and V. r. Algazi, "An adaptable ellipsoidal head model for the interaural time difference," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, AZ, mAR. 1999, vol. 104, pp. 965–968.

[124] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, 2000, vol. 5, pp. 2985–2988 vol.5.

[125] N. Roman, D. Wang, and G.J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2236, Oct. 2003.

[126] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, Sep. 2014.

[127] X. Zhao and D. Wang, "Analyzing noise robustness of mfcc and gfcc features in speaker identification," pp. 7204–7208, Apr. 2013.

[128] J. Chen, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," pp. IV–277–IV–280, Apr. 2007.

[129] Y. Shao, Z. Jin, and D. Wang, "An auditory-based feature for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Apr. 2009, pp. 4625–4628.

[130] K. Han, Y. Lu, Y. Hu, and P. C. Loizou, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, pp. 3475 –3284, Sep. 2012.

[131] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 114, pp. 1486–1494, Jul. 2009.

[132] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[133] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, Sep. 2014.

[134] X. Zhao and D. Shao, Y. Wang, "Casa-based robust speaker identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.