# SPATIAL AUDIO REPRODUCTION USING PRIMARY AMBIENT EXTRACTION

**JIANJUN HE**

School of Electrical & Electronic Engineering

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

**2016**

# Acknowledgements

First and foremost, I would like to thank my supervisor, Associate Professor Gan Woon-Seng, for his continued support and invaluable guidance. I thank him for bringing me to Singapore, taking care of me, and encouraging me throughout the course of this work. He has always been very supportive and provides the best resources for pursuing research work in DSP Lab. His professional expertise and great leadership impress me the most.

My sincere thanks also goes to Dr. Joseph Tan Ee Leng, for his guidance, ideas, discussions, and helps in improve my technical presentation and writing skills. This thesis work could not have been completed without his contributions. I would also like to thank Dr. Sam Shi Chuang for his sharing and encouragements. Thanks to Mr. Yeo Sung Kheng, for his continual support and help in any matters in the DSP Lab. Thanks to Dr. Mu Hao for taking so many nice photos, and being a sincere friend. My gratitude must also go to all my friends from the DSP Lab, Dr. Lee Kong Aik, Dr. Vincent Wang Liang, Mr. Ji Wei, Mr. Wang Tongwei, Dr. Kaushik Sunder, Mr. Rishabh Ranjan, Mr. Phyo Ko Ko, Mr. Ted Chen Chiu Hao, Mr. Kumar Dileep, Ms. James Anusha, Ms. Sabrina Rahmawati, Mr. Du Bo, Ms. Mahapatra Anushree, Mr. Apoorv Agha, Mr. Lam Bhan, Ms. Santi Peksi, Mr. Cao Yi, Mr. Ang Yi Yang, Mr. Nguyen Duy Hai, Mr. Zou Bingbing, Dr. Stefano Fasciani, Mr. Belyi Valiatsin, Dr. Tatsuya Murao, and Mr. Shi Dongyuan. They have made the lab as warm as home.

I would also like to thank my examiners and lecturers in NTU, including Assoc Prof. Ng Boon Poh, Assoc Prof. Soon Ing Yann, Assoc Prof. Chng Eng Siong, Assoc Prof. Chau Lap Pui, Assoc Prof. Andy Khong Wai Hoong, Assoc Prof. Chong Yong Kim, Assoc Prof. Saman S Abeysekera, Assoc Prof. Anamitra Makur, Assoc Prof. Jiang Xudong and so on. I have learned a lot from them.

To all my other friends in NTU and Singapore, Dr. Wang Dawei, Dr. Deng Jun, Dr. Qi Yongbing, Mr. Hou Junhui, Mr. Yin Jun, Mr. Jin Yichao, Mr. Xiong Qi, Mr. Wu Kai, Mr. Wang Bing, Mr. Zhao Lifan, Mr. Li Heng, to name a few, thank you for making my life colorful and memorable in Singapore.

The work was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, in collaboration with Phillips. I wish to thank all the financial supporters.

Last but not least, I would like to express my gratitude to my family for their unconditional love and consistent support. Thanks to Dad and Mum for raising me up and providing me the best opportunities and resources for my study. Thanks to Li Ya’s parents, for raising such a nice daughter and trusting me to take her from them. And above all, my deepest thanks goes to my wife Li Ya, for her tolerance, sacrifice, accompany, and sharing. She is my greatest inspiration and I dedicate this thesis to her.

# Table of Contents

# Summary

Recreating a natural listening experience is the aim of spatial audio reproduction of recorded audio content using playback devices, such as loudspeakers and headphones. Majority of the legacy audio contents are in channel-based format, which is dependent on the desired playback system. Considering the diversity of today's playback systems, the quality of reproduced sound scenes degrades significantly when mismatches between the audio content and the playback system occur. With the aim to solve this pressing issue and improve human's listening experience, this thesis focuses on the development of an efficient, flexible, and immersive spatial audio reproduction system based on primary ambient extraction (PAE).

Inspired by the human auditory system, the sound scene is considered as the mixture of a foreground sound (primary component, directional) and a background sound (ambient component, diffuse). The primary and ambient components are rendered separately to preserve their spatial characteristics, in accordance with the actual playback configurations. The core problem is how to extract the primary and ambient components from channel-based audio content efficiently. To answer this question, this thesis begins with the fundamentals of spatial hearing, and reviews existing spatial audio reproduction techniques, as well as prior arts in primary ambient extraction. The focus of this thesis is to enhance the performance of PAE in various scenarios encountered in practice.

Existing PAE approaches, such as the principal component analysis (PCA) and the least-squares (LS) method, though widely used, were not well studied.

To fill in this gap, these existing PAE approaches are generalized into a linear estimation framework. Under this framework, a series of performance measures are proposed to identify the components that contribute to the extraction error. Finally, a comprehensive comparative study and experimental testing of the linear estimation based PAE approaches, including PCA, LS, and three proposed variants of the LS, are presented.

Previous studies revealed that these state-of-the-art PAE approaches suffer from severe extraction error when dealing with sound mixtures that contain a relatively strong ambient component, a commonly encountered case in the sound scenes of digital media. To improve the PAE performance, we propose a novel ambient spectrum estimation (ASE) framework. The ASE framework exploits the equal magnitude of the uncorrelated ambient components in two channels of a stereo signal, and reformulates the PAE problem into the problem of estimating either ambient phase or magnitude. In particular, we take advantage of the sparse characteristic of the primary components to derive sparse-constrained solutions for ASE based PAE, as well as an approximate but efficient solution. The objective and subjective experiment results indicated a significantly better performance of the proposed ASE approaches over existing approaches, especially when the ambient component is relatively strong.

Considering most of these existing PAE approaches are mainly based on a basic stereo signal model, it is necessary to study PAE on input signals that do not satisfy the model assumptions. Taking PCA as a representative of PAE approaches, this thesis further investigates the performance degradation of PAE with respect to the correlation of the primary components in the cases with partially correlated primary components. To alleviate such performance degradation, a time-shifting technique is proposed by time-shifting the input signal according to the estimated inter-channel time difference of the primary component prior to the signal decomposition using conventional PAE approaches. The switching artifacts, caused by varied time-shifting in successive frames, can be avoided using overlapped output mapping. Comparative experimental results validate the improved performance of the time-shifting based PAE approaches.

In practice, the complex audio scenes could even include multiple concurrent sources in the primary components. Subband techniques are commonly implemented in PAE to deal with such signals. The effect of subband decomposition on PAE is investigated. The results indicate that the partitioning of the frequency bins is very critical in PAE and the proposed top-down adaptive partitioning method achieves superior performance, as compared to the conventional partitioning methods. Moreover, we also extended the time-shifting technique to multiple shifts. It is found that the consecutive multi-shift PAE with proper weighting yields more robust results. Finally, we introduce an important concept of natural sound rendering for rendering spatial sound over headphones, where PAE is one integral part.

In conclusion, several advancements on PAE are presented. Objective and subjective evaluations validate the feasibility of applying PAE in spatial audio reproduction. With these advanced PAE approaches readily applied, the listeners can thus immerse him/her-self in the reproduced sound scenes, without the limitation on the audio contents or playback systems.

# List of Figures

# List of Tables

# List of Abbreviations and Acronyms

| | |
|---|---|
| 3D | Three dimensional |
| AME | Ambient magnitude estimation |
| AMES | Ambient magnitude estimation with a sparsity constraint |
| ANC | Active noise control |
| ANOVA | Analysis of variance |
| APE | Ambient phase estimation |
| APES | Ambient phase estimation with a sparsity constraint |
| APEX | Ambient phase estimation using input signal |
| ASE | Ambient spectrum estimation |
| ASES | Ambient spectrum estimation with a sparsity constraint |
| BCC | Binaural cue coding |
| BRIR | Binaural room impulse response |
| BSS | Blind source separation |
| CASA | Computational auditory scene analysis |
| DirAC | Directional Audio Coding |
| DSR | Distortion-to-signal ratio |
| DRR | Direct to reverberation ratio |
| ESR | Error-to-signal ratio |
| ERB | Equivalent rectangular bandwidth |
| FFT | Fast Fourier transform |
| HPTF | Headphone transfer function |
| HRIR | Head-related impulse responses |
| HRTF | Head-related transfer function |
| HOA | High order Ambisonics |
| ICA | Independent component analysis |
| ICC | Inter-channel cross-correlation coefficient |
| ICLD | Inter-channel level difference |
| ICTD | Inter-channel time difference |
| ILD | Interaural level difference |
| ISR | Interference-to-signal ratio |
| ITD | Interaural time difference |
| JND | Just-noticeable difference |
| LoRo | Left only, Right only |
| LeSR | Leakage-to-extracted-signal ratio |
| LtRt | Left total, Right total |
| LS | Least-squares |
| LSR | Leakage-to-signal ratio |
| MAA | Minimum audible angle |
| MDLS | Minimum distortion LS |
| MLLS | Minimum leakage LS |
| MOS | Mean opinion score |
| MSE | Mean-square-error |
| MSPCA | Multi-shift PCA |
| MVDR | Minimum variance distortionless response |
| NLMS | Normalized least-mean-squares |
| NMF | Non-negative matrix factorization |
| PAE | Primary ambient extraction |
| PCA | Principal component analysis |
| PDR | Pressure division ratio |
| PPF | Primary panning factor |
| PPR | Primary power ratio |
| QMF | Quadrature mirror filter |
| RIR | Room impulse response |
| RMSE | Root-mean-square-error |
| SAOC | Spatial Audio Object Coding |
| SASC | Spatial audio scene coding |
| SD | Spectral distortion |
| SPCA | Time-shifted PCA |
| STFT | Short-time Fourier transform |
| USAC | Unified Speech and Audio Coding |
| VBAP | Vector base amplitude panning |
| WDO | W-disjoint orthogonality |
| WFS | Wave field synthesis |

# List of Symbols

| | |
|---|---|
| $\mathbf{a}_c$ | Ambient component |
| $\hat{\mathbf{a}}_{\mathrm{ALS},\,c}$ | Ambient component extracted using ALS |
| $\hat{\mathbf{a}}_{\mathrm{LS},\,c}$ | Ambient component extracted using LS |
| $\hat{\mathbf{a}}_{\mathrm{MDLS},\,c}$ | Ambient component extracted using MDLS |
| $\hat{\mathbf{a}}_{\mathrm{MLLS},\,c}$ | Ambient component extracted using MLLS |
| $\hat{\mathbf{a}}_{\mathrm{PCA},\,c}$ | Ambient component extracted using PCA |
| $\hat{a}_{\mathrm{SPCA},\,c}$ | Ambient component (one sample) extracted using SPCA |
| $\mathbf{A}$ | Anthropometric features |
| $\mathbf{A}_c$ | Ambient component in frequency domain |
| $b$ | Subband index |
| $c \in \{0,1\}$ | Channel index |
| $D$ | Number of phase or magnitude estimates in discrete searching |
| $Dist$ | Distortion in the extraction error |
| $\mathbf{e}$ | Extraction error |
| $fs$ | Sampling rate |
| $g_{ck}$ | Gain of source *k* in channel *c* |
| $G$ | Electrical transmission gain |
| $\mathbf{H}$ | HRTF magnitude |
| $H_L(f,\theta,\phi,r,lsn)$ | Left ear HRTF at frequency f, azimuth *θ,* Elevation ø, distance *r*, and listener *lsn* |
| $H_R(f,\theta,\phi,r,lsn)$ | Right ear HRTF at frequency f, azimuth *θ,* Elevation ø, distance *r*, and listener *lsn* |
| $h_{skL}(n),\ h_{skR}(n)$ | HRIR (Left, and right) for the source signal *k* |
| $h_{xcL}(n),\ h_{xcR}(n)$ | HRIR (Left, and right) for the mixture signal at channel *c* |
| $Intf$ | Interference in the extraction error |
| $J$ | Cost function |
| $k$ | Primary panning factor |
| $\hat{k}_{ic}, \hat{\gamma}_{ic}$ | Estimates of *k* and $\gamma$ in ideal case |
| $\hat{k}_{pc}, \hat{\gamma}_{pc}$ | Estimates of *k* and $\gamma$ in the primary-complex case |
| $l$ | Bin index |
| $Leak$ | Leakage in the extraction error |
| $m$ | Time frame index |
| $n$ | Sample index |
| $N$ | Frame length |
| $\mathbf{p}_c$ | Primary component |
| $\hat{\mathbf{p}}_{\mathrm{ALS},\,c}$ | Primary component extracted using ALS |
| $\hat{\mathbf{p}}_{\mathrm{LS},\,c}$ | Primary component extracted using LS |
| $\hat{\mathbf{p}}_{\mathrm{MDLS},\,c}$ | Primary component extracted using MDLS |
| $\hat{\mathbf{p}}_{\mathrm{MLLS},\,c}$ | Primary component extracted using MLLS |
| $\hat{\mathbf{p}}_{\mathrm{PCA},\,c}$ | Primary component extracted using PCA |
| $\hat{p}_{\mathrm{SPCA},\,c}$ | Primary component (one sample) extracted using SPCA |
| $\mathbf{P}_c$ | Primary component in frequency domain |

| | |
|---|---|
| $\hat{\mathbf{P}}_i$ | Extracted primary component from shifted signal |
| $Q$ | Duration of the overlapping samples in the stereo signals (in ms) |
| $r$ | Correlation or ambient magnitude |
| $s_k(n)$ | Source signal |
| $SP_L(f,\theta,\phi,r,lsn)$ | Left eardrum sound pressure at frequency *f*, azimuth *θ,* elevation *ø*, distance *r,* and listener *lsn* |
| $SP_0(f,r)$ | Sound pressure at the center of the head with the absence of the head, at frequency *f*, and distance *r* |
| $SP_R(f,\theta,\phi,r,lsn)$ | Right eardrum sound pressure at frequency *f*, azimuth *θ,* elevation *ø*, distance *r,* and listener *lsn* |
| $\mathbf{u}_{\mathrm{A}}$ | Ambient basis vector |
| $\mathbf{u}_{\mathrm{P}}$ | Primary basis vector |
| $\mathbf{W}$ | Weighting matrix in linear estimation based PAE |
| $\mathbf{W}_c$ | Complex exponential of ambient phase spectra $\boldsymbol{\theta}_c$ |
| $\mathbf{x}_c$ | Mixed signal |
| $\mathbf{X}_c$ | Mixed signal in frequency domain |
| $\mathbf{X}_l = \left\{\mathbf{x}_0, \overleftrightarrow{\mathbf{x}}_1^l\right\}$ | Shifted input signal |
| $y_L(n), y_R(n)$ | Signal fed to headphone |
| $Z$ | Impedance |
| $\gamma$ | Primary power ratio |
| $\theta$ | Azimuth |
| $ø$ | Elevation |
| $\lambda_{\mathrm{P}}$ | Eigenvalue that corresponds to primary basis |
| $\phi_{\mathrm{H}}$ | Higher bound for ICC |
| $\phi_{\mathrm{L}}$ | Lower bound for ICC |
| $\phi_{\mathrm{p}}$ | Correlation coefficient of the primary component (at zero lag) |
| $\phi_{\mathrm{x}}$ | Correlation coefficient of the mixed signal |
| $\tau$ | Lag index |
| $\tau_0$ | ICTD |
| $\tau_i$ | *i*th estimated ICTD |
| $\vartriangle k, \vartriangle\gamma$ | Ratio between the estimated *k*, $\gamma$ and their true values in the primary-complex case |
| $\odot$ | Element-wise Hadamard product |
| $*$ | Convolution |

# Chapter 1

# Introduction

## 1.1 Research area and motivation

Sound is an inherent part of our everyday lives for information, communication and interaction. Sound improves the situational awareness by providing feedback for actions and situations that are out of the view of the listener. An advantage of sound is that multiple sound sources can be perceived from any location around the head in the three dimensional (3D) space [Beg00]. The role of natural 3D sound, or spatial sound, is very essential in high stress applications, like flight navigation and communication systems [Air15], [BWG10]. Naturally rendered spatial sound has also been proven to be beneficial in personal route guidance for visually impaired people [LMG05], [Mic14] and in medical therapy for patients [DLH03], [ASI08], [SPL10]. Last but not least, the ever growing market of consumer electronics calls for spatial audio reproduction for digital media, such as movies, games, and virtual reality applications (e.g., Oculus Rift), augmented reality applications (e.g., Microsoft HoloLens).

Considering the variety of applications, spatial audio reproduction of digital media (especially the movies and video games) has gained significant popularity over the recent years [ITU12b]. The reproduction methods generally differ in the formats of audio content. Despite the growing interest in object-based audio formats [ITU12b], such as Dolby ATMOS [Dol13], DTS multi-dimensional audio (DTS: X) [JoF11], most existing digital media content is still in channel-based formats (such as stereo and multichannel signals). The channel-based audio is usually specific in its playback configuration, and it does not support flexible playback configurations in domestic or personal listening circumstances [ITU12b]. Considering the wide diversity of today's playback systems [HHK14], it becomes necessary to process audio signals such that the reproduction of the audio content is not only compatible with various playback systems, but also able to achieve the best quality (especially spatial quality [Rum02]) with the actual playback system [Rum11]. In line with the objective of the new MPEG-H standard for 3D audio [HHK14], this thesis aims to achieve a flexible, efficient, and immersive spatial audio reproduction.

Depending on the actual playback system, the challenges in spatial audio reproduction can be broadly categorized into two main types: loudspeaker playback and headphone playback [Rum13]. The challenge in loudspeaker playback mainly arises from the mismatch of loudspeaker playback systems in home theater applications, where the number of loudspeakers [Rum01] or even the type of loudspeakers [GTK11], [TaG12], [TGC12] between the intended loudspeaker system (based on the audio content) and the actual loudspeaker system is different. Conventional techniques to solve this challenge are often referred to as audio remixing (i.e., down-mix and up-mix), for example, "Left only, Right only (LoRo)", "Left total, Right total (LtRt)", matrix-based mixing surround sound systems, etc. [Rum01], [BaS07], [Ger92], [ITU93]. These audio remixing techniques basically compute the loudspeaker signals as the weighted sums of the input signals. For headphone playback, the challenge arises when the audio content is not tailored for headphone playback (usually intended for loudspeaker playback). Virtualization is often regarded as the technique to solve this challenge [Beg00], where virtualization of loudspeakers is achieved by binaural rendering, i.e., convolving the channel-based signals with head-related impulse responses (HRIRs) of the corresponding loudspeaker positions. These conventional techniques in spatial audio reproduction are capable of solving the compatibility issue, but the spatial quality of the reproduced sound scene is usually limited [BaS07], [BrS08], [BrF07], [ZiR03]. To improve the spatial quality of the sound reproduction, the MPEG audio standardization group proposed MPEG Surround and related techniques, which typically address the multichannel and binaural audio reproduction problem based on human perception [BrF07], [FaB03], [Fal04]. In the synthesis, these techniques usually employ the one-channel down-mixed signal and the spatial cues, which better suit the reproduction of the distinct directional source signals as compared to the diffuse signals [BrF07], [GoJ07b].

To further improve the quality of the reproduced sound scene, the perception of the sound scenes is considered as a combination of the foreground sound and background sound [StM15], which are often referred to as primary (or direct) and ambient (or diffuse) components, respectively [GoJ08], [HTG14], [SHT15], [KTT15]. The primary components consist of point-like directional sound sources, whereas the ambient components are made up of diffuse environmental sound, such as the reverberation, applause, or nature sound like waterfall [GoJ07b], [AvJ04]. Due to the perceptual differences between the primary and ambient components, different rendering schemes should be applied to the primary and ambient components for optimal spatial audio reproduction of sound scenes [GoJ07b], [MeF10]. However, the existing channel-based audio formats provide only the mixed signals [Hol08], which necessitate the process of extracting primary and ambient components from the mixed signals. This extraction process is usually known as the primary ambient extraction (PAE).

As a spatial audio processing tool [Rum01], [BrS08], [GoJ07b], [BrF07], [SHT15], [MeF10], PAE can also be incorporated into spatial audio coding systems, such as spatial audio scene coding [GoJ08], [JMG07], and directional audio coding [Pul07]. Essentially, PAE serves as a front-end to facilitate flexible, efficient, and immersive spatial audio reproduction. First, by decomposing the primary and ambient components of the sound scene, PAE enables the sound reproduction format to be independent of the input format, hence increasing the flexibility of spatial audio reproduction [JMG07], [Rum10]. Second, PAE based reproduction of sound scenes does not require the individual sound objects as in object-based format (which is the most flexible), but is able to recreate perceptually similar sound scenes, hence maintaining the efficiency of spatial audio reproduction [HTG14]. Last but not least, PAE extracts the two key components of the sound scenes, namely, directional and diffuse sound components. These components are highly useful in recreating an immersive listening experience of the sound scene [GoJ08], [JPL10], [UsB07], [Fal07], [KKM15].

Figure 1.1 illustrates the PAE based spatial audio reproduction system, where the primary and ambient components undergo different rendering schemes [HGT14]. The rendering schemes differ for loudspeaker or headphone

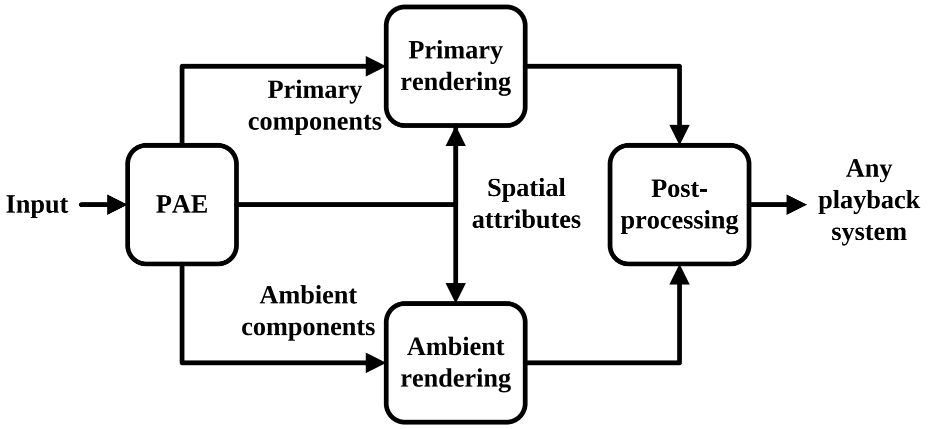


Figure 1.1 Block diagram of PAE based spatial audio reproduction

playback [AvJ04], [JPL10], [FaB11]. For loudspeaker playback, the primary components are reproduced using vector base amplitude panning (VBAP) [Pul97] or vector base intensity panning [GoJ06], [JLP99] to reproduce the accurate direction of the sound sources. The ambient components, on the other hand, are further decorrelated and distributed to all the loudspeaker channels to create an envelopment effect of the sound environment [GoJ08], [Fal06]. For headphone playback, the conventional virtualization that simply applies binaural rendering to the mixed channel-based signals suffers from virtual phantom effect as discussed in [BrS08], [GoJ07b]. PAE based virtualization resolves this problem by applying binaural rendering to the extracted primary components, creating accurate virtual sound sources in the desired directions [GoJ07b] for headphone playback [SHT15], [LBP14]. Similar to the loudspeaker playback case, the ambient components are decorrelated using artificial reverberation [BrF07], [GoJ08], [AvJ04], [MeF10] to create a more natural sound environment.

## 1.2 Objective

With the aim of improving humans' listening experience, an efficient, flexible, and immersive spatial audio reproduction using primary ambient extraction is the core objective of this thesis work, which can be divided into the following four aspects.

Firstly, due to the lack of systematical study of the existing PAE approaches, the relationships and the performance of these PAE approaches are unclear. This lack of theoretical knowledge hinders the development of better PAE approaches to improve the quality of spatial audio reproduction. Therefore, it is necessary to conduct a comprehensive evaluation of the existing PAE approaches.

Secondly, it is important to understand the drawbacks of the existing PAE approaches in certain cases and their appropriate application scenarios. Furthermore, the fundamental reasons for these drawbacks and how these drawbacks can be tackled need to be investigated.

Thirdly, an effective PAE approach must also be able to handle complex practical signals that may not match all the assumptions of the basic signal model. To improve the robustness of PAE, more complex signal models are considered and the performance of conventional PAE approaches (proposed for the basic signal model) shall be investigated. Techniques that can improve PAE performance in the complex cases are of particular interests in this thesis.

Lastly, it is important to address how to apply PAE in spatial audio reproduction systems. Take headphone playback as an example, it shall be investigated how PAE can be applied, in combination with other techniques, to achieve a more natural listening experience.

## 1.3 Major contributions of this thesis

This thesis focuses on the study and development of primary ambient extraction techniques for efficient, flexible, and immersive spatial audio reproduction. Its major contributions are highlighted as follows:

I. *Investigation of linear estimation based primary ambient extraction under the basic signal model.* First, we propose a linear estimation framework for PAE that generalizes existing PAE approaches, such as principal component analysis and least-squares. Secondly, two groups of measures are introduced to yield a more complete performance evaluation of the timbre and spatial quality of the PAE approaches. Thirdly, three variants of least-squares based PAE approaches are proposed, and a comprehensive evaluation and comparison of all these linear estimation based PAE approaches are conducted. Finally, we provide practical guidelines in selecting the proper PAE approaches in different spatial audio applications.

II. *Improving PAE performance for strong ambient power cases using ambient spectrum estimation techniques.* The performance of linear estimation based PAE approaches is inferior in strong ambient power cases, due to the limitations to cancel the uncorrelated ambient components without distorting the primary components. To circumvent this problem, we propose a new ambient spectrum estimation framework that reformulates the PAE problem as the problem of estimating ambient phase or magnitude. Solutions to ambient spectrum estimation are obtained by exploiting the sparsity of the primary components in the time-frequency domain. Computational complexity and robustness of the ambient spectrum estimation based PAE approaches are further investigated. To facilitate the detailed objective performance analysis using performance measures introduced in earlier work, an optimization method is proposed to compute these extraction performance measures for PAE approaches without analytic solutions (as is the case with the ambient spectrum estimation based PAE approaches). Finally, objective and subjective evaluations are performed to validate the performance of these PAE approaches.

III. *Employing time-shifting techniques for PAE with partially correlated primary components.* Practical signals are usually more complex than what is assumed in the basic signal model for PAE. One common case is the primary-complex case that considers the primary components to be partially correlated. Using conventional PAE approaches for these complex signals degrades PAE performance, as a function of primary correlation. Time-shifting techniques can be employed to increase the primary correlation to its maximum. Thus, the input signal is closest to the basic signal model, and conventional PAE approaches can be re-used. A corresponding output mapping can be employed to avoid the frame boundary switching artifacts due to time-shifting. Advantages of the proposed time-shifting based PAE approaches over conventional PAE approaches include lower extraction error and closer spatial cues, as shown in the experiments using synthetic signals and real recordings.

IV. *Adaptation of conventional PAE approaches to deal with primary components with multiple sources.* Though one dominant source in the primary components is found to be quite common in PAE, it is still possible to encounter the cases with multiple dominant (usually up to three) sources in some movies and games. Failing to consider this case will degrade the overall performance of the spatial audio rendering [ThH12]. To handle such cases, we investigate two ways to adapt the PAE approaches. The first technique considers subband decomposition of the full-band input signal before performing PAE on each subband signal. The partitioning of the frequency bins into subbands is found to be critical, where the adaptive top-down partitioning method outperforms other methods. The other way is the multi-shift technique that involves multiple instances of time-shifting, performing extraction for each shifted signals, and combining the extracted components from all shifting versions. The weighting method based on inter-channel cross-correlation is found to yield the best performance.

V. *Applying PAE in natural sound rendering headphone systems.* The application of PAE in headphone based spatial audio reproduction is discussed. Based on the comparative analysis of the differences between conventional headphone listening and natural listening, an important concept of natural sound rendering is proposed. Five types of signal processing techniques including PAE based sound scene decomposition are discussed to achieve natural sound rendering. We addressed the problem of integration of these signal processing techniques, which is explained using an exemplar 3D audio headphone system. Subjective listening tests were conducted to validate the improved performance brought by natural sound rendering.

## 1.4 Organization

This thesis is organized into eight chapters. Chapter 1 introduces the background, motivation, objective, and major contributions of this thesis work. Chapter 2 reviews the basics of spatial hearing. Based on the three types of audio representations, various spatial audio reproduction systems are discussed. Lastly, prior works on PAE are reviewed. In chapter 3, the widely used stereo signal model and the linear estimation framework for PAE are discussed. In-depth analysis on the extraction error leads to different objectives in PAE, and five linear estimation based PAE approaches are proposed and evaluated thoroughly. Based on the study in Chapter 3, we observed limited performance of these linear estimation based PAE approaches, especially when ambient power is relatively strong. Such a problem leads us to a new ambient spectrum estimation framework for PAE in Chapter 4, where the solutions can be obtained by exploiting the sparsity of the primary components. Simulations and subjective listening tests are conducted to validate the performance of these PAE approaches. Chapter 5 and Chapter 6 focus PAE in dealing with complex signals that are encountered in practice. In Chapter 5, we examine primary components with partial correlation at zero lag (i.e., primary-complex case). The performance of the conventional PAE approaches is investigated in the primary-complex case, leading to the proposed time-shifting technique. Following the study in Chapter 5, Chapter 6 proposes techniques based on subband decomposition and multi-shift techniques to handle complex primary components with multiple dominant sources. In Chapter 7, we discuss how PAE can be applied in spatial audio reproduction using headphones. An important concept of natural sound rendering is proposed, which integrates five

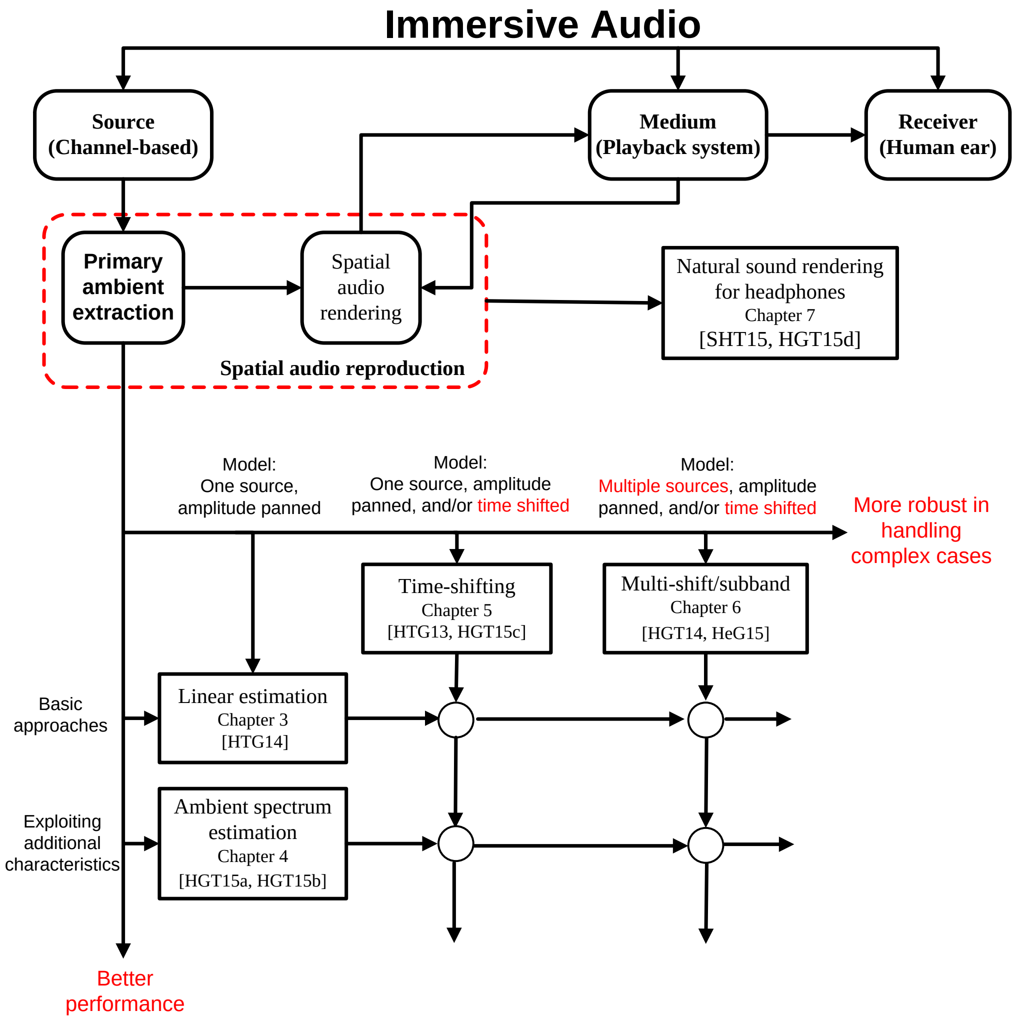


Figure 1.2 The overview and organization of this thesis. The circles denote that these approaches can be directly combined.

types of signal processing techniques, including sound scene decomposition using PAE. One example that implements the natural sound rendering concept, known as 3D audio headphones, is used for subjective evaluations. Finally, Chapter 8 concludes this thesis and points out some meaningful directions for future work.

Figure 1.2 shows how these chapters are linked to the major contributions and the related publications of the author. Given channel-based audio as the input (source), PAE is applied to achieve an immersive spatial audio

Table 1.1 Chapters of this thesis and authors' related publications[1]

| Chapter | Author's Publications | Published in |
|---|---|---|
| 3 | [J1] | TASLP, 2014 |
| 4 | [J3] | SPL, 2015 |
|  | [J4] | TASLP, 2015 |
| 5 | [C1] | ICASSP, 2013 |
|  | [J5] | TASLP, 2015 |
| 6 | [C2] | ICASSP, 2014 |
|  | [C5] | ICASSP, 2015 |
| 7 | [J2] | SPM, 2015 |
|  | [C4] | ICASSP, 2015 |
|  | [C8] | ICASSP, 2016 |

reproduction for any arbitrary playback systems (medium). The major contributions of this thesis lie in the development of a more robust PAE approach with enhanced performance, as illustrated in the lower part of Fig. 1.2. On the vertical axis, starting from the basic approaches (linear estimation, Chapter 3), we observe how the performance of PAE can be improved by exploiting more characteristics (ambient spectrum estimation, Chapter 4). On the horizontal axis, we improve the robustness of PAE in handling more complex cases, using time-shifting techniques (Chapter 5), and multi-shift/subband techniques (Chapter 6). The true advantage of these robustness enhancement techniques is that they can be inherently applied to any PAE approaches that were originally proposed for signals under the basic signal model. Thus, a complete network of PAE approaches can be established. Furthermore, an example of spatial audio reproduction systems that incorporate PAE is discussed in Chapter 7. Table 1.1 lists the related information of the publications for each chapter of this thesis.

[1] Refer to page 203 for the detailed information of the author's publications.

# Chapter 2

# Literature Review on Spatial Audio

Spatial audio, also known as three dimensional (3D) audio, refers to the perception of sound in 3D space and anything that is related to such a perception, including sound acquisition, production, mastering, processing, reproduction, and evaluation of the sound. This thesis describes the reproduction of 3D sound based on the formats of the audio content. For this purpose, we first review the fundamental principles of human's spatial hearing, and discuss various conventional, as well as advanced techniques for spatial audio reproduction. After that, a summary of the prior work on primary ambient extraction is presented.

## 2.1 Basics of spatial hearing

With the ears positioned on both sides of our head, humans are capable to perceive sound around us. The perceived sound can be processed by our brain to interpret the meaning of the sound. Equally amazing is our ability to localize sound in the 3D space. This capability of localizing sound in 3D space is often referred to as spatial hearing. In this section, we will review the fundamentals of spatial hearing.

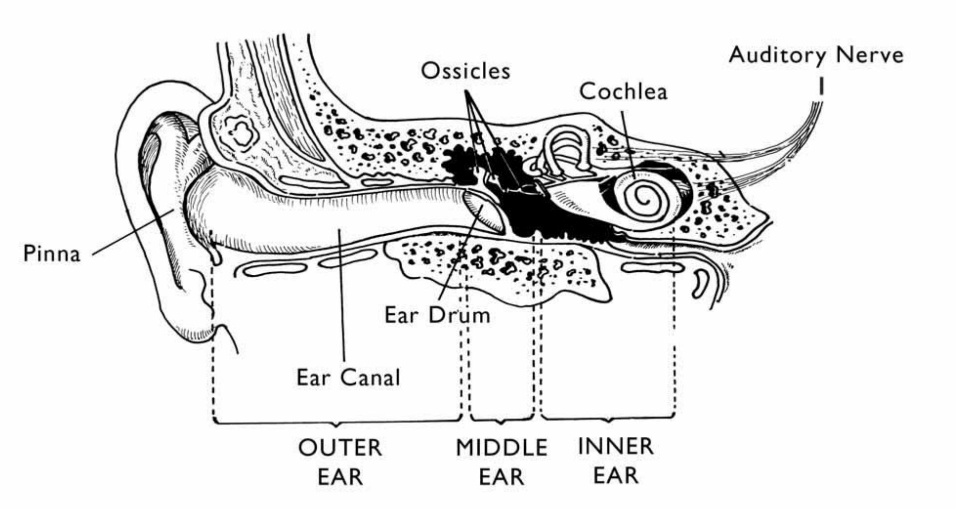


Figure 2.1 Structure of the human ear (extracted from [WHO06])

### 2.1.1 How do we hear sound

From a physical point of view, sound waves, emanating from a vibration process (a.k.a., sound source), travel through the air all the way into our ears. Human ears can be broadly separated into three parts: the outer ear, middle ear, and inner ear, as shown in Fig. 2.1 [WHO06]. The pinna of the outer ear picks up the sound and passes through the ear canal to the eardrum of the middle ear. The sound vibrations captured by the eardrum, are transformed into nerve signals by the cochlea. These nerve signals travel through the auditory nerve and reach our brain. Our brain can then interpret the sound we hear. Impairment to any parts of the ear would affect our hearing.

### 2.1.2 How do we localize sound

For a particular sound source in a 3D space, localization of this sound source would involve three dimensions. Clearly, take the listener (more specifically, the head of the listener) as the center of the space, a polar coordinate system is considered to be more appropriate to describe the 3D space. Hence, we describe the three dimensions as distance, azimuth, and elevation, as shown in Fig. 2.2. Distance is the length of the direct line path between the sound source and the center of the head. Horizontal plane refers to the plane that is horizontal to the ground at ear-level height. Median plane is a vertical plane that is perpendicular to the horizontal plane with the same origin at the center of the head. Azimuth $\theta$ refers to the angle between the median plane and the vector from center of the head to the source position. Azimuth is usually defined in clockwise direction, with 0 ° azimuth refers to the direction right in front of us. Elevation $\o$ is defined as the angle between the horizontal plane and the vector from center of the head to the source position. An elevation of 0 ° refers to a sound directly in front, and increasing elevation will first move the sound up, then behind, and finally under the listener.

In spatial hearing, sound localization can be considered in different perspectives. In terms of the position of the sound source, we usually consider the direction (i.e., azimuth and elevation) and distance of the sound. Perception of single sound source is different from multiple sound sources, where incoherent sound sources are perceived as separate auditory events and coherent sound sources are governed by summing localization (usually for sound sources with time difference under 1ms) or precedent effect (for time difference above 1ms, e.g., reflections) [Bla97]. Coherent sound sources that arrive after several milliseconds would be perceived as echo, which is quite common for sound in enclosed space. For sound localization task, human brains combine various cues from perceived sound and other sensory information such as visual images. It

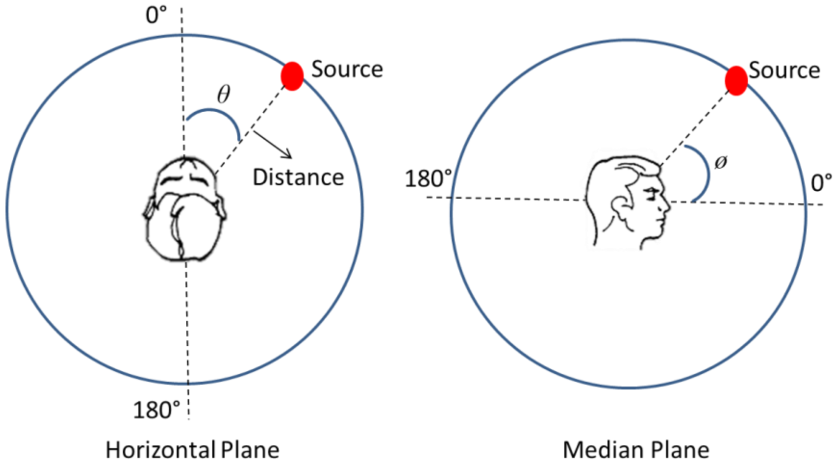


Figure 2.2 The coordinate system for sound localization

has been commonly known that the following cues contribute to sound localization [Bla97], [Beg00], [AlD11], [Xie13]:

1). Interaural time difference (ITD)

2). Interaural level difference (ILD)

3). Spectral cues (monaural, relevant to the anthropometry of the listener)

4). Head movement cues (a.k.a., dynamic cues)

5). Intensity, loudness cues

6). Familiarity to sound source

7). Direct to reverberation ratio (DRR)

8). Visual and other non-auditory cues

Among the seven auditory cues 1) to 7), the first four contribute to direction localization, whereas the last three affect distance perception.

### 2.1.3 Direction perception: azimuth and elevation

A variety of psychoacoustic experiments have demonstrated human’s ability to localize the direction of the sound source. The minimum audible angle (MAA) can reach as low as 1 °-3 ° for broadband sound (e.g., white noise) in the front horizontal plane (±90 ° azimuth), though it becomes worse for other directions and narrowband sound [Bla97]. The ITD and ILD are the two most important cues for azimuth direction localization. The ITD refers to the difference of time that the sound travels from the source to the left and right ears. Apparently, sound from different directions would have different traveling time durations to the two ears, resulting different ITDs. The ILD is mainly caused by the attenuation of the sound levels in the contralateral ear (further to the source) due to the head shadowing effect, compared to the ipsilateral ear (nearer to the source). According to the duplex theory [Ray07], ITD relates to the ability of human auditory system to detect interaural phase differences at low frequency and hence ITD is more dominant in low frequency, whereas ILD dominants at high frequency region. The cutoff frequency is determined by the distance between the two ears (typically 22-23cm), which is usually considered to be around 1,500 Hz.

For localization of sound in different elevations, ITD and ILD are not enough. This is because identical ITD and ILD values can be obtained from the sound source in a conical surface, as shown in Fig. 2.3 [Beg00]. This is the so-called “cone of confusion” phenomenon [Mil72]. One of the most common perceptual errors in cone of confusions is the front-back confusions, where one perceives a front (or back) sound in the back (or front). In order to perceive the elevation directions correctly, spectral cues are required. Spectral cues are

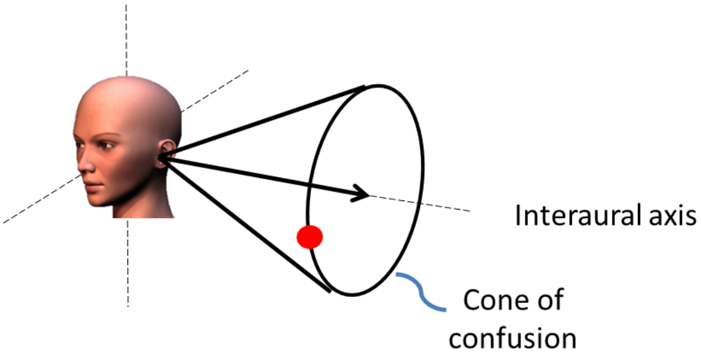


Figure 2.3 Cone of confusion due to identical ITD and ILD

mainly caused by head, torso, and pinna that filter the incoming sound waves. Sound from different elevations would reach different parts of our body (especially the pinna), and undergoes different reflections before entering the ear canal. Most of the spectral cues due to pinna occur at frequencies above 3 kHz, and the spectral cues due to head and torso appear in lower frequencies. It is worth mentioning that the spectral cues vary greatly from person to person due to the idiosyncratic anthropometry of the listener. In addition to the static cues mentioned above, dynamic cues due to head movement are extremely useful in resolving localization errors, especially front-back confusions.

The Head-related transfer function (HRTF) is usually introduced to describe the change on the sound spectra due to the interactions of the sound wave with the listener's head, torso, and pinna, which is defined as follows. In a free field environment, take the Fourier transform of the sound pressure ( $SP_L$ or $SP_R$ ) at the eardrums of the two ears and the sound pressure ( $SP_0$ ) at the center of the head with the listener absent. The HRTF is the ratio of these two Fourier representations. Since human has two ears, HRTF typically comes in pairs.

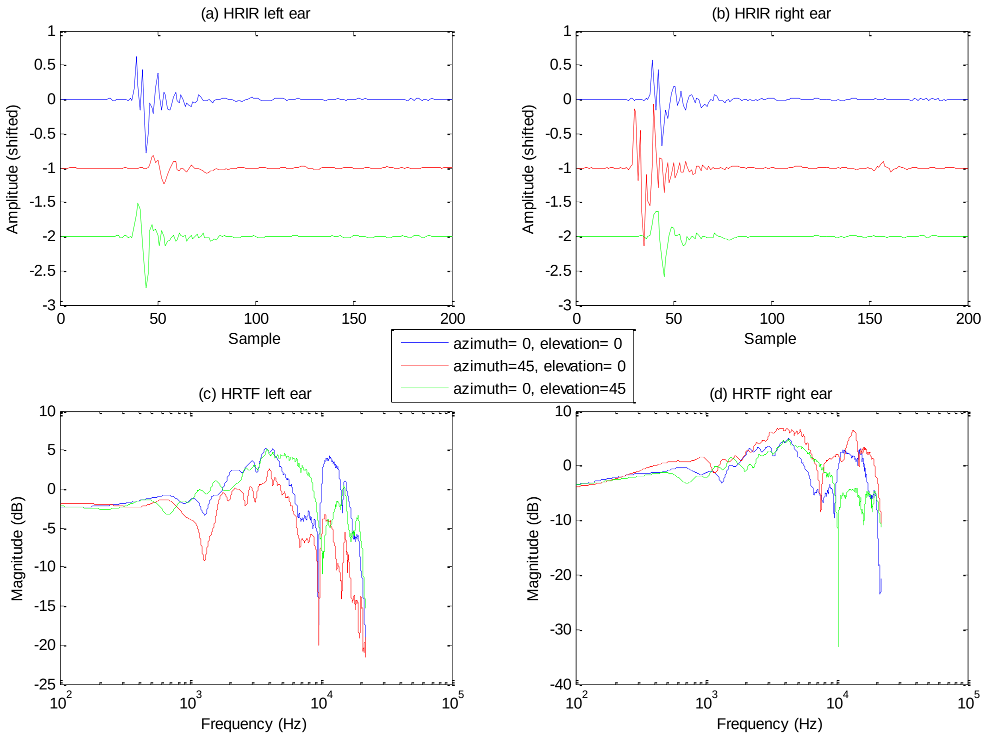


Figure 2.4 HRIR and HRTF of the same subject (CIPIC HRTF database subject 003 [ADT01]) in different directions.

Clearly, HRTF is a function of frequency (*f*), direction $(\theta,\phi)$, distance (*r*), and listener (*lsn*), and is expressed as

$$H_L\left(f,\theta,\phi,r,lsn\right)=\frac{SP_L\left(f,\theta,\phi,r,lsn\right)}{SP_0\left(f,r\right)},$$
$$H_R\left(f,\theta,\phi,r,lsn\right)=\frac{SP_R\left(f,\theta,\phi,r,lsn\right)}{SP_0\left(f,r\right)}. \tag{2.1}$$

where $P_L, P_R$ and $P_0$ are sound pressures in the frequency domain. According to Algazi *et al*. [ADM01], [BrD98], [ADD02], HRTF can be approximated by a structural composite of pinna-less head and torso, and the pinna, which is mainly effective at modifying the source spectra at low and high frequencies, respectively. In the far field, HRTF is usually considered to be independent of

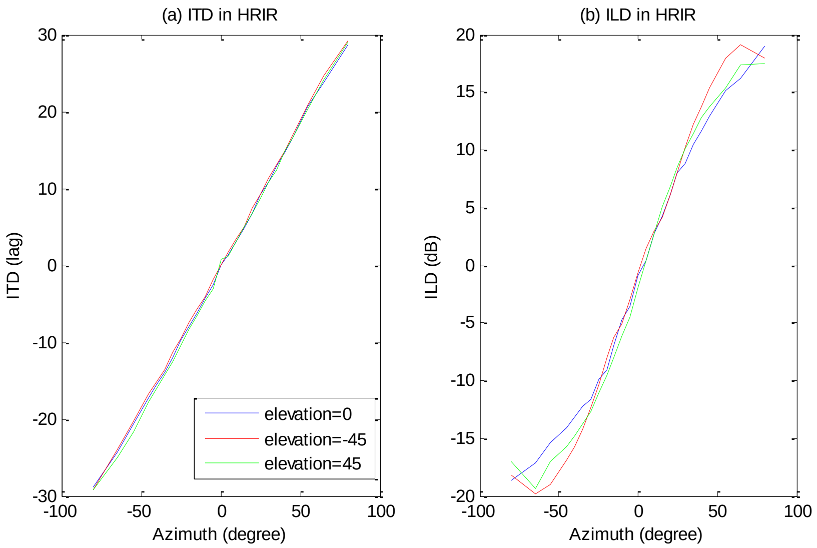


Figure 2.5 ITD and ILD of the same subject (CIPIC HRTF database subject 003 [ADT01]) in different directions.

distance [Ken95a]. The time domain representation of HRTF is referred to as head-related impulse response (HRIR).

In Figs 2.4, 2.5, and 2.6, the HRIR and HRTF of subjects from the CIPIC HRTF database are plotted [ADT01]. The HRIR and HRTF of the same subject at different directions are shown in Fig. 2.4. It is clear that the waveform and magnitude spectra shapes vary with the direction horizontally and vertically. In Fig. 2.5, we show the ITD and ILD (full-band) that are computed from the HRTFs of the same subject. It is clear that ITD and ILD exhibit a close-to-linear relationship with the azimuth, and the change across different elevations is minimal, especially at non-lateral azimuthal directions. The HRIR and HRTF of three different subjects are plotted in Fig. 2.6, which indicates that HRTF generally differs from individual to individual, especially the spectral notches in the high frequency range. The individual differences of HRTF among different subjects are indeed due to the anthropometric features of these subjects.

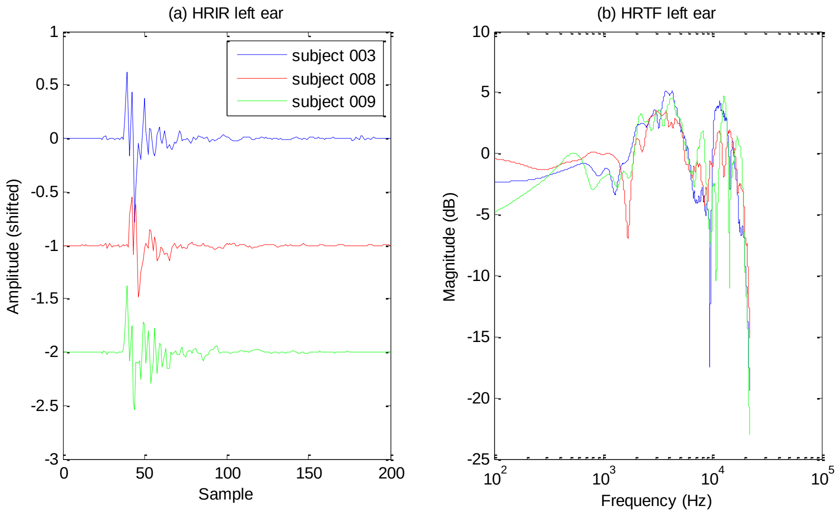


Figure 2.6 HRIR and HRTF (left ear, azimuth = 0 °, elevation = 0 °) of three different subjects (subjects 003, 008, 009 in the CIPIC HRTF database [ADT01]).

### 2.1.4 Distance perception

Perception of distance of sound sources is important in sound localization. In sound rendering, it is critical to recreate the perception of distance of the sources close to natural listening. However, the challenges in simulating accurate distance perception are numerous. Human beings' ability to accurately estimate the distance of a sound source has long been known to be poorer compared to our ability to estimate directions, even in the physical listening space [Zah02]. The experiments conducted by Zahorik showed that the perceived distance can usually be expressed in a power function of the actual distance [Zah02a]. The direct-to-reverberation energy ratio is found to be the most critical cue for absolute distance perception, even though the intensity, loudness, and binaural cues (including ILD, and interaural coherence) can provide relative cues for distance perception [Zar02b], [Beg00]. However,

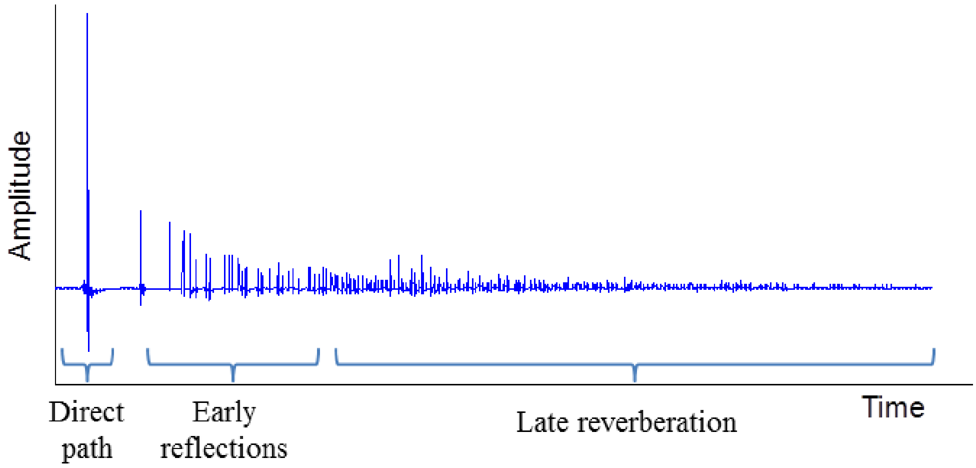


Figure 2.7 A schematic illustration of RIR (adopted from [VPS12])

accurate simulation of distance perception is challenging since reverberation depends on the room characteristics. The correct amount of reverberation to be added to simulate distance perception in a particular room can be obtained only by carrying out acoustical measurements.

### 2.1.5 Sound in rooms: reflections and reverberation

Though sound localization is discussed in free-field environment, the real-life sound environment is never free-field. The existing free-field environment can only be found in an anechoic chamber. Rooms that we live in everyday are filled with reflections and reverberation, usually characterized by the room impulse response (RIR). A schematic illustration of RIR is shown in Fig. 2.7. A typical RIR consists of three parts: the direct path, early reflections, and late reverberation (after 80ms). An important aspect of room acoustics is the reverberation time $RT_{60}$, which is defined by the time that it takes for the sound to attenuate by 60 dB once the sound source ceases. To simulate the perception of sound in rooms (or sound environment in general), RIRs that are derived or measured from the (approximately) geometrically identical room are usually used to add artificial reverberation to the dry sound sources [VPS12].

### 2.1.6 Psychoacoustics and critical band

Sound is meaningful when it is perceived by humans. Changes in the physical part of the sound (including frequency, intensity, phase, direction, etc.) may not always excite perceptual difference. This is mainly due to the limitation of human auditory system. Thus, in additional to objective evaluation, psychoacoustic experiments, which are in the form of subjective listening tests, are conducted to evaluate the performance of a sound reproduction system. The psychoacoustic experiments could help us better understand how the system actually performs in practice. The psychoacoustic experiments usually include localization of the sound sources, quality of the synthesized sound, quality of the reproduction system (e.g., loudspeakers and headphones), quality of the rendering methods, and so on.

One of the most important aspects of psychoacoustics is auditory masking, where a louder sound masks (fully or partially) a weaker sound when their spectra are close. Auditory masking happens in frequency domain (spectral masking) and time domain (temporal masking). The range of the spectra for spectral masking is defined based on its critical band, as per psychoacoustic experiments. According to Zwicker [Zwi61], 24 bands known as the Bark scale are defined to cover the frequency range of human listening. Each critical band has a center frequency with an approximate 1/3 octave bandwidth. The conversion from frequency (*f* in kHz) into the Bark can be described as:

$$Bark = 13\arctan\left(0.76f\right) + 3.5\arctan\left[\left(f/7.5\right)^2\right]. \tag{2.2}$$

Another example of critical band is the equivalent rectangular bandwidth (ERB) [Moo98], which is described as:

$$ERB = 24.7\left(4.37f + 1\right). \tag{2.3}$$

It is widely believed that the human auditory system is performing the critical band analysis of the incoming sound, in tasks like localization and separation of sound [Fle40], [Bre90]. Therefore, many audio processing systems are derived based on the concept of critical band (or its equivalents). For example, in binaural cue coding (BCC), 20 non-uniform filterbank based on ERB is employed [FaB03]. Furthermore, MPEG Surround employs a hybrid quadrature mirror filter (QMF) filterbanks [SBP04], [HPB05] that matches the frequency resolution of the human auditory system.

## 2.2 Spatial audio reproduction

Most of the time, we are not listening to real sound in a real environment, but are listening to a reproduced sound playback from a sound reproduction system. The reproduced sound is often referred to as virtual sound, as compared to real sound in natural listening.

### 2.2.1 A brief history of sound reproduction systems

Ever since the invention of phonograph by Edison in 1887, sound has been an essential part of telecommunication and media. The first stereo loudspeaker system was introduced by Blumlein [Blu31] in 1931, which has since then become the most popular sound reproduction system in homes. It takes humans some forty years to come up with new sound systems, including the first Dolby surround sound [ITU12] and Ambisonics invented by Gerzon [Ger73]. Though invented at almost the same time, these two systems undergo extremely different paths. The surround sound reproduction system, including 5.1, 7.1, as pushed by the film and music industry, has become the most prevalent home theater systems. The 5.1 surround sound system requires five speakers placed at center (0 ° azimuth), front left (-30 ° azimuth), front right (30 ° azimuth), surround left (-110 ° azimuth), surround right (110 ° azimuth), as well as a subwoofer. 7.1 surround sound system extends 2 surround speakers in 5.1 to 4 speakers. The multichannel surround sound system keeps evolving, from one layer to two layers (such as 9.1, 10.2) to even more layers (such as 22.2 [HMS11], Auro 3D). On the other hand, Ambisonics, despite its mathematical beauty (based on Huygens principle), was not well adopted in commercial systems. Nevertheless, the research on Ambisonics was never stopped in academia and it retrieves popularity in recent years, as shown in new MPEG-H standard [HHK14]. In 1993, another sound reproduction technique: wave field synthesis (WFS) was introduced [Ber88], [BVV93] and has found it presence in commercial products since 2001. Besides the development of loudspeaker systems, headphones are getting more and more widely used in recent years, which is mainly due to the rapid increase in mobile devices. The HRTFs are widely used in headphone based 3D sound reproduction [Beg00], [AlD11], [KHT15]. Today, we see a variety of sound reproduction systems in various applications, from cinema, home theater, to on the go. More and more 3D sound reproduction techniques have been studied and implemented in commercial products.

### 2.2.2 Representations of audio content

With the development of different recording and mixing techniques, different types of audio content representations have emerged in commercial market. Three main types are: channel-based, object-based, and transform-domain based.

Channel-based format has been the most common way of audio content representation. The channel-based format is playback-oriented as the channel signals can be directly fed to the loudspeakers based on the standard configuration (i.e., prescribed positions). Usually no additional processing (or very little processing like volume control) is required. This is because the channel-based format is usually the outcome of the sound mixing process (performed by the sound engineer). Besides the easy applicability for the playback, the channel-based format is also rather efficient at transmission and storage. The down side of channel-based format lies in its requirement to have a fixed playback system that corresponds to the number of channels. For example, stereo audio content requires the two speakers to be placed symmetrically at ±30 degrees azimuth on the two sides of the listener. 5.1 channel further adds a center and two rear channels, placed at 0 degree, and ±110 degrees azimuth, respectively, together with a subwoofer (low frequency effect channel). A matrix system that enables the downward compatibility of 5.1 is discussed in [ITU12]. Other channel-based formats include 7.1, 9.1, 10.2, and all the way up to 22.2 in three vertical layers. Adding height channels in channel-based audio is a fundamental improvement over horizontal loudspeaker setup to make the sound reproduction in full three dimensions. Commercial examples involving height channels on top of the conventional surround sound formats include Dolby ATMOS [Dol13] and Auro 3D [Aur15].

Object-based format is the most original format of a sound recording. Object-based format represents a sound scene using a combination of sound objects with the associated metadata [HHK15]. Sound objects are essentially individual sound sources. The metadata usually consists of two types: static metadata such as language, on/off time, etc., and dynamic metadata, such as position or direction, level, width or diffuseness of the sound object. Not all audio objects are separated. Those objects that collectively contribute to a fix sound effect or sound environment shall be grouped and regarded as one “larger” audio object. As a result, metadata can be specified for each audio object or a group of objects. The greatest benefit of object-based audio is that it can be rendered optimally for any arbitrary playback systems. Meanwhile, interactivity can be enabled, for example, changing to another language of speech, increasing the loudness of certain objects (e.g., speech level shall be higher for hearing impaired listeners), and adapting the position of the sound objects according to listener’s movement in virtual reality applications, etc. The object-based format is the best format in terms of reproduction flexibility and quality. However, two challenges that are found in practical implementation are high storage or transmission bandwidth, and high computation complexity for real-time rendering [MMS11]. Important aspects on implementation of audio objects coding and rendering were extensively studied in [Pot06]. Some work has been carried out by MPEG to achieve an efficient coding of sound objects based on perceptual features [HPK12].

The other type of audio representation is known as the transform-domain based format (or scene based, Ambisonics) [SWR13]. Transform-domain based format encodes the sound scene using orthogonal basis functions physically (using microphones) or digitally. In the reproduction, a corresponding rendering process is required. Though individual sound objects are not used, transform-domain based format can also achieve flexibility in reproduction for various playback setups, thanks to the sound field analysis and synthesis principle [Pol05]. However, the transform-domain representation is less common and less supported (e.g., recording/reproduction equipment) in industry than in academia.

### 2.2.3 Spatial audio reproduction techniques

These above-mentioned sound scene representations support different spatial audio reproduction techniques. Due to the nature of channel-based representations, conventional spatial audio reproduction techniques are straightforward as the audio signals of each channel are directly sent to drive the corresponding loudspeaker, resulting in stereo loudspeaker playback, 5.1, 7.1 surround sound playback, and stereo headphone playback. The simplicity of channel-based reproduction is achieved at the cost of strict requirement of exact match of the playback configuration. When there is a mismatch between the audio content and actual playback configuration, the performance is degraded, though simple down-mixing and up-mixing approaches can be applied.

In contrast to the channel-based format, the object-based and transform-domain based formats are more flexible in the playback and usually achieve better performance in spatial audio reproduction. Modern spatial audio reproduction techniques can usually be divided into two classes, namely, the physical reconstruction and perceptual reconstruction [HWZ14].

The first class of physical reconstruction aims at synthesizing the sound field in the listening area or point to be (approximately) equal to the desired sound field. Sound field synthesis is essentially based on the physical principle of synthesizing acoustic pressure using a weighted distribution of monopole sources [SWR13]. Two examples of sound field synthesis techniques are Ambisonics (4 channels) or high order Ambisonics (HOA, consists of more than 4 channels), and wave-field synthesis. Ambisonics or HOA decomposes (or encodes) a sound field using spherical harmonics, which results in the transform-domain based representation. With more channels, HOA can improve the spatial quality of reproduced sound field over Ambisonics. The best listening area in Ambisonics is usually limited to the central area of the sphere. In contrast, WFS can extend the sweet spot to a much wider area by approximating the propagation of the primary source using an array of secondary sources (loudspeakers). The loudspeaker driving signals are derived using a synthesis system function and source signals, which are expressed in object-based format. Compared to Ambisonics, WFS is not only well studied in academia, but also employed in some commercial sound systems such as IOSONO [Ios15] and Sonic Emotion [SoE15]. A major challenge in the physical reconstruction techniques is the requirement of large amount of loudspeakers and high computational complexity (especially in real-time rendering scenarios) [SWR13].

The other type of spatial audio reproduction techniques is based on the perceptual characteristics of human auditory system that our listening is not very sensitive. A good spatial audio reproduction is one that sounds good. The key idea of perceptual based spatial audio reproduction techniques is to have the sound captured by the listener’s eardrum to be perceptually close to the desired sound field. While the reproduced sound field does not always well match the desired sound field, perceptual based spatial audio reproduction techniques can greatly simplify the reproduction method. The simplest example of this category is the amplitude panning techniques, which are widely employed in sound mixing for stereo and surround sound [Hol08]. Techniques that extend amplitude panning to 3D space include the vector base amplitude panning [Pul97], [PuK08] and variants like distance based amplitude panning [LBH09]. Amplitude panning techniques are based on the ILD cues to recreate the correct direction of the sound sources. Similarly, time delay techniques that vary the ITD can also be used for spatial audio reproduction [SWR13].

However, the amplitude panning and time delay techniques are usually too simple to reproduce the correct impression of the sound sources with increased source width [MWC99], degraded location performance [ThP77], and coloration [PKV99]. A better approach is to consider the complete localization cues, which are included in the HRTFs [Beg00]. This approach is usually applied in headphone playback and it is known as binaural rendering. The key idea in binaural rendering is to consider the sound source propagation process (from sound source to listener’s eardrum) as a linear-time-invariant system and express this alteration of the source spectra due to human body as a filter. Therefore, the perception of any source from any direction can be recreated by convolving the sound source with the corresponding filters to obtain the driving signals that are sent to a compensated headphone (assumed transparent). The same concept of binaural rendering can also be applied in stereo loudspeakers, which is known as transaural rendering [Gar97]. Compared to binaural rendering, transaural rendering requires one additional process known as crosstalk cancellation. Multichannel extension of crosstalk cancellation and transaural system are discussed in [Gar00]. Crosstalk cancellation techniques are very sensitive of listener movement and small changes in sound environment, which limits the practical use of transaural systems. In contrast, binaural rendering over headphones are much more widely used. However, there remain two major challenges. The first one lies in the large variations of the HRTFs among different individuals. Use of non-individualized HRTF will degrade the localization accuracy. The other problem is the headphone itself, which is hardly completely transparent and headphone effect compensation varies not only from person to person, but also even after re-positioning. Due to the advent of virtual/augmented reality applications, many studies on HRTF individualization and headphone compensation are currently being carried out.

### 2.2.4 Spatial audio processing

In order to achieve the goal of efficient, flexible and immersive spatial audio reproduction, different spatial audio signal processing techniques are introduced. The aim of spatial audio processing (or coding) techniques is to complement the discrepancies in the above-mentioned spatial audio reproduction techniques, with a focus on channel-based signals and conventional multichannel reproduction techniques. Generally, these techniques are based on the concept of parametric spatial audio processing [KTT15] and exploit the perceptual characteristics of human auditory systems [Bla97]. In this

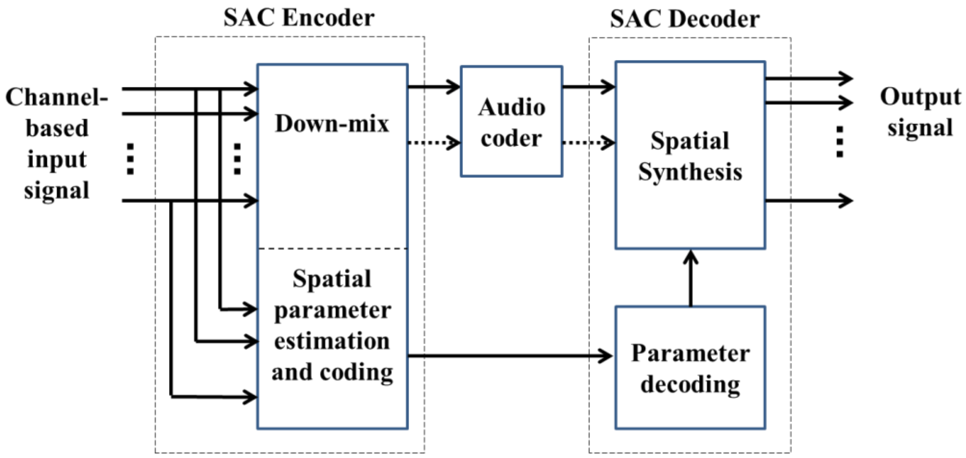


Figure 2.8 Basic concept of MPEG Surround (adapted from [HiD09])

part, we focus on five most widely studies frameworks, though other variations could also been found in the literature. Among the five frameworks discussed below, two of them deal with channel-based signals, one on the object-based signals, another one on the transform-domain based signals, and the latest one consolidates all three types of signals.

For channel-based signals, the objective of spatial audio processing is to achieve a more efficient representation that can reproduce perceptually plausible sound scenes. The most widely known framework comes from the MPEG audio group, known as MPEG Surround [HKB08], [BrF07], [HiD09]. In MPEG Surround, the multichannel signals go through a spatial analysis process and is represented using a down-mixed version together with the spatial parameters, as shown in Fig. 2.8. In the spatial synthesis, the original multichannel can be reconstructed using the spatial parameters in a way that the spatial perception is maximally preserved. Furthermore, other types of synthesized output include the direct down-mix for the playback with reduced number of loudspeakers and binaural signals for headphone playback [BrF07],

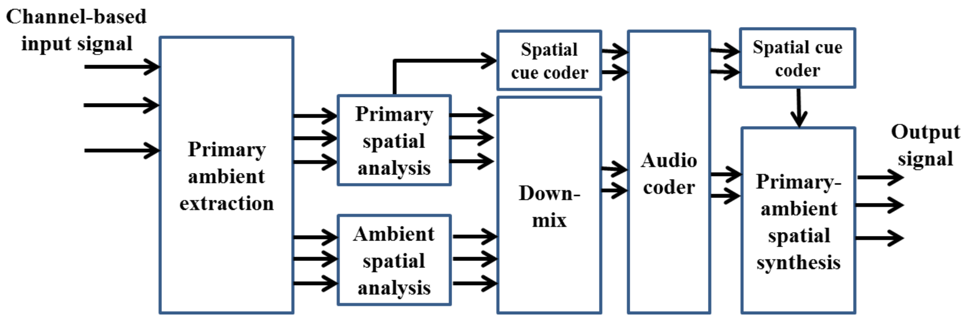


Figure 2.9 Block diagram of Spatial Audio Scene Coding (SASC) (adapted from [GoJ08])

[FaB03]. The details on the coding of the spatial parameters can be found in binaural cue coding (BCC) framework [FaB03], [BaF03].

Another framework that is also targeting channel-based audio is the so-called spatial audio scene coding (SASC) framework developed by Jot *et al.* [GoJ08], [JMG07]. [GoJ07a], [GoJ06a], [GoJ06b], [GoJ07c]. Compared to MPEG Surround, SASC was designed to address the pressing need to enhance sound reproduction over arbitrary playback configurations in loudspeakers and headphones. The detailed block diagram is shown in Fig. 2.9. In SASC, a sound scene is considered as a sum of primary and ambient components. Therefore, primary ambient extraction (or decomposition) is applied first, followed by the spatial analysis carried out independently for the primary and ambient components to obtain the spatial cues (i.e., localization information). In the spatial synthesis, the output is reconstructed using the primary and ambient components as well as the spatial cues. By taking into account the actual playback format, the reconstruction is able to fit any playback configuration. Due to this advantage of SASC, the primary ambient extraction work described in this thesis is essentially based on SASC. Details on the primary ambient extraction will be discussed throughout this thesis.

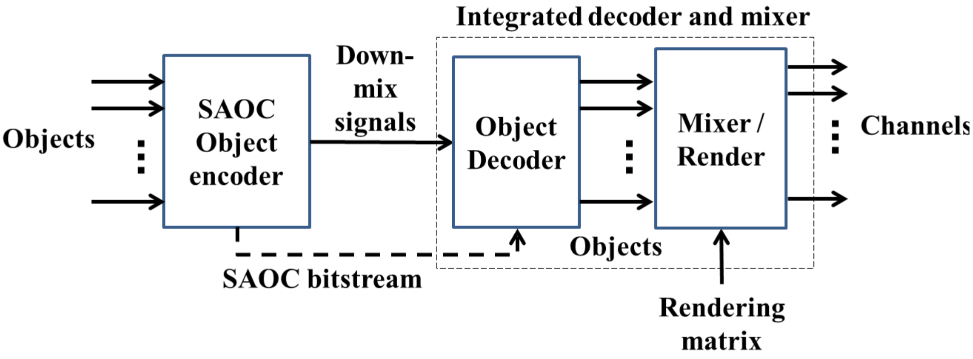


Figure 2.10 Basic concept of Spatial Audio Objects Coding (SAOC) (adapted from [HPK12])

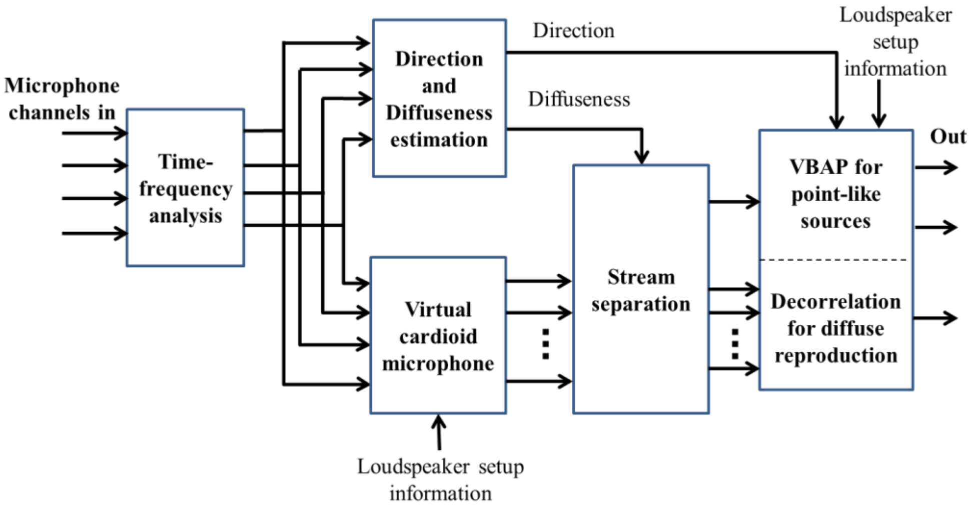

Microphone channels in
Time-frequency analysis
Direction and Diffuseness estimation
Direction
Diffuseness
Loudspeaker setup information
Out
VBAP for point-like sources
Decorrelation for diffuse reproduction
Stream separation
Virtual cardioid microphone
Loudspeaker setup information

Figure 2.11 Block diagram of Directional Audio Coding (DirAC) (adapted from [Pul07])

For object-based audio signals, MPEG introduced MPEG Spatial Audio Object Coding (SAOC) framework in 2012 [HPK12]. Similar to MPEG Surround, the MPEG SAOC aims to achieve an efficient representation of the object-based audio using a parametric approach that takes a down-mix of the audio objects in subband with supplementary inter-object information, as shown in Fig. 2.10. In the synthesis, the object decoder can be employed first before the render or can be combined into one block. Based on the information of the actual playback information, a rendering matrix is used to transform the audio

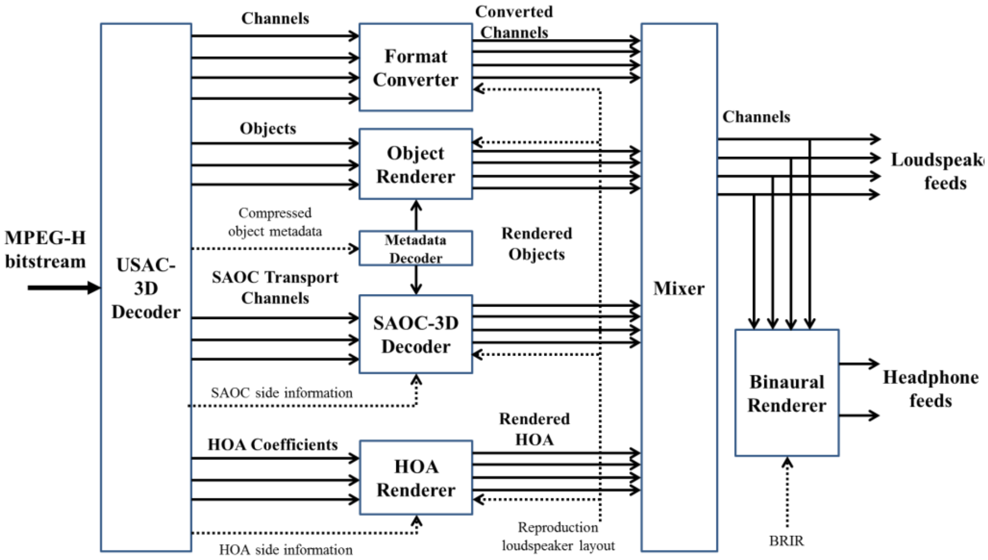


Figure 2.12 Overview of MPEG-H 3D audio coding (adapted from [HHK14])

objects into channel signals for playback. It shall be noted that SAOC can also achieve the flexibility and interactivity of the object-based format.

For transform-based signals, a parametric spatial audio processing framework known as Directional Audio Coding (DirAC) was introduced by Pulkki *et al.* [Pul07]. As shown in Fig. 2.11, DirAC analyzes the direction and diffuseness information of the microphone signals (in B-format) and then decomposes the microphone signals into two streams, namely, diffuse streams and non-diffuse streams. As shown in Fig. 2.11, these two streams go through different rendering process, where the non-diffuse streams is processing using VBAP with the loudspeaker setup information provided, and diffuse streams are decorrelated and played back over all the channels. The advantage of such decomposition, similar to SASC, is to be able to achieve flexible reproduction over arbitrary playback configurations.

Finally, MPEG-H [HHK14], [HHK15], introduced in 2014, aims to handle all three types of audio content (channel-based, object-based, and transform-domain based, presenting a complete solution for universal spatial audio reproduction. An overview of MPEG-H framework is depicted in Fig. 2.12. In the first step, the input bit stream is converted to their respective format using Unified Speech and Audio Coding (USAC)-3D core decoder. Next, different content types go through corresponding processing before they were mixed into channel signals that match the actual playback system layout. Finally, in the case of headphone playback, a binaural rendering of loudspeaker signals based on binaural room impulse response (BRIR) is employed. With such a unified framework, MPEG-H 3D audio can be employed for any content type, any playback configuration, while achieving the highest spatial audio quality.

### 2.2.5 Spatial audio evaluation

In spatial audio reproduction, the quality of the reproduced sound scene is usually evaluated on human perception. Perceptual evaluation of audio quality is often achieved using subjective listening tests [BeZ07]. Unlike conventional sound quality evaluation that usually only considers the timbre quality [GaS79] (e.g., evaluation of the quality of audio codec [ITU03]), the spatial quality is equally important in spatial audio evaluation [Rum02]. Referring to these two aspects of audio quality for spatial audio evaluation, Table 2.1 below summaries the various attributes that can be considered in each category [SWR13]. Among the timbre attributes, timbre fidelity, coloration, and distortion are more widely used. For spatial attributes, spatial fidelity,

Table 2.1 Attributes used for perceptual spatial audio evaluation (adapted from [SWR13])

| Category | Attribute | Description |
|---|---|---|
| Timbre | Timbral fidelity | Degree to which timbral attributes agree with reference |
| | coloration | Timbre-change considered as degradation of auditory event |
| | Timbre, color of tone | Timbre of auditory events |
| | Volume, richness | Perceived thickness |
| | Brightness | Perceived brightness or darkness |
| | Clarity | Absence of distortion, clean sound |
| | Distortion, artifacts | Noise or other disturbances in auditory event |
| Spatial | Spatial fidelity | Degree to which spatial attributes agree with the reference |
| | Spaciousness | Perceived size of environment |
| | Width | Individual or apparent source width |
| | Ensemble width | Width of the set of sources present in the scene |
| | Envelopment | Degree to which the auditory scene is enveloping the listener |
| | Distance | Sense of perspective in the auditory scene as a whole |
| | Externalization | Degree to which the auditory event is localized in- or outside of the head |
| | Localization | Measure of how well a spatial location can be attributed to an auditory event |
| | Robustness | Degree to which the position of an auditory event changes with listener movements |
| | Stability | Degree to which the location of an auditory event changes over time |

envelopment, distance, and localization are more important. Relative importance between the spatial quality and timbre quality is investigated in [RZK05], and it was summarized that the overall sound quality can be explained by the sum of 70% of the timbre quality and 30% of the spatial quality. Beyond these "perceptive domain" attributes as listed in Table 2.1, the highest level of perception is in the "affective domain" [BeZ07], where the listeners indicate their preference of the perceived sound scenes. In spatial audio reproduction where virtual audio is presented to the listener, an importance affective feature is the immersiveness. In other words, while listening to the reproduced sound, how much the listener feels as if him/her-self is inside the virtual scene (a.k.a., being there). Pursuing an immersive reproduction is the common aim of all spatial audio reproduction systems including the primary ambient extraction based spatial audio reproduction.

### 2.2.6 Summary and comparison of spatial audio reproduction

Table 2.2 summarizes the advantages, disadvantages and the status of the three audio formats discussed in this section. Furthermore, the spatial audio reproduction systems that correspond to each audio format are listed, together with the possible spatial audio processing techniques. It shall be noted that though classified in Table 2.2, there are still exceptions that link one audio format with other reproduction systems or processing techniques. For example, channel-based signals can also be employed in binaural/transaural rendering by considering one channel as one audio object with a fixed position. Ambisonics reproduction can also be extended to object-based audio by encoding the sound objects using spherical harmonics. It could be foreseen that with the advancement of semiconductor industry, the efficiency problem in object-based audio could be greatly alleviated and object-based audio will overtake channel-based to become the most commonly used audio format. Thus, advanced spatial audio reproduction system can essentially be employed in homes and mobile platforms. Nevertheless, there is still a need to ensure the compatible playback of channel-based audio signals due to the large amount of content available today.

Table 2.2 A summary of the characteristics of three audio content formats and their relationships with the spatial audio reproduction systems and processing techniques

| Audio content format | Channel-based | Object-based | Transform-domain based |
|---|---|---|---|
| Advantages | Easy to set up; no processing for the matched playback configurations | Flexible for arbitrary playback configuration; accurate sound image;<br>enable interactivity | Flexible for arbitrary playback configuration; full 3D sound image |
| Disadvantages | Difficult to fit in different playback configurations; 3D sound image limited | High transmission or storage; high computation complexity | Require a large number of speakers placed on the surface of a sphere |
| Status | Legacy audio format, still dominant | Emerging audio format; gaining popularity | Not well adopted commercially |
| Desired reproduction system | Stereo and multichannel surround sound system | Amplitude panning, WFS, binaural, transaural rendering | Ambisonics, and HOA |
| Typical spatial audio processing | MPEG Surround [HiD09], SASC [GoJ08] | SAOC [HPK12] | DirAC [Pul07] |

## 2.3 Prior work in primary ambient extraction

In this section, we will summarize various existing works on PAE and highlight how our works differ from those in the literature.

As discussed above, PAE is an integral part of spatial audio scene coding framework that considers the audio scene as a sum of the primary components and ambient components. The primary components are usually composed of directional point-like sources, whereas the ambient components are diffuse sound determined by the sound environment. The target audio format of PAE is channel-based signals. Therefore, we classify the PAE approaches based on the number of channels in the input signals: single channel (or mono), stereo, and multichannel. From another perspective, the complexity of the audio scenes affects the performance of PAE greatly. Based on the existing PAE work, the complexity of audio scenes can generally be classified into three levels, namely, basic, medium, and complex. The basic complexity level refers to the audio scene where there is usually one dominant source in the primary components, with its direction created using only amplitude panning techniques. More specific conditions for the basic level will be detailed in Chapter 3. The medium complexity level requires only the condition of one dominant sources, without restricting how its direction (using amplitude panning, delay, or HRTF, etc.) can be created. In the complex audio scene level, we consider multiple dominant sources in the primary components. The number of dominant sources in this case is also usually limited to 2-3 since it is impractical for listeners to concentrate on too many sources at one time and listeners would rather consider those sources as ambient components. Note that those PAE approaches that claimed to work in multiple sources using subband techniques, but without detailed study, will not be classified in the complex level category. From these two perspectives, we shall classify the existing PAE approaches into different categories, as summarized in Table 2.3.

With a glance of this table, it is observed that most of the PAE works are mainly focused on the stereo signals, due to the large amount of stereo content. There are some works carried out for multichannel signals, whereas very limited works are on single channel signals. This makes sense because dealing multichannel signals is much less challenging than dealing with single channel signals, where there is very limited information (especially the inter-channel relations). Next, we will summary the PAE work in each category.

Table 2.3 An overview of recent work in PAE

| No. of channels | Complexity of audio scenes | | |
|---|---|---|---|
| | Basic<br>(single source, only amplitude panning) | Medium<br>(single source) | Complex<br>(multiple sources) |
| Stereo | **Time frequency masking:**<br>[AvJ02], [AvJ04], [MGJ07], [Pul07]<br>**PCA:**<br>[IrA02], [BVM06], [MGJ07], [GoJ07b], [BaS07], [God08], [JHS10], [BJP12], [TaG12], [TGC12], [LBP14], [HTG14]<br>**Least-squares:** [Fal06], [Fal07], [JPL10], [FaB11], [HTG14], [UhH15]<br>**Ambient spectrum estimation**:<br>[HGT15a], [HGT15b]<br>**Others:** [BrS08], [MeF10], [Har11] | **LMS:** [UsB07]<br>**Shifted PCA:**<br>[HTG13]<br>**Time-shifting:**<br>[HGT15c] | **PCA:**<br>[DHT12],<br>[HGT14],<br>[HeG15], |
| Multichannel | **PCA:**<br>[GoJ07b]<br>**Others:**<br>[GoJ07a], [WaF11], [TGC12], [CCK14] | **ICA and time-frequency masking:**<br>[SAM06]<br>**Pairwise correlations:**<br>[TSW12]<br>**Others:**<br>[StM15] | **ICA:**<br>[HKO04] |
| Single | **NMF**: [UWH07]<br>**Neural network**: [UhP08] | | |

Notes:

1. Those papers that does not explicitly study and evaluate complex signals will be classified into the basic or medium complexity categories.
2. Blue color represents application papers, where no detailed study is carried out on PAE.
3. Red color represents our works, which are described in the following chapters of this thesis.

## 2.3.1 Stereo signals

PAE for stereo signals in the basic complexity category can be classified into four types: (i) time frequency masking, (ii) principal component analysis,

(iii) least-squares, and (iv) ambient spectrum estimation, as well as some other techniques.

One of the earliest works in primary or ambient extraction was from Avendano and Jot in 2002 [AvJ02]. In this work, a time-frequency masking approaches was proposed to extract ambient components $\hat{A}_c$ from stereo signals $X_c$, as

$$\hat{A}_c(m,l) = X_c(m,l)\Psi_A(m,l), \tag{2.4}$$

where $c$ denotes the channel index, and $0 \le \Psi_A(m,l) \le 1$ is the real-valued ambient mask at time-frequency bin $(m,l)$.The time-frequency regions that present high coherence correspond to stronger primary components, and low coherence time-frequency regions can be attributed to stronger ambient components [AvJ04]. Thus, they derived the ambient mask using a nonlinear function of the inter-channel coherence. Following works on time-frequency masking derives the ambient mask based on the characteristic that ambient components have equal level in the two channels of the stereo signal [MGJ07] or using diffuseness measured from B-format microphone recordings [Pul07].

Principal component analysis (PCA) has been the most widely studied PAE approach [IrA02], [BVM06], [MGJ07], [GoJ07b], [BaS07], [God08], [JHS10], [BJP12], [TaG12], [TGC12], [LBP14], [HTG14]. The key idea behind the PCA based PAE approach is to extract the principal component with the largest variance as the primary components (as the name suggests). Variants of PCA include the modified PCA that ensures uncorrelated ambience extraction [God08], enhanced post-scaling to restore the correct primary-to-ambient energy ratio [JHS10] and correct power of primary and ambient components

[BJP12]. In our work [HTG14], we derived a simplified solution for PCA and conducted a comprehensive objective evaluation of PCA, which leads us to the applications of PCA in PAE.

Least-squares is another type of commonly used PAE approaches [Fal06], [Fal07], [JPL10], [FaB11], [HTG14], [UhH15]. Based on the basic stereo signal model, least-squares algorithm derives the estimated primary and ambient components by minimizing the mean-square-error (MSE) of the estimation of these components [Fal06]. Several variants of least-squares have been proposed and studied in our work [HTG14]. Combining PCA with least-squares, we proposed a unified linear estimation framework for PAE [HTG14], where details of liner estimation based PAE can be found in Chapter 3. Furthermore, other least-squares variants were introduced to improve the spatial quality of the extracted primary and ambient components [JPL10], [UhH15].

To solve the problem of removing uncorrelated (undesired) ambient components from the extraction output, a new framework based on ambient spectrum estimation was introduced recently [HGT15a], [HGT15b]. Details on the ambient spectrum estimation approaches can be found in Chapter 4 of this thesis. Other PAE approaches that fall into this category include [BrS08] that derives an out-of-phase signal as ambient components; [MeF10] that considers ambient components as the sum of a common component and an independent component; and [Har11] that classifies various signal models for respective extraction.

In order to handle stereo signals that consist of primary components whose directions are created using time/phase differences (i.e., medium complexity), several works can be found in the literature. Usher and Benesty proposed an adaptive approach using normalized least-mean-squares (NLMS) to extract reverberation from stereo microphone recordings [UsB07]. However, this adaptive approach cannot always yield a good performance in a short time. In contrast, our proposed shifted PCA [HTG13] and extended time-shifting technique [HGT15c] is much simpler in solving this problem. Details on this approach can be found in Chapter 5 of this thesis.

With respect to stereo signals with multiple sources, there is less work reported in the literature of PAE. One prior work by Dong *et al.* applied PCA in polar coordinates to reduce the coding noise of stereo signals for multiple source case [DHT12]. However, the extraction performance was not studied. To fill in this gap, we conducted two works that studied PCA with different frequency partitioning methods in frequency domain [HGT14], and PCA with multiple time shifts in time domain [HeG15]. Details are described in Chapter 6 of this thesis.

### 2.3.2 Multichannel signals

Besides the extensive study on PAE for stereo signals, PAE on multichannel signals is less well studied. PCA was originally proposed to work for multichannel signals with only one dominant amplitude-panned source in [GoJ07b]. There are several works [GoJ07a], [WaF11], [TGC12], [CCK14] that only briefly mention the idea for multichannel PAE without in-depth studies. For other multichannel signals with one dominant source, independent component analysis (ICA) can be combined with time-frequency masking to extract the dominant sources [Sam06]. Another approach that was extended from [AvJ04], achieves primary ambient extraction using a system of pairwise correlation. Recently, Stefanakis introduced W-disjoint orthogonality (WDO) and PCA based foreground suppression techniques in multichannel microphone recordings [StM15]. In the case of multiple sources in multichannel signals, blind source separation techniques can be employed for the purpose of primary ambient extraction. When the number of dominant sources is equal to or less than the number of channels (as it is the case for PAE), ICA is a common technique [HKO04]. Compared to stereo signals, PAE with multichannel signals is in fact easier to solve since there are more information available. Moreover, PAE approaches based on stereo signals can be extended to multichannel signals. Some discussions on this topic can be found in [HeG15b].

### 2.3.3 Single channel signals

In contrast to stereo and multichannel signals, PAE with single channel signals is quite challenging due to the limited amount of information available. A critical problem in the single channel case is that how primary and ambient components can be defined and characterized since there are no inter-channel cues. Nevertheless, two works from Uhle shed some light on solving such a problem. In [UWH07], it is considered that ambient components exhibit a less repetitive and constructive spectra structure than primary components. Therefore, when applying non-negative matrix factorization (NMF) on the single channel signal, primary components are better explained and factorized, and the residue can thus be considered as ambient components. However, the NMF method suffers from high computational complexity and latency. To avoid this problem, Uhle and Paul introduced a supervised learning approach for ambient extraction from single channel signals [UhP08], where a neural network is trained to obtain an ambient spectra mask. Subjective listening tests in [UhP08] validated the improved perceptual quality of the up-mix systems employing these PAE approaches.

## 2.4 Conclusions

In this chapter, we reviewed the basics on spatial hearing of humans, where the binaural cues are very important. Various aspects on spatial audio reproduction are further discussed, which begins with the history of spatial audio reproduction. Three types of audio representations are explained and found to be deterministic in choosing the appropriate spatial audio reproduction techniques as well as spatial audio processing techniques. With the aim to improve the reproduction flexibility and quality of channel-based audio, primary ambient extraction is introduced. Various existing PAE approaches are classified and reviewed in this section. The details on our work to improve the performance of PAE in various circumstances as well as applying PAE in spatial audio reproduction will be presented in the following chapters.

# Chapter 3

# Linear Estimation based Primary Ambient Extraction

In this chapter[1], we focus on primary ambient extraction approaches that can be considered in a unified linear estimation framework, with the assumption that the primary and ambient components are linearly mixed in the stereo signal model [GoJ07b]. Based on the linear estimation, PCA and least-squares (LS) are designed to minimize the correlation between the primary and ambient components and the extraction error, respectively. Our analysis reveals that the extraction error consists of three error components, namely, distortion, interference, and leakage. Distortion relates to the amount of amplitude scaling of the extracted primary (or ambient) component as compared to the true primary (or ambient) component. Interference measures the amount of uncorrelated primary (or ambient) component that is extracted from the stereo signal. Leakage measures the amount of undesired ambient (or primary) components in the extracted primary (or ambient) component. The characteristics of these three error components indicate that the leakage and distortion are perceptually more noticeable than interference in most of the applications. Taking this into consideration, different solutions for PAE can be obtained by minimizing these components. By minimizing the leakage and

[1] The work reported in this chapter is an extension from the author's Journal paper [HTG14] published in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, February 2014 issue.

distortion, two variant LS approaches, namely, minimum leakage LS (MLLS) and minimum distortion LS (MDLS) are proposed in this chapter, respectively. This derivation is followed by a comparative study on the performance of these PAE approaches. Based on our observations of this comparison, another approach referred to as the adjustable LS (ALS) is proposed, which offers adjustable error performance between the distortion and extraction error.

The rest of this chapter is organized as follows. In Section 3.1, we review the stereo signal model, and the key assumptions of this signal model. Subsequently, the linear estimation framework of PAE and two groups of performance measures are presented in Section 3.2. Section 3.3 discusses several approaches applied in PAE. Section 3.4 presents our discussion on the simulation results, which leads to our recommendations in applying the PAE approaches in different applications. Section 3.5 concludes this work.

## 3.1 Stereo signal model

Sound scenes in moving pictures and video games usually comprise several point-like sound sources (or primary component) and the environmental ambient sound (or ambient component) [Hol08]. PAE aims to separate the primary component from the ambient component based on their perceptual spatial features. The perceptual spatial features can be characterized by the inter-channel relationships, including inter-channel time difference (ICTD), inter-channel level difference (ICLD), and inter-channel cross-correlation coefficient (ICC) [BaF03]. Since the number of primary sources is usually unknown and might be varying, a common practice in spatial audio processing is to convert the signals into time-frequency domain using short-time Fourier transform (STFT) [AvJ04], [GoJ07b], [GoJ07a], [Pul07], [Fal06], [MGJ07], [FaB03] or subband via filter banks like hybrid quadrature mirror filter banks [BHK07]. For each frequency band or subband, it is generally assumed that the primary component of the input signal is composed of only one dominant source [AvJ04], [GoJ07b], [Fal06], [MGJ07]. Denoting the *b*th subband of input stereo signals (denoted by the subscript 0, and 1) at time frame index *m* as $\mathbf{x}_0[m,b]=\left[x_0(mN,b),\ldots,x_0(mN+N-1,b)\right]^T$, and $\mathbf{x}_1[m,b]=\left[x_1(mN,b),\ldots,x_1(mN+N-1,b)\right]^T$, where $N$ is the length of one frame. PAE is carried out in each subband of each frame independently, and the extracted primary and ambient components are combined via inverse STFT or synthesis filter banks. Here, a non-overlapping case of the signal frames is considered, though extension to the overlapping case is quite straightforward. In this chapter, the time-domain stereo signal model is expressed as:

$$\begin{aligned}\mathbf{x}_0[m,b]&=\mathbf{p}_0[m,b]+\mathbf{a}_0[m,b],\\ \mathbf{x}_1[m,b]&=\mathbf{p}_1[m,b]+\mathbf{a}_1[m,b],\end{aligned} \tag{3.1}$$

where $\mathbf{p}_0,\mathbf{p}_1$ and $\mathbf{a}_0,\mathbf{a}_1$ are the primary and ambient components in the two channels of the stereo signal, respectively. Since the subband of the input signal is generally used in the analysis of PAE approaches, the indices $[m,b]$ are omitted for brevity. Fig. 3.1 shows the stereo signal model and the input and output of PAE.

The stereo signal model also assumes that the primary components in the two channels are correlated, whereas the ambient components in the two channels are uncorrelated. The correlation coefficient between the two channels

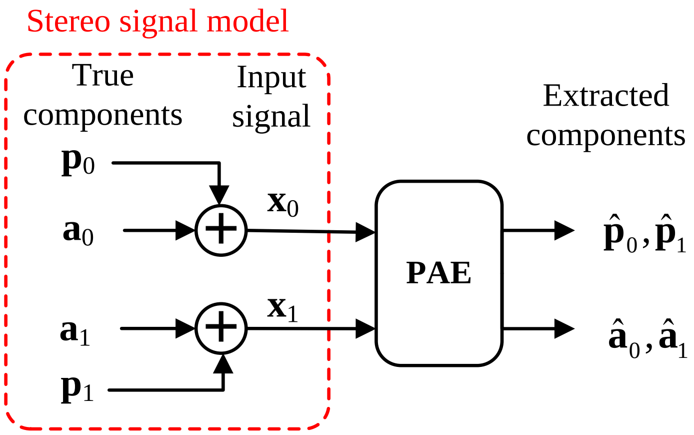


Figure 3.1 Extraction of the primary and ambient components using PAE, where $\mathbf{x}_0, \mathbf{x}_1$ are the input stereo signals; $\mathbf{p}_0, \mathbf{p}_1$ and $\mathbf{a}_0, \mathbf{a}_1$ are the true primary and ambient components, respectively; $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1$ and $\hat{\mathbf{a}}_0, \hat{\mathbf{a}}_1$ are the extracted primary and ambient components, respectively.

of the signal $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as $\phi_{ij}(\tau) = r_{ij}(\tau) \big/ \sqrt{r_{ii}(0) r_{jj}(0)}$, where $r_{ij}(\tau) = \sum_{n=mN}^{mN+N-1} \left[ x_i(n,b) x_j(n+\tau,b) \right]$ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ at lag $\tau$. Two signals are considered correlated when $\max_{\tau} \left| \phi_{ij}(\tau) \right| = 1$; uncorrelated when $\max_{\tau} \left| \phi_{ij}(\tau) \right| = 0$; and partially correlated when $0 < \max_{\tau} \left| \phi_{ij}(\tau) \right| < 1$.

Correlated primary component in the stereo signal can be described by one of the following conditions [Bla97]:

(i) amplitude panned, i.e., $\mathbf{p}_1 = k\mathbf{p}_0$, where $k$ is referred to as the primary panning factor (PPF);

(ii) time shifted, i.e., $p_1(n) = p_0(n + \tau_0)$, where $p_1(n)$ is the $n$th sample of $\mathbf{p}_1$ and $\tau_0$ is the ICTD (in samples); and

(iii) amplitude panned and time shifted, i.e., $p_1(n) = kp_0(n+\tau_0)$.

In this signal model, we only consider the primary component to be amplitude panned by PPF *k* [GoJ07b], [Fal06], [MGJ07]. This amplitude panned primary component is commonly found in stereo recordings using coincident techniques and sound mixes using conventional amplitude panning techniques [Hol08]. For an ambient component that consists of environmental sound, it is usually considered to be uncorrelated with the primary component [UsB07], [KDN09], [HGC09]. The ambient component in the two channels is also assumed to be uncorrelated and relatively balanced in terms of power, considering the diffuseness of ambient component. To quantify the power difference between the primary and ambient components, we introduce the primary power ratio (PPR) $\gamma$, which is defined as the ratio of total primary power to total signal power in two channels:

$$\gamma = \left(P_{\mathbf{p}_0} + P_{\mathbf{p}_1}\right) \Big/ \left(P_{\mathbf{x}_0} + P_{\mathbf{x}_1}\right), \tag{3.2}$$

where $P_{(.)}$ denotes the mean square power of the signal in the subscript. From (3.2), it is clear that $\gamma$ ranges from zero to one. Summarizing the assumptions for the stereo signal model, we have

$$\mathbf{p}_1 = k\mathbf{p}_0,\ \mathbf{a}_0 \perp \mathbf{a}_1,\ \mathbf{p}_i \perp \mathbf{a}_j, \forall i, j \in \{0,1\}, \tag{3.3}$$

$$P_{\mathbf{p}_1} = k^2 P_{\mathbf{p}_0},\ \ P_{\mathbf{a}_1} = P_{\mathbf{a}_0}, \tag{3.4}$$

where $\perp$ represents that two signals are uncorrelated.

Given any stereo input signal that fulfills the above conditions, the relationships between the auto-correlations $r_{00}, r_{11}$ and cross-correlation $r_{01}$ at zero-lag and the power of these components can be expressed as

$$r_{00} = \mathbf{x}_0{}^H \mathbf{x}_0 = NP_{\mathbf{x}_0} = N\left(P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right), \tag{3.5}$$

$$r_{11} = \mathbf{x}_1{}^H \mathbf{x}_1 = NP_{\mathbf{x}_1} = N\left(k^2 P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right), \tag{3.6}$$

$$r_{01} = \mathbf{x}_0{}^H \mathbf{x}_1 = \mathbf{p}_0{}^H \mathbf{p}_1 = NkP_{\mathbf{p}_0}, \tag{3.7}$$

where $H$ is the Hermitian transpose operator. From (3.5)-(3.7), the PPF($k$) and PPR($\gamma$) of the stereo signal are derived as:

$$k = \frac{r_{11} - r_{00}}{2r_{01}} + \sqrt{\left(\frac{r_{11} - r_{00}}{2r_{01}}\right)^2 + 1}, \tag{3.8}$$

$$\gamma = \frac{2r_{01} + \left(r_{11} - r_{00}\right)k}{\left(r_{11} + r_{00}\right)k}. \tag{3.9}$$

The primary component is panned to channel 1 for $k > 1$ and to channel 0 for $k < 1.$ In spatial audio, the PPF is considered as the square root of ICLD. Only the primary or ambient component is found in the stereo signal for $\gamma = 1$ or $\gamma = 0,$ respectively. In other words, the primary component becomes more prominent as $\gamma$ increases. In the following sections, we shall see that PPF and PPR are useful parameters for the extraction of the primary and ambient components, as well as to evaluate the performance of the PAE approaches.

## 3.2 Linear estimation framework and performance measures

In this chapter, we examine the blind extraction of primary and ambient components from a stereo input signal. Inspired by the mixing signal model given in (3.1), we address the PAE problem based on a linear estimation framework, where the primary and ambient components are estimated as weighted sums of the stereo signals in two channels. Thus, the extracted primary and ambient components are expressed as

$$\begin{bmatrix} \hat{\mathbf{p}}_0{}^T \\ \hat{\mathbf{p}}_1{}^T \\ \hat{\mathbf{a}}_0{}^T \\ \hat{\mathbf{a}}_1{}^T \end{bmatrix} = \begin{bmatrix} w_{\mathrm{P}0,0} & w_{\mathrm{P}0,1} \\ w_{\mathrm{P}1,0} & w_{\mathrm{P}1,1} \\ w_{\mathrm{A}0,0} & w_{\mathrm{A}0,1} \\ w_{\mathrm{A}1,0} & w_{\mathrm{A}1,1} \end{bmatrix} \begin{bmatrix} \mathbf{x}_0{}^T \\ \mathbf{x}_1{}^T \end{bmatrix} = \mathbf{W} \begin{bmatrix} \mathbf{x}_0{}^T \\ \mathbf{x}_1{}^T \end{bmatrix}, \tag{3.10}$$

where $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1$ and $\hat{\mathbf{a}}_0, \hat{\mathbf{a}}_1$ are the extracted primary and ambient components in the two channels, respectively; $T$ is the transpose operator; and $w_{(\cdot)}$ is the estimated weight of the extracted component, where the first subscript refers to the output signal, with “P” or “A” denotes the primary or ambient component, respectively, and the following number denotes the channel of the extracted component; and the second subscript denotes the channel of the input signal. Using this formulation, the PAE problem is simplified to the estimation of weighting matrix **W**.

Based on the weighting matrix **W**, we shall introduce two groups of measures to evaluate the objective performance of the linear estimation based PAE approaches. The first group measures the extraction accuracy of the primary and ambient components, whereas the second group examines the accuracy of the localization cues for the primary component and diffuseness for the ambient component.

### 3.2.1 Group 1: measures for extraction accuracy

In [MGJ07], the extraction accuracy of PAE approaches is evaluated by the similarity measures based on the cross-correlation coefficient between the extracted and true components. While these measures quantify the overall performance of the PAE approaches, these measures are unable to provide in-depth insights on possible causes for the performance degradation. In this subsection, we shall analyze the components that form the extraction error of the PAE approaches, and propose four performance measures to quantify the extraction error. A similar decomposition on the error components with corresponding measures can be found in source separation [VGF06] and speech enhancement [HaR09], where different beamformers are derived using different error components as criteria. In the following, we discuss the error measures for the primary component, followed by the ambient component.

Considering the error between the extracted primary component $\hat{\mathbf{p}}_0$ and its true component $\mathbf{p}_0$, we have

$$\mathbf{e}_{\mathrm{P}} = \hat{\mathbf{p}}_0 - \mathbf{p}_0. \tag{3.11}$$

Based on (3.11), we compute the error-to-signal ratio (ESR) for the primary component, which is defined as the ratio of the power of the extraction error to the power of the true primary component:

$$\mathrm{ESR}_{\mathrm{P}} = P_{\mathbf{e}_{\mathrm{P}}} \big/ P_{\mathbf{p}_0}. \tag{3.12}$$

Note that the ESR is equivalent to the normalized mean-squar-error (NMSE).

Based on (3.10), $\hat{\mathbf{p}}_0$ can be expressed as

$$\hat{\mathbf{p}}_0 = w_{\mathrm{P}0,0}\mathbf{x}_0 + w_{\mathrm{P}0,1}\mathbf{x}_1. \tag{3.13}$$

According to the assumptions stated in (3.3) and substituting (3.1) into (3.13), we have

$$\begin{aligned}\hat{\mathbf{p}}_0 &= \left(w_{\mathrm{P}0,0}\mathbf{p}_0 + w_{\mathrm{P}0,1}\mathbf{p}_1\right) + \left(w_{\mathrm{P}0,0}\mathbf{a}_0 + w_{\mathrm{P}0,1}\mathbf{a}_1\right) \\ &= w_{\mathrm{P}0}\mathbf{p}_0 + \left(w_{\mathrm{P}0,0}\mathbf{a}_0 + w_{\mathrm{P}0,1}\mathbf{a}_1\right) \\ &= \mathbf{p}_0 + \left(w_{\mathrm{P}0} - 1\right)\mathbf{p}_0 + \left(w_{\mathrm{P}0,0}\mathbf{a}_0 + w_{\mathrm{P}0,1}\mathbf{a}_1\right),\end{aligned} \tag{3.14}$$

where $w_{\mathrm{P}0} = w_{\mathrm{P}0,0} + kw_{\mathrm{P}0,1}$ is the weight of $\mathbf{p}_0$ in the extracted component $\hat{\mathbf{p}}_0$. Substituting (3.14) into (3.11), the extraction error becomes

$$\mathbf{e}_{\mathrm{P}} = \left(w_{\mathrm{P}0} - 1\right)\mathbf{p}_0 + \left(w_{\mathrm{P}0,\,0}\mathbf{a}_0 + w_{\mathrm{P}0,\,1}\mathbf{a}_1\right) = Dist_{\mathrm{P}} + Leak_{\mathrm{P}}, \tag{3.15}$$

where $Dist_{\mathrm{P}} = \left(w_{\mathrm{P}0} - 1\right)\mathbf{p}_0$ and $Leak_{\mathrm{P}} = w_{\mathrm{P}0,0}\mathbf{a}_0 + w_{\mathrm{P}0,1}\mathbf{a}_1$ are the distortion and leakage in the extraction error, respectively. The distortion comes from the extraction weight $w_{\mathrm{P}0}$, which fluctuates from frame to frame, causing variations in sound timbre or level. We consider the primary component to be completely extracted and hence distortionless when $w_{\mathrm{P}0} = 1$. On the other hand, the leakage of the extracted primary component $Leak_{\mathrm{P}}$ originates from the true ambient components $\mathbf{a}_0$ and $\mathbf{a}_1$ of the stereo signal. We consider the ratios of the distortion and leakage power to the power of true primary component, as the distortion-to-signal ratio (DSR) [BCH11] and the leakage-to-signal ratio (LSR), respectively:

$$\begin{aligned} \mathrm{DSR}_{\mathrm{P}} &= P_{Dist_{\mathrm{P}}} \big/ P_{\mathbf{p}_0}, \\ \mathrm{LSR}_{\mathrm{P}} &= P_{Leak_{\mathrm{P}}} \big/ P_{\mathbf{p}_0}. \end{aligned} \tag{3.16}$$

Similar performance measures are also obtained to quantify the ambient extraction error. Based on (3.10), the extraction error of the ambient component is rewritten as

$$\begin{aligned} \mathbf{e}_{\mathrm{A}} &= \hat{\mathbf{a}}_0 - \mathbf{a}_0 \\ &= \left(w_{\mathrm{A}0,0} - 1\right)\mathbf{a}_0 + w_{\mathrm{A}0,1}\mathbf{a}_1 + \left(w_{\mathrm{A}0,0}\mathbf{p}_0 + w_{\mathrm{A}0,1}\mathbf{p}_1\right) \\ &= Dist_{\mathrm{A}} + Intf_{\mathrm{A}} + Leak_{\mathrm{A}}, \end{aligned} \tag{3.17}$$

where the three components in $\mathbf{e}_{\mathrm{A}}$: $Dist_{\mathrm{A}} = \left(w_{\mathrm{A}0,0} - 1\right)\mathbf{a}_0$, $Intf_{\mathrm{A}} = w_{\mathrm{A}0,1}\mathbf{a}_1$, and $Leak_{\mathrm{A}} = w_{\mathrm{A}0,0}\mathbf{p}_0 + w_{\mathrm{A}0,1}\mathbf{p}_1$ are the distortion, interference, and leakage, respectively. Similar to primary extraction, the distortion comes from the extraction weight $w_{\mathrm{A}0,0}$, and the ambient component is considered to be distortionless when $w_{\mathrm{A}0,0} = 1$. Interference $Intf_{\mathrm{A}}$ is produced by the uncorrelated ambient component in the counterpart channel $\mathbf{a}_1$, whereas the leakage of the extracted ambient component $Leak_{\mathrm{A}}$ originates from true primary components $\mathbf{p}_0$ and $\mathbf{p}_1$. The extraction error of the ambient component and its three error components are quantified by the ratios of their power to the power of true ambient component, as ESR, DSR, interference-to-signal ratio (ISR), and LSR, which are given as

$$\begin{aligned} \mathrm{ESR}_{\mathrm{A}} &= P_{\mathbf{e}_{\mathrm{A}}} \big/ P_{\mathbf{a}_0} , \\ \mathrm{DSR}_{\mathrm{A}} &= P_{Dist_{\mathrm{A}}} \big/ P_{\mathbf{a}_0} , \\ \mathrm{ISR}_{\mathrm{A}} &= P_{Intf_{\mathrm{A}}} \big/ P_{\mathbf{a}_0} , \\ \mathrm{LSR}_{\mathrm{A}} &= P_{Leak_{\mathrm{A}}} \big/ P_{\mathbf{a}_0} . \end{aligned} \tag{3.18}$$

Comparing the measures of extraction error for the primary and ambient components, we find that no interference is found in the extracted primary component due to the unity correlation of the primary component. For both the primary and ambient components, ESR quantifies the overall error of the extracted component, and DSR, ISR, LSR provide detailed information on the extraction performance. In particular, LSR corresponds to the perceptual difference between the primary and ambient components. Both the interference and distortion in the extracted primary (or ambient) component come from the differences in this primary (or ambient) component between the two channels, hence they often exhibit some perceptual similarity with the true primary (or ambient) component. However, leakage solely comes from the ambient (or primary) component. Consequently, leakage is much more noticeable and undesirable than interference and distortion. Thus, we consider LSR to be the most important measure among DSR, ISR, and LSR for many applications. Nevertheless, more emphasis should be placed on DSR when sound timbre or amplitude is of high importance.

### 3.2.2 Group 2: measures for spatial accuracy

In the second group of measures, we consider the spatial accuracy of the extracted primary component based on three widely used spatial cues, namely, ICC, ICTD, and ICLD. These cues are used to evaluate the sound localization accuracy of the extracted primary component [Rum01], [Bla97]. There have been many studies to estimate ICTD after the coincidence model proposed by Jeffress (see [Jef48], [JSY98] and references therein). Based on the Jeffress model [Jef48], the ICC at different time lags is calculated and the lag index corresponds to the maximum ICC is the estimated ICTD. ICLD is obtained by taking the ratio of the power between the signals in two channels.

As the ambient component is assumed to be uncorrelated and balanced in the two channels, ICC and ICLD are selected as the measures to determine the diffuseness of the extracted ambient component [AnC09]. A better extraction of the ambient component is obtained when the ICC and ICLD of the extracted ambient component are closer to zero and one, respectively.

## 3.3 Linear estimation based PAE approaches

Following the discussions in Section 3.2, we shall derive the solutions for PAE approaches using linear estimation. These solutions are obtained by optimizing the weights in **W** for different criteria in PAE, including the minimization of the correlation between primary and ambient components, and the minimization of different error components. In this section, an analytic study and comparison of five linear estimation based PAE approaches including three proposed approaches will be presented.

### 3.3.1 PAE using principal component analysis

Principal component analysis is a widely used method in multivariate analysis [Jol02]. The central idea of PCA is to linearly transform its input sequence into orthogonal principal components with descending variances. PCA was first introduced to solve the PAE problem in [IrA02]. In general, the primary component is assumed to possess more power than the ambient component, i.e., $\gamma > 0.5$. Hence, it is a common practice to relate the larger eigenvalue to the primary component and the smaller eigenvalue to the ambient component. Based on the stereo signal model, PAE using PCA decomposition can be mathematically described as [MGJ07]:

$$\begin{aligned} &\mathbf{u}_{\mathrm{P}} = \arg\max_{\mathbf{u}_{\mathrm{P}}} \left( \left\| {\mathbf{u}_{\mathrm{P}}}^T \mathbf{x}_0 \right\|^2 + \left\| {\mathbf{u}_{\mathrm{P}}}^T \mathbf{x}_1 \right\|^2 \right), \\ &\mathbf{u}_{\mathrm{A}} = \arg\min_{\mathbf{u}_{\mathrm{A}}} \left( \left\| {\mathbf{u}_{\mathrm{A}}}^T \mathbf{x}_0 \right\|^2 + \left\| {\mathbf{u}_{\mathrm{A}}}^T \mathbf{x}_1 \right\|^2 \right), \\ &\text{s.t. } \mathbf{u}_{\mathrm{P}} \perp \mathbf{u}_{\mathrm{A}},\ \left\| \mathbf{u}_{\mathrm{P}} \right\| = \left\| \mathbf{u}_{\mathrm{A}} \right\| = 1, \end{aligned} \tag{3.19}$$

where $\mathbf{u}_{\mathrm{P}}$ and $\mathbf{u}_{\mathrm{A}}$ are the primary and ambient basis vectors, respectively. As depicted in Fig. 3.2, $\mathbf{u}_{\mathrm{P}}$ and $\mathbf{u}_{\mathrm{A}}$ maximizes and minimizes the total projection energy of the input signal vectors, respectively. The solution to (3.19) can be obtained by eigenvalue decomposition of the input covariance matrix [GoJ07b].

First, we find the larger eigenvalue and its corresponding primary basis

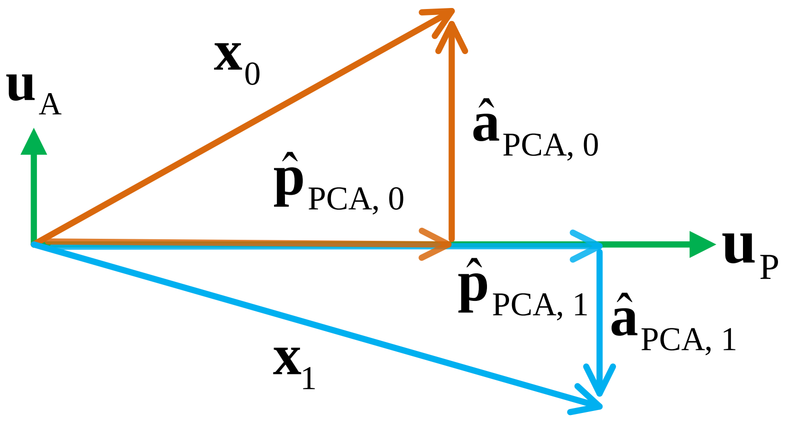

$\mathbf{u}_A$
$\mathbf{x}_0$
$\hat{\mathbf{a}}_{\text{PCA, 0}}$
$\hat{\mathbf{p}}_{\text{PCA, 0}}$
$\mathbf{u}_P$
$\hat{\mathbf{p}}_{\text{PCA, 1}}$
$\hat{\mathbf{a}}_{\text{PCA, 1}}$
$\mathbf{x}_1$

Figure 3.2 A geometric representation of PCA based PAE

vector [GoJ07b], [MGJ07] as

$$\lambda_{\text{P}} = 0.5\left[ r_{00} + r_{11} + \sqrt{\left(r_{00} - r_{11}\right)^2 + 4r_{01}^{\ 2}} \right], \tag{3.20}$$

$$\mathbf{u}_{\text{P}} = r_{01}\mathbf{x}_0 + \left(\lambda_{\text{P}} - r_{00}\right)\mathbf{x}_1. \tag{3.21}$$

Next, we compute the extracted primary components as

$$\begin{aligned} \hat{\mathbf{p}}_{\text{PCA},0} &= \frac{\mathbf{u}_{\text{P}}^{\ H}\mathbf{x}_0}{\mathbf{u}_{\text{P}}^{\ H}\mathbf{u}_{\text{P}}}\mathbf{u}_{\text{P}}, \\ \hat{\mathbf{p}}_{\text{PCA},1} &= \frac{\mathbf{u}_{\text{P}}^{\ H}\mathbf{x}_1}{\mathbf{u}_{\text{P}}^{\ H}\mathbf{u}_{\text{P}}}\mathbf{u}_{\text{P}}. \end{aligned} \tag{3.22}$$

However, the above solution of the extracted primary components is too complex in terms of its computation. Using (3.5)-(3.9), we can simplify the expressions for the extracted primary components using PCA as follows (detailed derivation can be found in Appendix A):

$$\begin{aligned} \hat{\mathbf{p}}_{\text{PCA},0} &= \frac{1}{1+k^2}\left(\mathbf{x}_0 + k\mathbf{x}_1\right), \\ \hat{\mathbf{p}}_{\text{PCA},1} &= \frac{k}{1+k^2}\left(\mathbf{x}_0 + k\mathbf{x}_1\right) = k\hat{\mathbf{p}}_{\text{PCA},0}. \end{aligned} \tag{3.23}$$

Similarly, the extracted ambient components are obtained as

$$\begin{aligned}\hat{\mathbf{a}}_{\text{PCA},0} &= \frac{k}{1+k^2}\left(k\mathbf{x}_0 - \mathbf{x}_1\right),\\ \hat{\mathbf{a}}_{\text{PCA},1} &= -\frac{1}{1+k^2}\left(k\mathbf{x}_0 - \mathbf{x}_1\right) = -\frac{1}{k}\hat{\mathbf{a}}_{\text{PCA},0}.\end{aligned} \tag{3.24}$$

From (3.23)-(3.24), we observe that the weights for the extracted primary and ambient components are solely dependent on the PPF *k*. Between the two channels, the primary components are amplitude panned by a factor of *k*, whereas the ambient components are negatively correlated and panned to the opposite direction of the primary components, as indicated by the scaling factor $-1/k$. Clearly, the assumption of the uncorrelated ambient components in the stereo signal model does not hold considering the ambient components extracted using PCA. This drawback is inevitable in PCA since the ambient components in two channels are obtained from the same basis vector. Nevertheless, as the primary and ambient components are derived from different basis vectors, the assumption that the primary components are uncorrelated with the ambient components is well satisfied in PCA.

By substituting the true primary and ambient components into (3.23) and (3.24), we have

$$\begin{aligned}\hat{\mathbf{p}}_{\text{PCA},0} &= \mathbf{p}_0 + \frac{1}{1+k^2}\left(\mathbf{a}_0 + k\mathbf{a}_1\right),\\ \hat{\mathbf{p}}_{\text{PCA},1} &= \mathbf{p}_1 + \frac{k}{1+k^2}\left(\mathbf{a}_0 + k\mathbf{a}_1\right),\end{aligned} \tag{3.25}$$

$$\begin{aligned}\hat{\mathbf{a}}_{\text{PCA},0} &= \frac{k^2}{1+k^2}\mathbf{a}_0 - \frac{k}{1+k^2}\mathbf{a}_1,\\ \hat{\mathbf{a}}_{\text{PCA},1} &= \frac{1}{1+k^2}\mathbf{a}_1 - \frac{k}{1+k^2}\mathbf{a}_0.\end{aligned} \tag{3.26}$$

Since there is no primary component in (3.26), (3.26) or (3.24) that comes from the basis vector with the smaller eigenvalue cannot be related with the extraction of the primary components. That is to say, the basis vector with larger eigenvalue always corresponds to the primary component regardless of the value of the primary power ratio *γ*. This observation reveals that the assumption $\gamma > 0.5$ in PCA is redundant. However, if this assumption is not satisfied in the stereo input signal, the extraction error of the extracted primary component becomes higher, as inferred from (3.25).

Furthermore, it is observed from (3.25) that the primary component is completely extracted by PCA, and no primary components are found in the extracted ambient components. On the other hand, the extracted primary components suffer from the ambient leakage, i.e., $\frac{1}{1+k^2}(\mathbf{a}_0 + k\mathbf{a}_1)$, and $\frac{k}{1+k^2}(\mathbf{a}_0 + k\mathbf{a}_1)$. The severity of ambient leakage increases as the ambient power increases. In other words, dominant primary components lead to better extraction performance using PCA. Some variants of PCA based PAE approaches that improve the PAE performance for stereo signal containing non-dominant primary component are discussed in [God08], [JHS10], [BJP12].

### 3.3.2 PAE using least-squares

Least-squares estimation is frequently used to approximate solutions for over-determined systems. According to the stereo signal model, Faller introduced LS to extract the primary and ambient components by minimizing the MSE of the extracted components [Fal06]. Considering the extraction of the primary component, the extraction error expressed in (3.15) can then be rewritten as

$$\mathbf{e}_{\mathrm{P}} = \hat{\mathbf{p}}_0 - \mathbf{p}_0 = \left(w_{\mathrm{P}0,0} + k w_{\mathrm{P}0,1} - 1\right)\mathbf{p}_0 + w_{\mathrm{P}0,0}\mathbf{a}_0 + w_{\mathrm{P}0,1}\mathbf{a}_1, \tag{3.27}$$

and the MSE is $J = E\left[\mathbf{e}_{\mathrm{P}}{}^{H}\mathbf{e}_{\mathrm{P}}\right]$. By substituting the assumptions and relationships of the signal model stated in (3.2)-(3.4) and (3.27), the MSE becomes

$$\begin{aligned} J = {} & P_{\mathbf{p}_0}\left[1+(k^2+1)\frac{1-\gamma}{2\gamma}\right]{w_{\mathrm{P}0,0}}^2 \\ & +P_{\mathbf{p}_0}\left\{\left[k^2+(k^2+1)\frac{1-\gamma}{2\gamma}\right]{w_{\mathrm{P}0,1}}^2-2w_{\mathrm{P}0,0}-2kw_{\mathrm{P}0,1}+2kw_{\mathrm{P}0,0}w_{\mathrm{P}0,1}+1\right\}. \end{aligned} \tag{3.28}$$

Hence, the weights can be easily obtained by taking the gradients of *J* with respect to $w_{\mathrm{P}0,0}, w_{\mathrm{P}0,1}$ and equating their results to zero. The weights of the primary component extracted by LS are found to be

$$w_{\mathrm{P}0,0}=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2},\ w_{\mathrm{P}0,1}=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}. \tag{3.29}$$

Similarly, the weights for the remaining components can also be derived. The extracted primary and ambient components using LS are thus expressed as

$$\begin{aligned} \hat{\mathbf{p}}_{\mathrm{LS},\,0} &= \frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right), \\ \hat{\mathbf{p}}_{\mathrm{LS},\,1} &= \frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right), \end{aligned} \tag{3.30}$$

$$\begin{aligned} \hat{\mathbf{a}}_{\mathrm{LS},0} &= \frac{1+k^2+\left(k^2-1\right)\gamma}{1+\gamma}\frac{1}{1+k^2}\mathbf{x}_0-\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\mathbf{x}_1, \\ \hat{\mathbf{a}}_{\mathrm{LS},1} &= -\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\mathbf{x}_0+\frac{1+k^2+\left(1-k^2\right)\gamma}{1+\gamma}\frac{1}{1+k^2}\mathbf{x}_1. \end{aligned} \tag{3.31}$$

From (3.30)-(3.31), we observe that the weights for the extracted primary and ambient components are not only dependent on *k*, but also related to *γ*. As compared with PCA, the panning relationship of *k* between the extracted primary components in the two channels still holds, but no explicit panning is found in the extracted ambient components using LS.

### 3.3.3 PAE using minimum leakage least-squares

As discussed in Section 3.2, three types of error may be found in the extracted components, namely, the distortion, interference, and leakage. The leakage is the most undesirable among the three, and priority should be given to the minimization of the leakage in the extraction process. We therefore propose MLLS, which minimizes the extraction error with the constraint that the leakage is minimum in the extracted components. The amount of leakage power in the extracted primary or ambient component can be quantified by the leakage-to-extracted-signal ratio (LeSR), which is given as

$$\mathrm{LeSR}_{\mathrm{P}} = P_{Leak_{\mathrm{P}}} \Big/ P_{\hat{\mathbf{p}}_0}, \;\; \mathrm{LeSR}_{\mathrm{A}} = P_{Leak_{\mathrm{A}}} \Big/ P_{\hat{\mathbf{a}}_0}. \tag{3.32}$$

Minimum leakage in the extracted components is achieved by minimizing LeSR. For the extracted primary component, the leakage comes from the ambient components. Using (3.15) and (3.26), the $\mathrm{LeSR}_{\mathrm{P}}$ is computed as:

$$\mathrm{LeSR}_{\mathrm{P}} = \frac{\left(w_{\mathrm{P}0,0}{}^2 + w_{\mathrm{P}0,1}{}^2\right) P_{\mathbf{a}_0}}{\left(w_{\mathrm{P}0,0} + k w_{\mathrm{P}0,1}\right)^2 P_{\mathbf{p}_0} + \left(w_{\mathrm{P}0,0}{}^2 + w_{\mathrm{P}0,1}{}^2\right) P_{\mathbf{a}_0}}. \tag{3.33}$$

Minimizing $\mathrm{LeSR}_{\mathrm{P}}$ with respect to $w_{\mathrm{P}0,0}, w_{\mathrm{P}0,1}$, we have

$$w_{\mathrm{P}0,1} = k w_{\mathrm{P}0,0}. \tag{3.34}$$

Next, we substitute (3.34) into the extraction error given by (3.15), and the extraction error becomes

$$\mathbf{e}_{\mathrm{P}} = \left[\left(1+k^2\right) w_{\mathrm{P}0,0} - 1\right] \mathbf{p}_0 + w_{\mathrm{P}0,0} \mathbf{a}_0 + k w_{\mathrm{P}0,0} \mathbf{a}_1. \tag{3.35}$$

Based on (3.12) and (3.35), the $\mathrm{ESR}_{\mathrm{P}}$ is expressed as

$$\mathrm{ESR}_{\mathrm{P}} = \frac{\left[\left(1+k^2\right) w_{\mathrm{P}0,0} - 1\right]^2 P_{\mathbf{p}_0} + \left(w_{\mathrm{P}0,0}{}^2 + k^2 w_{\mathrm{P}0,0}{}^2\right) P_{\mathbf{a}_0}}{P_{\mathbf{p}_0}}. \tag{3.36}$$

By minimizing ESR$_\mathrm{P}$, we arrive at $w_{\mathrm{P}0,0}=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}$, and $w_{\mathrm{P}0,1}=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}$.

Finally, we can express the primary component in channel 0 extracted by MLLS as

$$\hat{\mathbf{p}}_{\mathrm{MLLS},0}=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right), \tag{3.37}$$

The remaining components extracted by MLLS can be obtained similarly, and are found to be

$$\hat{\mathbf{p}}_{\mathrm{MLLS},1}=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right), \tag{3.38}$$

$$\begin{aligned}\hat{\mathbf{a}}_{\mathrm{MLLS},0}&=\frac{k}{1+k^2}\left(k\mathbf{x}_0-\mathbf{x}_1\right),\\ \hat{\mathbf{a}}_{\mathrm{MLLS},1}&=-\frac{1}{1+k^2}\left(k\mathbf{x}_0-\mathbf{x}_1\right).\end{aligned} \tag{3.39}$$

### 3.3.4 PAE using minimum distortion least-squares

Inspired by the popular minimum variance distortionless response (MVDR) filter [Cap69], we propose the minimum distortion least-squares in PAE by minimizing the extraction error ESR, with the constraint that the extracted component is distortionless. Mathematically, we can express the objective function of MDLS as $\min_{\mathbf{w}} \mathrm{ESR}\ \text{s.t. DSR}=0.$ Similar to the steps in MLLS, the solution for each extracted component can be derived as:

$$\begin{aligned}\hat{\mathbf{p}}_{\mathrm{MDLS},0}&=\frac{1}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right),\\ \hat{\mathbf{p}}_{\mathrm{MDLS},1}&=\frac{k}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right).\end{aligned} \tag{3.40}$$

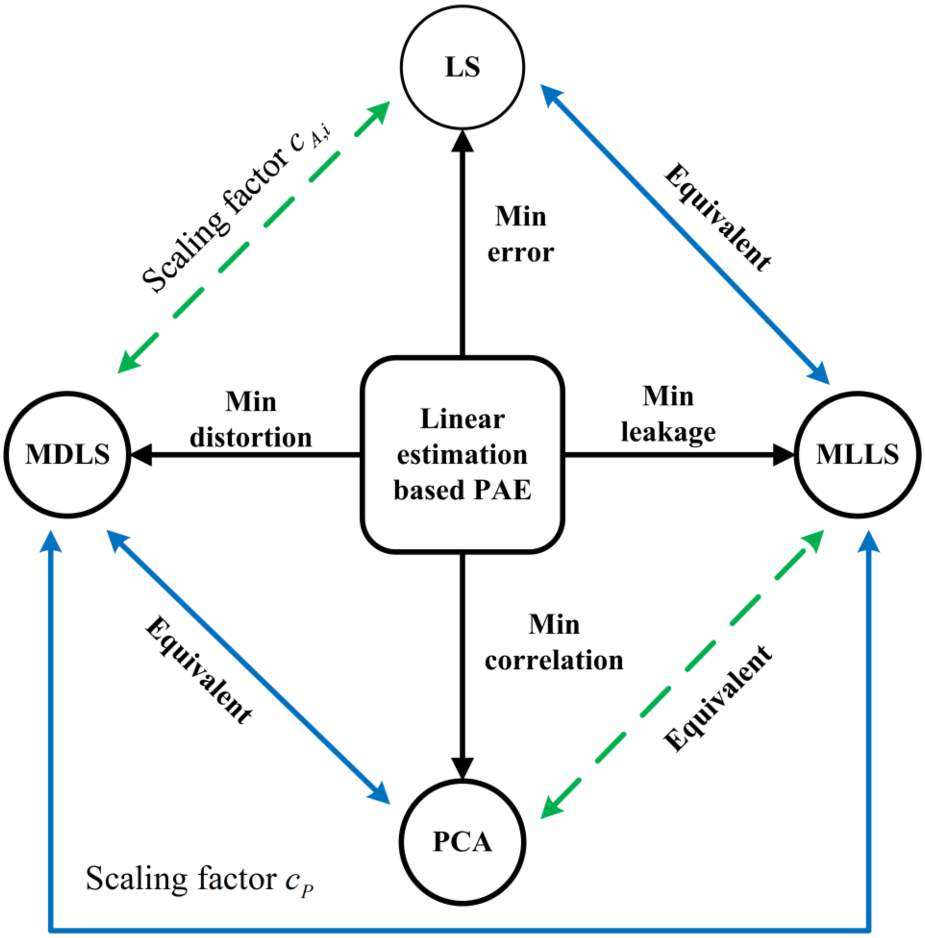


Figure 3.3 Objectives and relationships of four linear estimation based PAE approaches. Blue solid lines represent the relationships in the primary component, and green dotted lines represent the relationships in the ambient component.

$$
\begin{aligned}
\hat{\mathbf{a}}_{\mathrm{MDLS},0} &= \mathbf{x}_0 - \frac{2k\gamma}{\left(k^2-1\right)\gamma + k^2 + 1}\mathbf{x}_1, \\
\hat{\mathbf{a}}_{\mathrm{MDLS},1} &= -\frac{2k\gamma}{\left(1-k^2\right)\gamma + k^2 + 1}\mathbf{x}_0 + \mathbf{x}_1.
\end{aligned}
\qquad (3.41)
$$

### 3.3.5 Comparison among PCA, LS, MLLS, and MDLS in PAE

In this subsection, we compare the relationships and differences, as well as the performance among the four linear estimation based PAE approaches. The key minimization criteria and relationships of these approaches are illustrated in Fig. 3.3. Based on the linear estimation framework, PCA minimizes the correlation between the primary and ambient components, whereas LS, MLLS, and MDLS aim to minimize the extraction error, leakage, and distortion, respectively, for both the primary and ambient components. Some interesting relationships can be found for the primary components extracted using these approaches. From (3.23) and (3.40), we find that $\hat{\mathbf{p}}_{\text{PCA},i} = \hat{\mathbf{p}}_{\text{MDLS},i}, \forall i \in \{0,1\}$. This equivalence implies that PCA extracts the primary component with minimum distortion, even though PCA does not explicitly specify this constraint as found in MDLS. From (3.30) and (3.37)-(3.38), we observe that $\hat{\mathbf{p}}_{\text{LS},i} = \hat{\mathbf{p}}_{\text{MLLS},i}$. This equivalence implies that LS extracts the primary component with minimum leakage, even though LS does not explicitly specify this constraint as found in MLLS. There is an amplitude difference between the primary components extracted by MLLS and by MDLS, i.e.,

$$\hat{\mathbf{p}}_{\text{MLLS},i} = c_{\text{P}} \hat{\mathbf{p}}_{\text{MDLS},i}, \tag{3.42}$$

where the scaling factor $c_{\text{P}} = 2\gamma / (1+\gamma)$. Since $\gamma \in [0,1]$, $c_{\text{P}} \le 1$, it is clear that the primary component extracted by MLLS has lower power than the primary component extracted by MDLS for all $\gamma \ne 1$.

Similarly, we noted a few interesting relationships for the extracted ambient component. Based on (3.24) and (3.39), it is interesting to find that $\hat{\mathbf{a}}_{\text{PCA},i} = \hat{\mathbf{a}}_{\text{MLLS},i}$. This equivalence implies that PCA extracts the ambient component with minimum leakage, even though PCA does not explicitly specify this constraint as found in MLLS. From (3.31) and (3.41), there is also

Table 3.1 Results of performance measures for PCA, LS, minimum leakage LS, and minimum distortion LS in PAE.

| **Measures** | | **Primary component** | | | **Ambient component** | | |
|---|---|---|---|---|---|---|---|
| | | **MDLS / PCA** | **MLLS/LS** | | **MLLS /PCA** | **LS** | **MDLS** |
| Group 1: Extraction Accuracy | ESR | $\frac{1-\gamma}{2\gamma}$ | $\frac{1-\gamma}{1+\gamma}$ | | $\frac{1}{1+k^2}$ | $\frac{1}{1+k^2}\frac{2\gamma}{1+\gamma}$ | $\frac{2\gamma}{\left(k^2-1\right)\gamma+k^2+1}$ |
| | LSR | $\frac{1-\gamma}{2\gamma}$ | $\frac{1-\gamma}{2\gamma}\left(\frac{2\gamma}{1+\gamma}\right)^2$ | | 0 | $\frac{1}{1+k^2}\frac{2\gamma(1-\gamma)}{(1+\gamma)^2}$ | $\frac{\left(1+k^2\right)(1-\gamma)2\gamma}{\left[\left(1+k^2\right)(1+\gamma)-2\gamma\right]^2}$ |
| | DSR | 0 | $\left(\frac{1-\gamma}{1+\gamma}\right)^2$ | | $\left(\frac{1}{1+k^2}\right)^2$ | $\left(\frac{1}{1+k^2}\frac{2\gamma}{1+\gamma}\right)^2$ | 0 |
| | ISR | 0 | | | $\left(\frac{k}{1+k^2}\right)^2$ | $\left(\frac{k}{1+k^2}\frac{2\gamma}{1+\gamma}\right)^2$ | $\left[\frac{2k\gamma}{\left(1+k^2\right)(1+\gamma)-2\gamma}\right]^2$ |
| Group 2: Spatial Accuracy | ICC (ICTD) | 1(0) | | | 1 | $\frac{2k\gamma}{\sqrt{\left(1+k^2\right)^2-\left(1-k^2\right)^2\gamma^2}}$ | |
| | ICLD | $k^2$ | | | $\frac{1}{k^2}$ | $\frac{1}{k^2}\frac{1+\gamma+k^2(1-\gamma)}{1+\gamma+\frac{1}{k^2}(1-\gamma)}$ | $\frac{1}{k^2}\frac{1-\gamma+k^2(1+\gamma)}{1-\gamma+\frac{1}{k^2}(1+\gamma)}$ |

an amplitude difference between the ambient components extracted by MDLS and LS, which is given by

$$\hat{\mathbf{a}}_{\mathrm{LS},i}=c_{\mathrm{A},i}\hat{\mathbf{a}}_{\mathrm{MDLS},i}, \tag{3.43}$$

where $c_{\mathrm{A},i}=\frac{1+k^2+(-1)^i\left(k^2-1\right)\gamma}{\left(1+k^2\right)(1+\gamma)}$. As compared to (3.42), the scaling factor in the extracted ambient components differs from channel 0 to channel 1.

Next, we present a comparative analysis on the performance of these four PAE approaches. Here, we summarize the results of the performance measures obtained with channel 0 in Table 3.1. Due to the symmetry in the stereo signal model, the measures for channel 1 can be obtained by replacing *k* in the results in Table 3.1 with its reciprocal. From Table 3.1, it is clear that the two groups of measures are highly dependent on *γ* and/or *k*.

For the primary extraction, we have the following observations of MDLS (or PCA) and MLLS (or LS) based on the measures in Table 3.1. In Group 1, lower ESR and LSR of the extracted primary component are observed in MLLS as compared to MDLS. The distortion measure DSR = 0 indicates that primary component extracted using MDLS (or PCA) is free of distortion, whereas the distortion in MLLS (or LS) increases as $\gamma$ decreases. Hence, MLLS (or LS) extracts primary component with minimum leakage and error at the expense of introducing some distortion in the extracted primary component. All four approaches extract primary component without interference. According to the spatial cues (ICC, ICTD, and ICLD) of the primary component in Group 2, all four approaches are capable of preserving the correct spatial information in the extracted primary component.

For the ambient extraction, we have the following observations of MLLS (or PCA), LS, and MDLS based on the measures in Table 3.1. In Group 1, we observe that LS has the lowest ESR. The measure LSR = 0 found in MLLS indicates that no primary components are leaked into the extracted ambient component. In contrast, a certain amount of primary leakage is found in ambient component extracted using LS or MDLS. As for DSR, only MDLS extracts the ambient component without distortion. The overall best performance on the ambient extraction is achieved using LS based on the measures of diffuseness in Group 2, but none of the approaches is able to extract an uncorrelated and balanced ambient component. Therefore, some post-processing techniques such as decorrelation [Fal06b] and post-scaling [Fal06] should be used to enhance the ambient extraction.

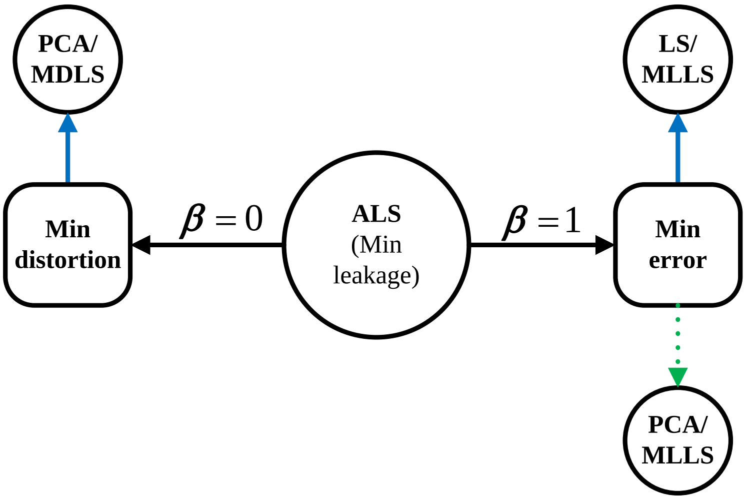


Figure 3.4 Characteristics and relationships of adjustable least-squares. Blue solid lines represent the relationships in the primary component, and green dotted lines represent the relationships in the ambient component.

### 3.3.6 PAE using adjustable least-squares

In this subsection, we propose the adjustable least-squares, which is designed to achieve an adjustable performance in terms of extraction error and distortion, as well as producing minimum leakage in the extracted primary and ambient components. Similar to (3.34), by minimizing the leakage LeSR in the extracted primary and ambient components, we have $\left[w_{\mathrm{P0,1}}, w_{\mathrm{P1,1}}\right] = k\left[w_{\mathrm{P0,0}}, w_{\mathrm{P1,0}}\right]$, and $\left[w_{\mathrm{A0,1}}, w_{\mathrm{A1,1}}\right] = -k^{-1}\left[w_{\mathrm{A0,0}}, w_{\mathrm{A1,0}}\right]$, respectively. To achieve the adjustable performance in terms of extraction error and distortion, we introduce the adjustable factor $\beta$ where $0 \le \beta \le 1$. By letting $\beta$ = 0, and $\beta$ = 1, we can achieve the minimum distortion and extraction error, respectively. Based on our analysis of the four PAE approaches, the weights in ALS are obtained as

$$\begin{bmatrix} w_{\mathrm{P}0,0} & w_{\mathrm{P}0,1} \\ w_{\mathrm{P}1,0} & w_{\mathrm{P}1,1} \end{bmatrix} = \frac{1}{1+k^2}\left(1-\beta\frac{1-\gamma}{1+\gamma}\right)\begin{bmatrix} 1 & k \\ k & k^2 \end{bmatrix}, \tag{3.44}$$

$$\begin{bmatrix} w_{\mathrm{A}0,0} & w_{\mathrm{A}0,1} \\ w_{\mathrm{A}1,0} & w_{\mathrm{A}1,1} \end{bmatrix} = \begin{bmatrix} 1-\beta\dfrac{1}{1+k^2} & -\dfrac{1}{k}\left(1-\beta\dfrac{1}{1+k^2}\right) \\ -k\left(1-\beta\dfrac{k^2}{1+k^2}\right) & 1-\beta\dfrac{k^2}{1+k^2} \end{bmatrix}. \tag{3.45}$$

Next, the three key performance measures for PAE using ALS are expressed as

$$\begin{aligned} &\mathrm{ESR}_{\mathrm{P}} = \frac{1-\gamma}{2\gamma} + \beta(\beta-2)\frac{(1-\gamma)^2}{2\gamma(1+\gamma)}, \\ &\mathrm{DSR}_{\mathrm{P}} = \beta^2\left(\frac{1-\gamma}{1+\gamma}\right)^2, \ \mathrm{LeSR}_{\mathrm{P}} = \frac{1-\gamma}{1+\gamma}, \\ &\mathrm{ESR}_{\mathrm{A}} = \frac{1}{k^2} + \beta(\beta-2)\frac{1}{k^2\left(k^2+1\right)}, \\ &\mathrm{DSR}_{\mathrm{A}} = \beta^2\left(\frac{1}{1+k^2}\right)^2, \ \mathrm{LeSR}_{\mathrm{A}} = 0. \end{aligned} \tag{3.46}$$

From the above measures, it can be inferred that the extraction error ESR decreases and the distortion DSR increases gradually as $\beta$ increases, whereas the measure for leakage LeSR remains constant and small. Since the adjustable factor $\beta$ = 0 and $\beta$ = 1 leads to minimum distortion and extraction error, respectively, other values of $\beta$ between 0 and 1 yield an adjustable performance in terms of extraction error and distortion. For example, ALS with $\beta$ = 0.5 produces 75% reduction of extraction error and distortion in PAE. The characteristics of ALS and its relationships with other PAE approaches are illustrated in Fig. 3.4. By adjusting the value of $\beta$, ALS can achieve the performance of the previously discussed PAE approaches. Specifically, in primary extraction, ALS with $\beta$ = 0 is equivalent to MDLS (or PCA), whereas ALS with $\beta$ = 1 is equivalent to MLLS (or LS). In ambient extraction, ALS can be linked with MLLS (or PCA) by letting $\beta$ = 1.

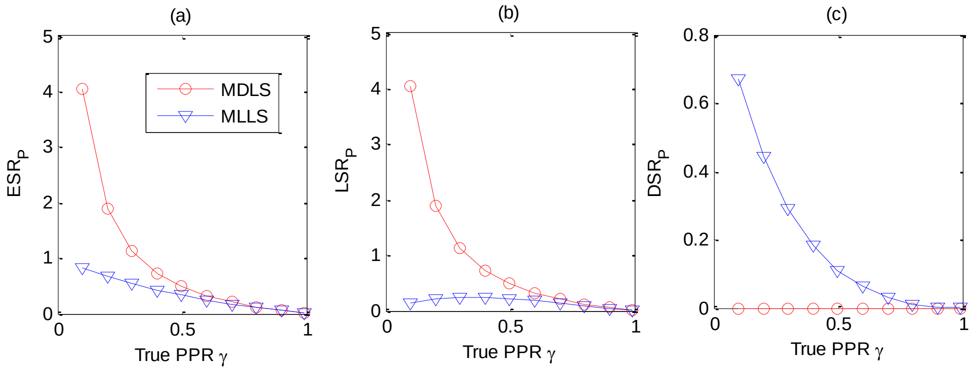


Figure 3.5 Comparison of MDLS (or PCA) and MLLS (or LS) in primary extraction, (a) error-to-signal ratio ESR; (b) leakage-to-signal ratio LSR, (c) distortion-to-signal ratio DSR. Legend in (a) applies to all plots.

## 3.4 Experiments and discussions

Since our focus in this chapter is to compare different linear estimation based PAE approaches, instead of the subband decomposition of the stereo signal, we shall consider only one primary component in the stereo signal in our simulations. A speech signal is selected as the primary component and uncorrelated white Gaussian noise with equal variance in two channels is synthesized as the ambient component in our simulations. To simulate the source panned to channel 1, the primary component is scaled by $k$ = 5. Subsequently, the stereo signals are synthesized by linearly mixing the primary and ambient components using different values of primary power ratio PPR, ranging from zero to one. The performance of these PAE approaches is then evaluated using the performance measures introduced in Section 3.2. Based on our simulations, we provide some recommendations for the applications using these PAE approaches.

### 3.4.1 Comparison of PAE Using PCA, LS, MLLS and MDLS

The simulation results of PAE using PCA, LS, MLLS, and MDLS are shown in Figs. 3.5-3.8. Recall that the extraction performance of the primary component is identical: (i) between PCA and MDLS, (ii) between LS and MLLS, we shall discuss the primary extraction for MLLS and MDLS only in this subsection. The extraction accuracy of the extracted primary components using MLLS and MDLS (same for the two channels) is shown in Fig. 3.5. Several observations from Fig. 3.5 are as follows. The extraction error given by $\mathrm{ESR_P}$ reduces gradually as $\gamma$ increases. The $\mathrm{ESR_P}$ and $\mathrm{LSR_P}$ for MLLS are relatively lower than those in MDLS, which indicates that MLLS is superior to MDLS in extracting the primary component in terms of the extraction error and leakage. However, the distortion of extracted primary component using MLLS increases as $\gamma$ decreases, whereas no distortion is found with MDLS. These observations can be directly related to the objectives of these approaches.

The difference in the performance for the extracted primary component between MLLS and MDLS is caused by the scaling difference, as expressed in (3.42). This scaling factor depends solely on PPR, which is determined by the power difference between true primary and ambient components in each frame. In the case of stationary primary and ambient components, the scaling factor is almost constant and leading to similar performance between MLLS and MDLS. However, there is a noticeable difference in the primary components extracted using MLLS and MDLS when the primary component is non-stationary. An

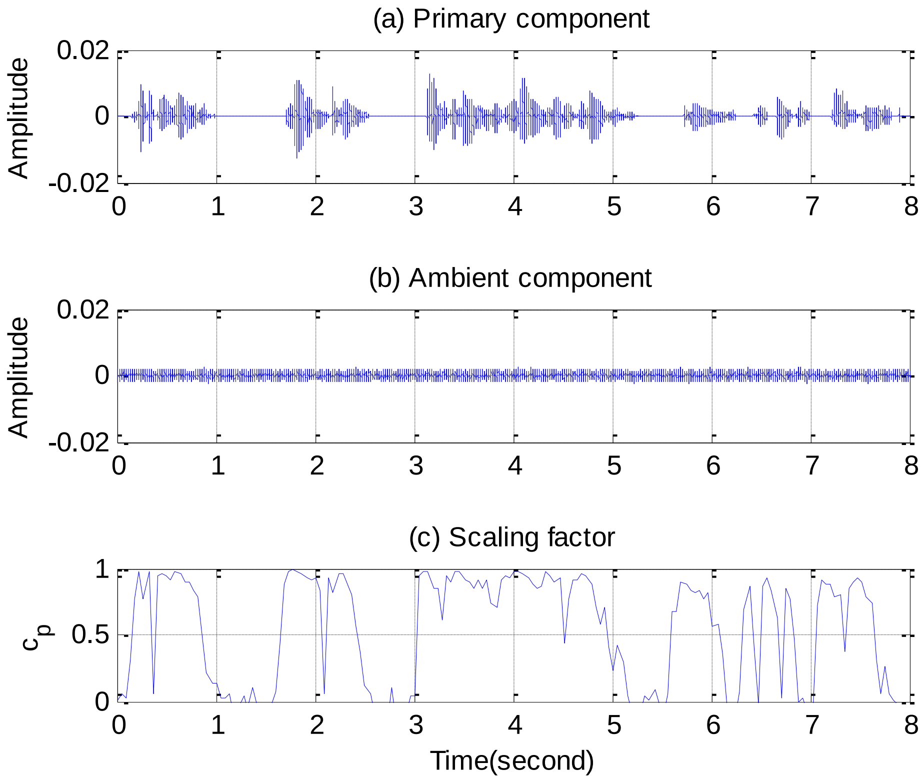


Figure 3.6 Scaling difference between the primary components extracted using MLLS and MDLS.

example to illustrate the variation of the scaling factor is shown in Fig. 3.6. It is observed that the scaling factor is fluctuating according to the power difference between primary and ambient components. The scaling factor rises closer to one when the primary component power is comparably stronger than the ambient component power, and the scaling factor drops to zero when the primary component becomes relatively weak compared to the ambient component. This example reveals that MLLS and MDLS behave similarly when primary component is dominant and only MLLS can extract weak primary component at the ambient-dominant periods of the signal. As a result, MLLS has lower $ESR_P$ but the extracted primary component may possess some discontinuity and more distortion, compared to MDLS.

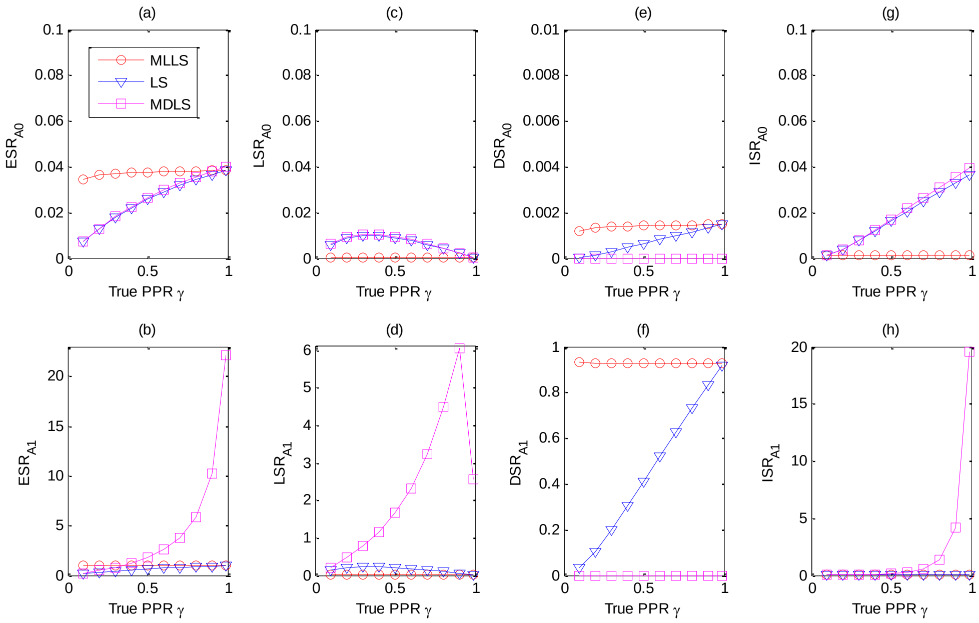


Figure 3.7 Comparison of ambient extraction ($k$ = 5) with MLLS (or PCA), LS and MDLS for channel 0 (top row) and channel 1 (bottom row). (a)-(b) error-to-signal ratio ESR; (c)-(d) leakage-to-signal ratio LSR; (e)-(f) distortion-to-signal ratio DSR; (g)-(h) interference-to-signal ratio ISR. Legend in (a) applies to all plots.

The performance of ambient extraction using PCA, LS, MLLS and MDLS is illustrated in Fig. 3.7. Unlike the primary extraction, the performance of ambient extraction has significant variation between the two channels. Due to the weaker primary component in channel 0, the performance of ambient extraction in channel 0 is better than that in channel 1 as shown in our simulations. Nevertheless, some common characteristics in the performance of ambient extraction in the two channels are observed. We found that LS has the lowest extraction error (Fig. 3.7(a)-(b)), whereas MLLS (or PCA), and MDLS can completely remove the leakage (Fig. 3.7(c)-(d)) and distortion (Fig. 3.7(e)-(f)), respectively. However, MDLS extracts the ambient component in

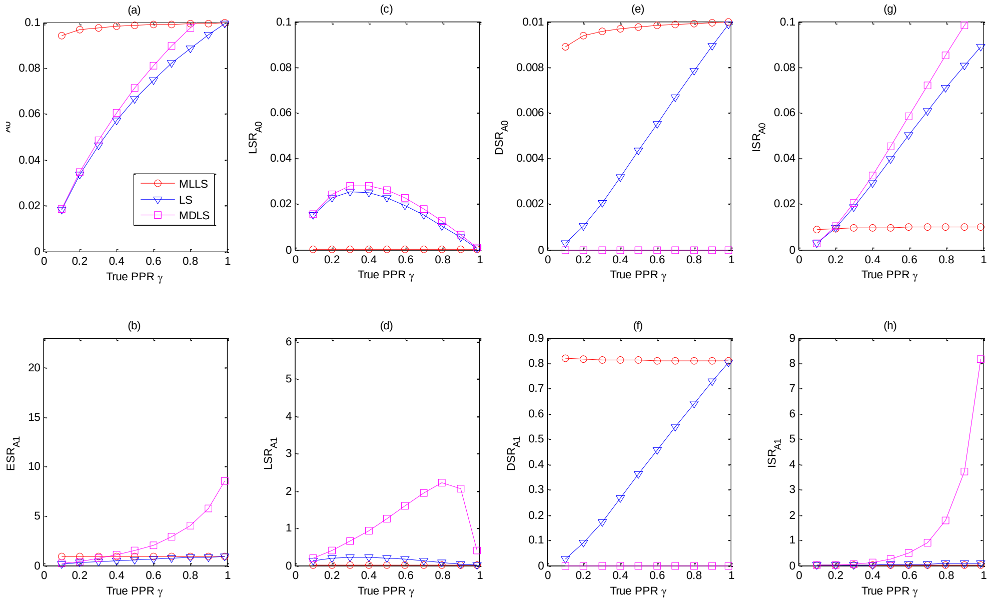


Figure 3.8 Comparison of ambient extraction ($k$ = 3) with MLLS (or PCA), LS and MDLS for channel 0 (top row) and channel 1 (bottom row). (a)-(b) ESR; (c)-(d) LSR; (e)-(f) DSR; (g)-(h) ISR. Legend in (a) applies to all plots.

channel 1 with much higher extraction error, leakage, and interference than the other PAE approaches.

In Figs. 3.8 and 3.9, we show the results of ambient extraction under different values of PPF, i.e., $k$ = 3 and $k$ = 1, respectively. With a smaller k, the extraction error performance between the two channels becomes closer. For $k = 3, 5$, we can observe a very similar relation in various error performance, with the difference on the scale. Whereas for $k$ = 1, the performance of MDLS becomes worse than MLLS or LS when PPR is high. Nevertheless, these methods still achieve the respective optimal performance in terms of different performance measures. That is, LS minimizes ESR, MLLS (or PCA) minimizes LSR, and MDLS minimizes DSR.

Finally, we examine the spatial accuracy of the extracted primary and ambient components ($k$ = 5), as shown in Fig. 3.10. Since the extracted primary

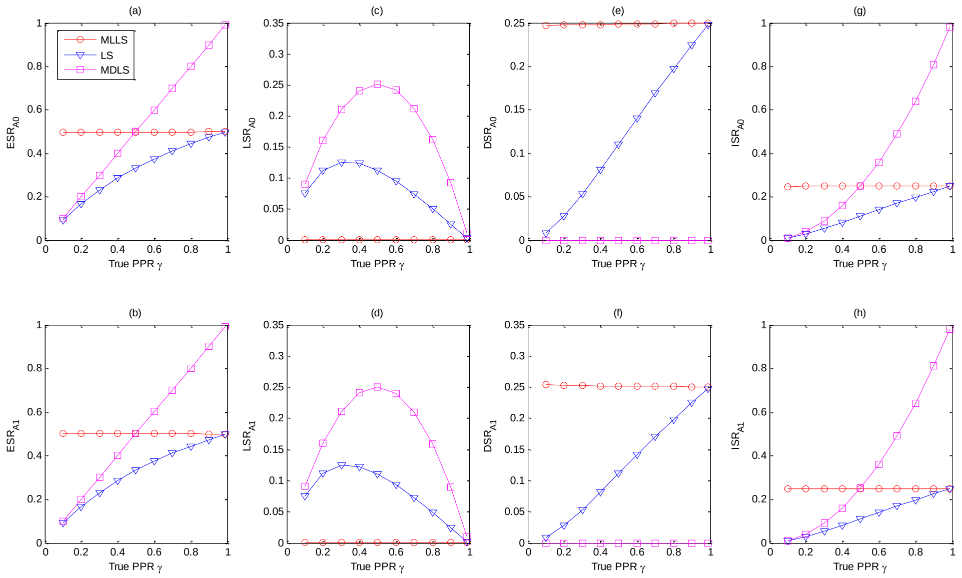


Figure 3.10 Comparison of ambient extraction (*k* = 1) with MLLS (or PCA), LS and MDLS for channel 0 (top row) and channel 1 (bottom row). (a)-(b) ESR; (c)-(d) LSR; (e)-(f) DSR; (g)-(h) ISR. Legend in (a) applies to all plots.

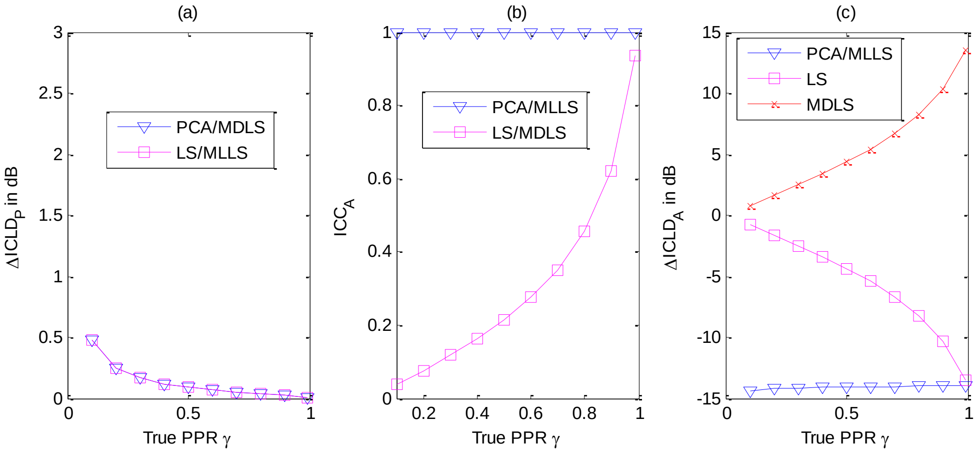


Figure 3.9 Comparison of spatial accuracy (*k* = 5) in PCA, LS, MLLS, and MDLS. (a) ICLD estimation error in the extracted primary component; (b) ICC of the extracted ambient component; and (c) ICLD estimation error in the extracted ambient component.

components are all scaled by *k* between the two channels, the ICC and ICTD of the primary components are the same as the true values, and the $ICLD_P$ is also very close to its true value, as shown in Fig. 3.10(a). However, from the results

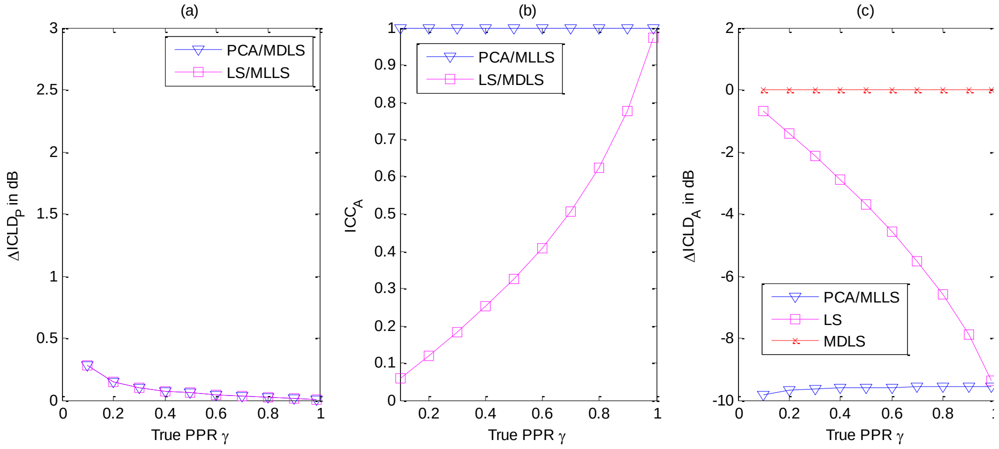


Figure 3.11 Comparison of spatial accuracy ($k$ = 3) in PCA, LS, MLLS, and MDLS. (a) ICLD estimation error in the extracted primary component; (b) ICC of the extracted ambient component; and (c) ICLD estimation error in the extracted ambient component.

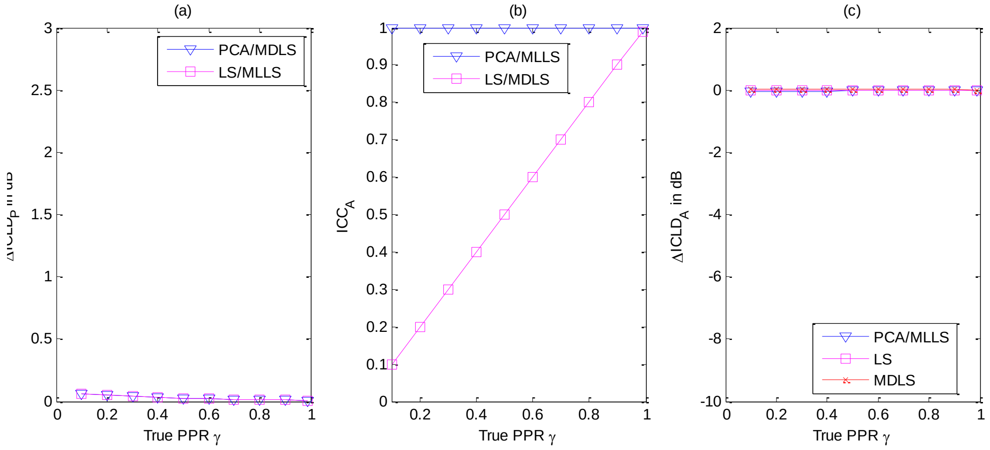


Figure 3.12 Comparison of spatial accuracy ($k$ = 1) in PCA, LS, MLLS, and MDLS. (a) ICLD estimation error in the extracted primary component; (b) ICC of the extracted ambient component; and (c) ICLD estimation error in the extracted ambient component.

of $\mathrm{ICC_A}$ and $\mathrm{ICLD_A}$ shown in Fig. 3.10(b) and 3.10(c), respectively, we found that none of these approaches is able to extract uncorrelated and balanced ambient components. In Figs. 3.11 and 3.12, we also show the spatial accuracy with other values of PPF, i.e., $k$ = 3 and $k$ = 1, respectively. Similar trends can

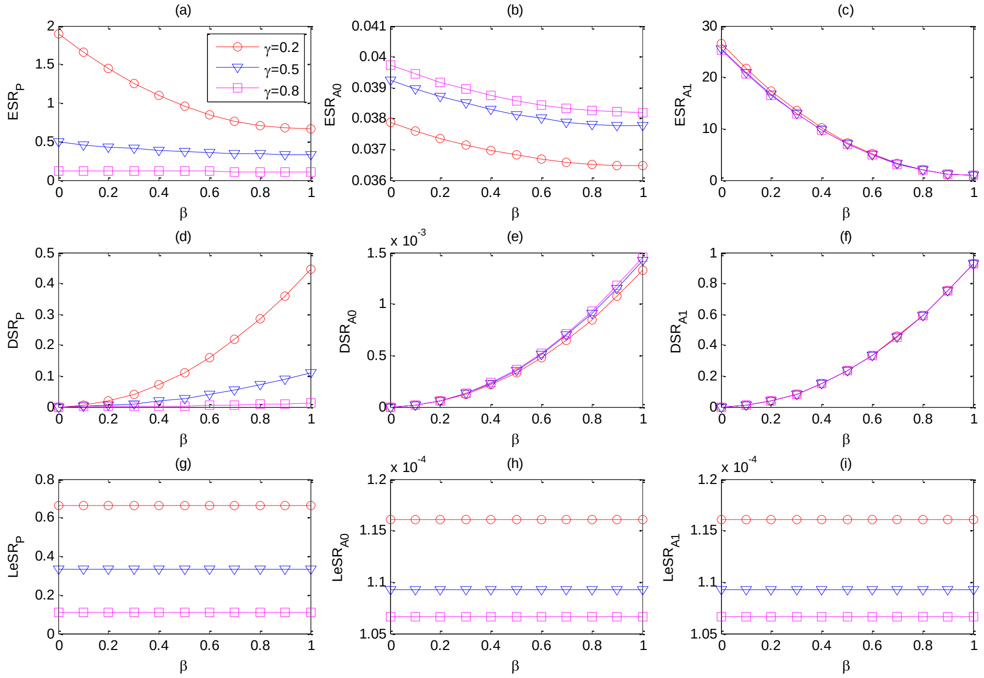


Figure 3.13 Measures for ALS with different values of adjustable factor *β*, error-to-signal ratio ESR (top row), distortion-to-signal ratio DSR (middle row), and leakage-to-extracted-signal ratio LeSR (bottom row), for the primary component (left column), the ambient component in channel 0 (middle column), and the ambient component in channel 1 (right column). Legend in (a) applies to all plots. Three lines in each plot represent different values of PPR *γ*.

be found with $\text{ICLD}_\text{P}$, $\text{ICC}_\text{A}$ and $\text{ICLD}_\text{A}$ for the three different values of PPF. One exceptional is that the $\text{ICLD}_\text{A}$ estimation difference is very small when PPF $k = 1$.

### 3.4.2 Performance of ALS in PAE

The performance of PAE using ALS is shown in Fig. 3.13. The measures for extraction error, distortion, and leakage are examined with respect to the adjustable factor *β*. These measures for the primary components for both channels are presented in the plots in the left column. The results of the measures for ambient extraction for the channels 0 and 1 are presented in the plots in the middle and right columns, respectively. From the plots of the top and middle rows, we observed that larger values of $\beta$ lead to lower extraction error (as shown by ESR) but higher distortion (as shown by DSR). Nevertheless, the leakage as quantified by LeSR remains at a very low level for all values of $\beta$, as shown in the plots in the bottom row. These observations verified that the adjustable performance in terms of extraction error and distortion using ALS is achieved by adjusting $\beta$.

### 3.4.3 General guidelines in selecting PAE approaches

Generally, the selection of the PAE approaches depends on the post-processing techniques and playback systems that are associated with the specific audio application, as well as the audio content and user preferences. Several guidelines on the applications of these PAE approaches can be drawn from our analysis and discussions. In Table 3.2, we summarize the strengths, weaknesses of different PAE approaches, and provide some recommendations on their applications. in Table 3.2. In applications like spatial audio coding and interactive audio in gaming, where the primary component is usually more important than the ambient component, PCA would be a better choice. In the case where both the primary and ambient components are extracted, processed, and finally mixed together, the extraction error becomes more critical and hence LS is recommended. In some spatial audio enhancement systems, where the extracted primary or ambient component is added back to the original signal to emphasize the extracted component, accurate extraction of the primary or ambient component becomes the key consideration. For such systems, MLLS is

Table 3.2 Strengths, weaknesses, and recommendations of different PAE approaches [HTG14]

| Approaches | Strengths | Weaknesses | Recommendations |
|---|---|---|---|
| PCA | • No distortion in the extracted primary component;<br>• No primary leakage in the extracted ambient component;<br>• Primary and ambient components are uncorrelated; | Ambient component severely panned; | Spatial audio coding and interactive audio in gaming, where the primary component is more important than the ambient component. |
| LS | Minimum MSE in the extracted primary and ambient components; | Severe primary leakage in the extracted ambient component; | Applications in which both the primary and ambient components are extracted, processed, and finally mixed together. |
| MLLS | • Minimum leakage in the extracted primary and ambient components;<br>• Primary and ambient components are uncorrelated; | Ambient component severely panned; | Spatial audio enhancement systems and applications in which different rendering or playback techniques are employed on the extracted primary and ambient components. |
| MDLS | No distortion in the extracted primary and ambient components; | Severe interference and primary leakage in the extracted ambient component; | High-fidelity applications in which timbre is of high importance. |
| ALS | Performance adjustable; | Need to adjust the value of the adjustable factor; | For applications without explicit requirements. |

preferred as the leakage becomes the most important consideration. MLLS is also recommended when different rendering and playback techniques are employed on the extracted primary and ambient components. MDLS is more suitable for high-fidelity applications, where timbre is of high importance, such as in musical application. When there is no explicit requirement, ALS can be employed by setting the proper adjustable factor.

## 3.5 Conclusions

In this chapter, we revisited the problem of primary ambient extraction (PAE) of stereo signals using linear estimation based approaches. Based on the stereo signal model, we formulated PAE as a problem to determine the weighting matrix under our linear estimation framework. Under this framework, we introduced two groups of performance measures and derived the solutions for two existing approaches, namely, principal component analysis (PCA), and least-squares (LS). Based on the objectives of minimum leakage, minimum distortion, and adjustable performance, we proposed three additional LS-based PAE approaches, namely, minimum leakage LS (MLLS), minimum distortion LS (MDLS), and adjustable LS (ALS). The relationships and differences of these PAE approaches are extensively studied. For primary extraction, PCA was found to be equivalent to MDLS in terms of minimum distortion; and LS is equivalent to MLLS in terms of minimum extraction error and leakage. The difference between extracted primary components using MDLS and MLLS is found to be a scaling factor, which is solely related to primary power ratio (PPR). All the discussed PAE approaches perform well for primary extraction but perform poorly in extracting ambient component when PPR is high. In ambient extraction, MLLS (or PCA), LS, and MDLS minimize the leakage, extraction error, and distortion, respectively. Adjustable LS offers an adjustable performance in terms of extraction error and distortion with the constraint of minimum leakage. Based on our discussions in this chapter, these PAE approaches are suggested in different spatial audio applications. In the following chapter, a different PAE framework will be discussed and compared with the linear estimation framework.

# Chapter 4

# Ambient Spectrum Estimation based Primary Ambient Extraction

Due to the nature of summing input signals directly [HTG14], the aforementioned PAE approaches, as studied in previous chapter, often have difficulty in removing uncorrelated ambient component in the extracted primary and ambient components. The extraction error in these PAE approaches is more severe when the ambient component is relatively strong compared to the primary component [HTG14], as often encountered in digital media content, including busy sound scenes with many discrete sound sources that contribute to the environment as well as strong reverberation indoor environment. According to [HGT15b], it is found that the percentage for the cases with over half of the time frames having relative strong ambient power is around 70% in these digital media content examples. Since high occurrence of strong ambient power case degrades the overall performance of PAE, a PAE approach that also performs well in the presence of strong ambient power is desired and investigated in this chapter[1].

In Section 4.1, we propose a new ambient spectrum estimation (ASE) framework to improve the performance of PAE. The ASE framework exploits

[1] The work reported in this chapter is an extension from the author's Journal papers [HGT15a] published in *IEEE Signal Processing Letters*, August 2015 issue, and [HGT15b] published in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, September 2015 issue.

the equal-magnitude characteristic of uncorrelated ambient components in the mixed signals of digital media content. These ASE problems are solved by pursuing the sparsity of the primary components [PBD10], as detailed in Section 4.2. To perform an in-depth evaluation of these PAE approaches, a novel technique to compute these measures for PAE approaches without analytic solutions, as is the case with the proposed ASE approaches, is proposed in Section 4.3. This is followed by the experiments to evaluate the PAE approaches in Section 4.4. Besides the comprehensive evaluation of these PAE approaches in ideal case, statistical variations are introduced to the ambient magnitudes to examine the robustness of the proposed ASE approaches. Furthermore, subjective listening tests are conducted to complement the objective evaluation. Finally, Section 4.5 concludes this chapter.

## 4.1 Ambient spectrum estimation framework

In this chapter, we denote the stereo signal in time-frequency domain at time frame index *m* and frequency bin index *l* as $X_c(m,l)$, where the channel index $c \in \{0,1\}$. Hence, the stereo signal at subband *b* that consists of bins from $l_{b-1}+1$ to $l_b$ (where $l_b$ is the upper boundary of bin index at subband *b*) is expressed as $\mathbf{X}_c[m,b] = \left[X_c(m,l_{b-1}+1), X_c(m,l_{b-1}+2), \ldots, X_c(m,l_b)\right]^T$ [GoJ06b]. The stereo signal model is expressed as:

$$\mathbf{X}_c[m,b] = \mathbf{P}_c[m,b] + \mathbf{A}_c[m,b] \quad \forall c \in \{0,1\}, \tag{4.1}$$

where $\mathbf{P}_c$ and $\mathbf{A}_c$ are the primary and ambient components in the *c*th channel of the stereo signal, respectively. Since the frequency band of the input signal is generally used in the analysis of PAE approaches, the indices $[m,b]$ are omitted for brevity.

The diffuseness of ambient components usually leads to low cross-correlation between the two channels of the ambient components in the stereo signal. During the mixing process, the sound engineers synthesize the ambient component using various decorrelation techniques, such as introducing delay [Rum99], all-pass filtering [Sch58], [PoB04], [Ken95b], artificial reverberation [Beg00], and binaural artificial reverberation [MeF09]. These decorrelation techniques often maintain the magnitude of ambient components in the two channels of the stereo signal. As such, we can express the spectrum of ambient components as

$$\mathbf{A}_c = |\mathbf{A}_c| \odot \mathbf{W}_c, \ \forall c \in \{0,1\}, \tag{4.2}$$

where $\odot$ denotes element-wise Hadamard product, $|\mathbf{A}_0| = |\mathbf{A}_1| = |\mathbf{A}|$ is the equal magnitude of the ambient components, and the element in the bin (*m, l*) of $\mathbf{W}_c$ is $W_c(m,l) = e^{j\theta_c(m,l)}$, where $\theta_c(m,l)$ is the bin (*m, l*) of $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_c = \angle\mathbf{A}_c$ is the vector of phase samples (in radians) of the ambient components. Following these discussions, we shall derive the ASE framework for PAE in two ways: ambient phase estimation (APE) and ambient magnitude estimation (AME).

### 4.1.1 Ambient phase estimation

Considering the panning of the primary component $\mathbf{P}_1 = k\mathbf{P}_0$, the primary component in (4.1) can be cancelled out and we arrive at

$$\mathbf{X}_1 - k\mathbf{X}_0 = \mathbf{A}_1 - k\mathbf{A}_0. \tag{4.3}$$

By substituting (4.2) into (4.3), we have

$$|\mathbf{A}| = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0), \tag{4.4}$$

where $./$ represents the element-wise division. Because ambient magnitude $|\mathbf{A}|$ is real and non-negative, we derive the relation between the phases of the two ambient components. First, we rewrite $\mathbf{W}_1 - k\mathbf{W}_0 = (\cos\boldsymbol{\theta}_1 - k\cos\boldsymbol{\theta}_0) + j(\sin\boldsymbol{\theta}_1 - k\sin\boldsymbol{\theta}_0)$. Since $|\mathbf{A}|$ is real, we have the following relation: $\sin\boldsymbol{\theta}./\cos\boldsymbol{\theta} = (\sin\boldsymbol{\theta}_1 - k\sin\boldsymbol{\theta}_0)./(\cos\boldsymbol{\theta}_1 - k\cos\boldsymbol{\theta}_0)$, which can be further rewritten as

$$\sin(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = k^{-1}\sin(\boldsymbol{\theta} - \boldsymbol{\theta}_1). \tag{4.5}$$

Two solutions arise when solving for $\boldsymbol{\theta}_0$:

$$\boldsymbol{\theta}_0^{(1)} = \boldsymbol{\theta} - \boldsymbol{\alpha}\ ,\ \boldsymbol{\theta}_0^{(2)} = \boldsymbol{\theta} + \boldsymbol{\alpha} + \boldsymbol{\pi}, \tag{4.6}$$

where $\boldsymbol{\alpha} = \arcsin\left[k^{-1}\sin(\boldsymbol{\theta} - \boldsymbol{\theta}_1)\right]$ and $\boldsymbol{\alpha} \in [-0.5\boldsymbol{\pi}, 0.5\boldsymbol{\pi}]$. Then we have $\sin\boldsymbol{\alpha} = k^{-1}\sin(\boldsymbol{\theta} - \boldsymbol{\theta}_1)$ and $\cos\boldsymbol{\alpha} = \sqrt{1 - k^{-2}\sin^2(\boldsymbol{\theta} - \boldsymbol{\theta}_1)}$. Based on the other condition that ambient magnitude $|\mathbf{A}|$ is nonnegative, the imaginary (or real) part of $\mathbf{W}_1 - k\mathbf{W}_0$ must have the same sign as the imaginary (or real) part of $\mathbf{X}_1 - k\mathbf{X}_0$. Next, we examine the two solutions for this condition. We take the first solution $\boldsymbol{\theta}_0^{(1)}$ and rewrite the ratio of imaginary part of $\mathbf{W}_1 - k\mathbf{W}_0$ to the imaginary part of $\mathbf{X}_1 - k\mathbf{X}_0$ as

$$\begin{aligned}\left.\frac{\operatorname{Im}\{\mathbf{W}_1 - k\mathbf{W}_0\}}{\operatorname{Im}\{\mathbf{X}_1 - k\mathbf{X}_0\}}\right|_{\boldsymbol{\theta}_0^{(1)}=\boldsymbol{\theta}-\alpha} &= \left.\frac{\sin\boldsymbol{\theta}_1 - k\sin\boldsymbol{\theta}_0}{\sin\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0^{(1)}=\boldsymbol{\theta}-\alpha} \\ &= \frac{\sin\boldsymbol{\theta}_1 - k\sin(\boldsymbol{\theta}-\boldsymbol{\alpha})}{\sin\boldsymbol{\theta}} \\ &= -\left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1) + k\cos\boldsymbol{\alpha}\right] \\ &= -\left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1) + \sqrt{k^2 - 1 + \cos^2(\boldsymbol{\theta}-\boldsymbol{\theta}_1)}\right] \\ &\leq 0.\end{aligned} \tag{4.7}$$

Therefore, the sign of the imaginary part of $\mathbf{W}_1 - k\mathbf{W}_0$ is different from the sign of imaginary part of $\mathbf{X}_1 - k\mathbf{X}_0$, resulting in negative values for ambient magnitude $|\mathbf{A}|$. Therefore, the first solution in (4.6) is inadmissible. Similarly, we take the second solution $\boldsymbol{\theta}_0^{(2)}$ and derive the ratio of imaginary part of $\mathbf{W}_1 - k\mathbf{W}_0$ to the imaginary part of $\mathbf{X}_1 - k\mathbf{X}_0$ as

$$\left.\frac{\operatorname{Im}\{\mathbf{W}_1 - k\mathbf{W}_0\}}{\operatorname{Im}\{\mathbf{X}_1 - k\mathbf{X}_0\}}\right|_{\boldsymbol{\theta}_0^{(2)}=\boldsymbol{\theta}+\alpha+\pi} = \left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1) + \sqrt{k^2 - 1 + \cos^2(\boldsymbol{\theta}-\boldsymbol{\theta}_1)}\right] \geq 0. \tag{4.8}$$

Therefore, the sign of the imaginary part of $\mathbf{W}_1 - k\mathbf{W}_0$ is the same from the sign of imaginary part of $\mathbf{X}_1 - k\mathbf{X}_0$, ensuring nonnegative values in ambient magnitude $|\mathbf{A}|$. Hence, we can conclude that based on the second solution, the relation between the ambient phases in two channels is

$$\left.\frac{\operatorname{Im}\{\mathbf{W}_1 - k\mathbf{W}_0\}}{\operatorname{Im}\{\mathbf{X}_1 - k\mathbf{X}_0\}}\right|_{\boldsymbol{\theta}_0^{(2)}=\boldsymbol{\theta}+\alpha+\pi} = \left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1) + \sqrt{k^2 - 1 + \cos^2(\boldsymbol{\theta}-\boldsymbol{\theta}_1)}\right] \geq 0. \tag{4.9}$$

where $\boldsymbol{\theta} = \angle(\mathbf{X}_1 - k\mathbf{X}_0)$. Furthermore, by substituting (4.4) and (4.2) into (4.1), we have

$$\begin{aligned}\mathbf{A}_c &= (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c, \\ \mathbf{P}_c &= \mathbf{X}_c - (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c,\ c \in \{0,1\}.\end{aligned} \tag{4.10}$$

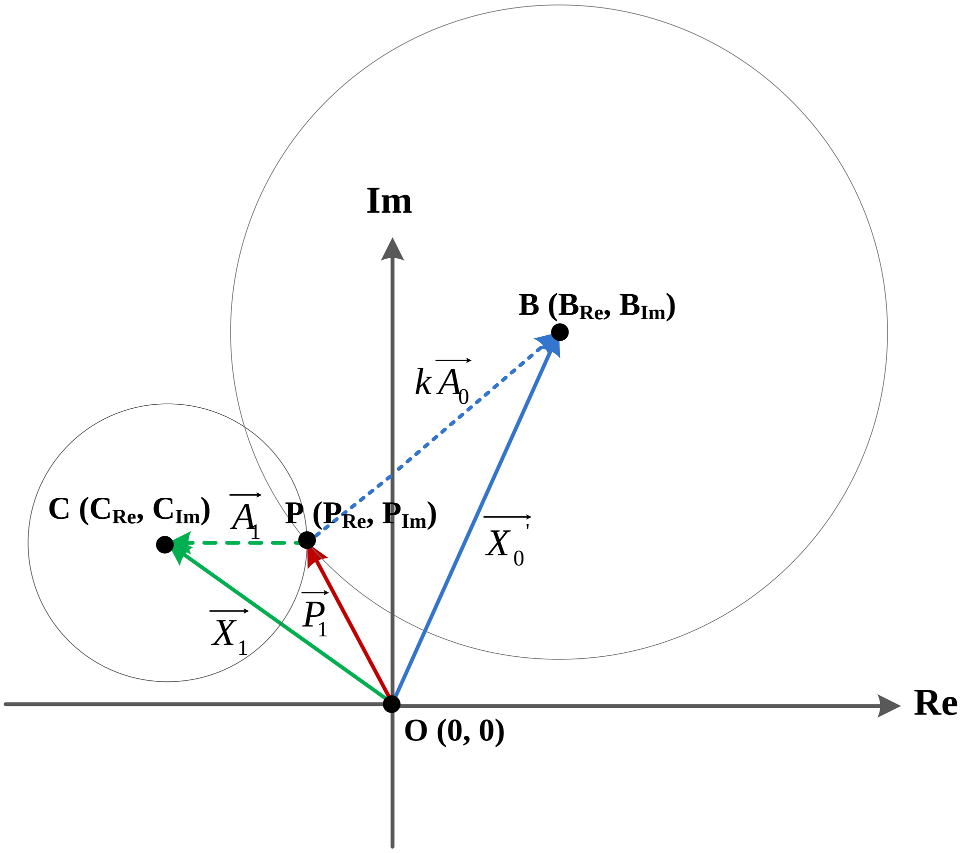

Im
Re
B (B$_{\mathrm{Re}}$, B$_{\mathrm{Im}}$)
$k\vec{A}_0$
C (C$_{\mathrm{Re}}$, C$_{\mathrm{Im}}$)
$\vec{A}_1$
P (P$_{\mathrm{Re}}$, P$_{\mathrm{Im}}$)
$\vec{X}_0$
$\vec{X}_1$
$\vec{P}_1$
O (0, 0)

Figure 4.1 Geometric representation of (4.11) in complex plane in AME

Since $\mathbf{X}_c$ and $k$ can be directly computed using the correlations of the input signals (refer to equation 3.8 in Chapter 3) [HTG14], $\mathbf{W}_0$, and $\mathbf{W}_1$ are the only unknown variables on the right hand side of the expressions in (4.10). In other words, the primary and ambient components are determined by $\mathbf{W}_0$, and $\mathbf{W}_1$, which are solely related to the phases of the ambient components. Therefore, we reformulate the PAE problem into an ambient phase estimation (APE) problem. Based on the relation between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ stated in (4.9), only one ambient phase $\boldsymbol{\theta}_1$ needs to be estimated.

### 4.1.2 Ambient magnitude estimation

To reformulate the PAE problem as an ambient magnitude estimation problem, we rewrite (4.1) for every time-frequency bin as:

$$\begin{aligned} X_0^{'} &= kX_0 = P_1 + kA_0, \\ X_1 &= P_1 + A_1. \end{aligned} \tag{4.11}$$

Consider these bin-wise spectra stated in (4.11) as vectors in complex plane (represented by an arrow on top), we can express their geometric relations in Fig. 4.1 as

$$\begin{aligned} &\overrightarrow{X_0^{'}} = \overrightarrow{\mathrm{OB}} = \left(\mathrm{B}_{\mathrm{Re}}, \mathrm{B}_{\mathrm{Im}}\right),\ \overrightarrow{X_1} = \overrightarrow{\mathrm{OC}} = \left(\mathrm{C}_{\mathrm{Re}}, \mathrm{C}_{\mathrm{Im}}\right), \\ &\overrightarrow{P_1} = \overrightarrow{\mathrm{OP}} = \left(\mathrm{P}_{\mathrm{Re}}, \mathrm{P}_{\mathrm{Im}}\right), \\ &k\overrightarrow{A_0} = \overrightarrow{\mathrm{PB}},\ \overrightarrow{A_1} = \overrightarrow{\mathrm{PC}}. \end{aligned} \tag{4.12}$$

Let $r$ denote the magnitude of the ambient component, i.e., $r = \left|\overrightarrow{A_0}\right| = \left|\overrightarrow{A_1}\right|$. Then we have $\left|\overrightarrow{\mathrm{PC}}\right| = r,\ \left|\overrightarrow{\mathrm{PB}}\right| = kr.$ Therefore, by drawing two circles with their origins at B and C, we can find their intersection point P (select one point when there are two intersection points), which corresponds to the spectrum of the primary component and leads to the solution for the extracted primary and ambient components. For any estimate of ambient magnitude $\hat{r}$, the coordinates of point P shall satisfy

$$\begin{aligned} \left(\mathrm{P}_{\mathrm{Re}} - \mathrm{B}_{\mathrm{Re}}\right)^2 + \left(\mathrm{P}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}}\right)^2 &= k^2\hat{r}^2, \\ \left(\mathrm{P}_{\mathrm{Re}} - \mathrm{C}_{\mathrm{Re}}\right)^2 + \left(\mathrm{P}_{\mathrm{Im}} - \mathrm{C}_{\mathrm{Im}}\right)^2 &= \hat{r}^2. \end{aligned} \tag{4.13}$$

The solution of $\left(\mathrm{P}_{\mathrm{Re}}, \mathrm{P}_{\mathrm{Im}}\right)$ for (4.13) is given by:

$$\begin{aligned}\hat{\mathrm{P}}_{\mathrm{Re}} &= \frac{\mathrm{B}_{\mathrm{Re}}+\mathrm{C}_{\mathrm{Re}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)\left(k^2-1\right)\hat{r}^2 \pm \left(\mathrm{B}_{\mathrm{Im}}-\mathrm{C}_{\mathrm{Im}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}, \\ \hat{\mathrm{P}}_{\mathrm{Im}} &= \frac{\mathrm{B}_{\mathrm{Im}}+\mathrm{C}_{\mathrm{Im}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)\left(k^2-1\right)\hat{r}^2 \mp \left(\mathrm{B}_{\mathrm{Re}}-\mathrm{C}_{\mathrm{Re}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},\end{aligned} \tag{4.14}$$

where the Euclidean distance between the points B and C, $\left|\overrightarrow{\mathrm{BC}}\right| = \sqrt{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)^2+\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)^2}$ and $\beta = \sqrt{\left[\left(k+1\right)^2\hat{r}^2-\left|\overrightarrow{\mathrm{BC}}\right|^2\right]\left[\left(k-1\right)^2\hat{r}^2-\left|\overrightarrow{\mathrm{BC}}\right|^2\right]}$. Based on (4.12), the spectra of the primary and ambient components can then be derived as:

$$\begin{aligned}&\hat{P}_1 = \hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}},\ \hat{P}_0 = k^{-1}\left(\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}\right), \\ &\hat{A}_1 = X_1 - \left(\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}\right),\ \hat{A}_0 = X_0 - k^{-1}\left(\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}\right).\end{aligned} \tag{4.15}$$

Therefore, the PAE problem becomes the problem of determining *r*, i.e., ambient magnitude estimation. The approach to determine *r* and select one of the two solutions in (4.14) will be discussed in Section 4.2. It can be inferred from Fig. 4.1 that determining the ambient magnitude is equivalent to determine the ambient phase as either of them will lead to the other. Therefore, we conclude that APE and AME are equivalent and they are collectively termed as ambient spectrum estimation. The block diagram of the ASE based PAE is illustrated in Fig. 4.2. The input signals are transformed into time-frequency domain using e.g., STFT, and followed by the proposed ambient spectrum estimation stage, where either APE or AME can be used. After estimating the ambient phase or magnitude, the extracted primary and ambient components can be derived in time-frequency domain, which are finally transformed into time domain. We argue that in theory, by accurately obtaining the spectra of ambient components, it is possible to achieve perfect extraction (i.e., error-free)

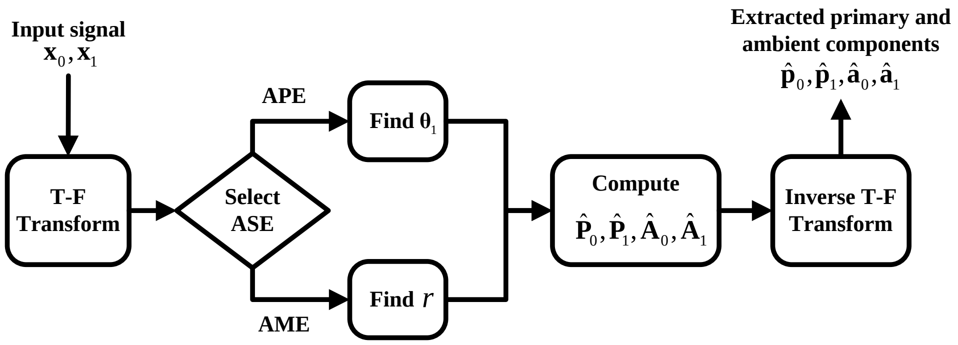

Input signal
$\mathbf{x}_0, \mathbf{x}_1$
T-F Transform
Select ASE
APE
AME
Find $\theta_1$
Find $r$
Compute $\hat{\mathbf{P}}_0, \hat{\mathbf{P}}_1, \hat{\mathbf{A}}_0, \hat{\mathbf{A}}_1$
Inverse T-F Transform
Extracted primary and ambient components
$\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \hat{\mathbf{a}}_0, \hat{\mathbf{a}}_1$

Figure 4.2 Block diagram of ASE based PAE

of the primary and ambient components using the formulation of ASE, which is not possible with existing PAE approaches as a consequence of residue error from the uncorrelated ambient component [HTG14].

## 4.2 Ambient spectrum estimation with a sparsity constraint

The proposed ambient spectrum estimation framework can greatly simplify the PAE problem into an estimation problem with only one unknown parameter per time-frequency bin. To estimate these parameters, we shall exploit other characteristics of the primary and ambient components that have not been used in previous derivations. One of the most important characteristics of sound source signals is sparsity, which has been widely used as a critical criterion in finding optimal solutions in many audio and music signal processing applications [PBD10]. In PAE, since the primary components are essentially directional sound sources, they can be considered to be sparse in the time-frequency domain [PBD10]. Therefore, we estimate the ambient phase or magnitude spectrum by restricting that the extracted primary component is sparse. We refer to these approaches as ambient spectrum estimation with a sparsity constraint (ASES). By applying the sparsity constraint in APE and AME, ASES can be divided into two approaches, namely, APES and AMES.

### 4.2.1 Ambient phase estimation with a sparsity constraint

With a sparsity constraint, the ambient phase estimation problem can be expressed as follows:

$$\hat{\boldsymbol{\theta}}_1^* = \arg\min_{\hat{\boldsymbol{\theta}}_1} \left\| \hat{\mathbf{P}}_1 \right\|_1 , \tag{4.16}$$

where $\left\| \hat{\mathbf{P}}_1 \right\|_1$ is the 1-norm of the primary component, which is equal to the sum of the magnitudes of the primary component over all the time-frequency bins. Since the objective function in (4.16) is not convex, convex optimization techniques are inapplicable. Heuristic methods, like simulated annealing [LaA87], require optimization to be performed for all the phase variables, and hence might be inefficient in solving APES. Therefore, a more efficient method referred to as discrete searching (DS) to estimate ambient phase is proposed. DS is proposed based on the following two observations. First, the magnitude of the primary component at one time-frequency bin is solely determined by the phase of the ambient component at the same time-frequency bin and hence, the estimation in (4.16) can be independently performed for each time-frequency bin. Second, the phase variable is bounded to $\left(-\pi, \pi\right]$ and high precision of the estimated phase may not be necessary. Thus, the optimal phase estimates can be selected from an array of discrete phase values $\hat{\theta}_1(d) = (2\pi d/D - \pi)$, where $d \in \{1, 2, \ldots, D\}$ with $D$ being the total number of phase values to be considered. In general, the value of $D$ affects the extraction and the computational performance of APES using DS. Following (4.9) and (4.10), a total number of $D$ estimates of the primary components can be computed. The estimated phase then corresponds to the minimum of magnitudes of the primary component, i.e., $\hat{\theta}_1^* = \hat{\theta}_1\left(d^*\right)$, where $d^* = \arg\min_{d\in\{1,2,\ldots,D\}}\left|\hat{P}_1\left(d\right)\right|$, Finally, the extracted primary and ambient components are computed using (4.10). It shall be noted that in DS, a sufficient condition of the sparsity constraint was employed in solving the APES problem in (4.16). The detailed steps of APES are listed in Table 4.1.

In addition to the proposed APES, we also consider a simple way to estimate the ambient phase based on the uniform distribution, i.e., $\hat{\boldsymbol{\theta}}_1^U \sim U\left(-\boldsymbol{\pi},\boldsymbol{\pi}\right]$. This approach is referred to as APEU, and is compared with the APES to examine the necessity of having a more accurate ambient phase estimation in the next section. Developing a complete probabilistic model to estimate the ambient phase, though desirable, is beyond the scope of the present study.

### 4.2.2 Ambient magnitude estimation with a sparsity constraint

Similarly to APES that is solved using the sparsity constraint, the ambient magnitude estimation problem can be expressed as:

$$\hat{\mathbf{r}}^* = \arg\min_{\hat{\mathbf{r}}}\left\|\hat{\mathbf{P}}_1\right\|_1, \tag{4.17}$$

where $\hat{\mathbf{r}}$ is the estimated ambient magnitude of all the time-frequency bins. As no constraints are placed on the ambient magnitude spectra among the time-frequency bins in one frame, the estimation of ambient magnitude can also be considered to be independent for every time-frequency bin. Therefore, the

Table 4.1 Steps in APES

```
1. Transform the input signal into time-frequency domain $X_0$, $X_1$, pre-compute $k$, choose $D$, **repeat** steps 2-7 for every time-frequency bin
2. Set $d = 1$, compute $\boldsymbol{\theta} = \angle\left(X_1 - kX_0\right)$, **repeat** steps 3-6
3. $\hat{\theta}_1\left(d\right) = 2\pi d/D - \pi$
4. Compute $\hat{\theta}_0\left(d\right)$ using eq. (4.9), and $\hat{W}_0\left(d\right), \hat{W}_1\left(d\right)$
5. Compute $\hat{P}_1\left(d\right)$ using eq. (4.10) and $\left|\hat{P}_1\left(d\right)\right|$
6. $d \leftarrow d + 1$, **Until** $d = D$
7. Find $d^* = \arg\min_{d\in\{1,2,\ldots,D\}} \left|\hat{P}_1\left(d\right)\right|$. **repeat** steps 3-5 with $d = d^*$ and compute the other components using eq. (4.10)
8. Finally, compute the time-domain primary and ambient components using inverse time-frequency transform.
```

estimation of ambient magnitude can be obtained individually for every time-frequency bin by minimizing the primary magnitude under the AMES framework.

To derive the solution for AMES, we follow the geometric relation illustrated in Fig. 4.1. To ensure the existence of intersection point P, the following constraint

$$\left|\overrightarrow{\mathrm{PC}}\right| - \left|\overrightarrow{\mathrm{PB}}\right| \le \left|\overrightarrow{\mathrm{BC}}\right| \le \left|\overrightarrow{\mathrm{PB}}\right| + \left|\overrightarrow{\mathrm{PC}}\right|, \tag{4.18}$$

has to be satisfied, which leads to:

$$r \in \left[r_{lb}, r_{ub}\right], \tag{4.19}$$

where $r_{lb} = \frac{\left|\overrightarrow{\mathrm{BC}}\right|}{k+1}, r_{ub} = \frac{\left|\overrightarrow{\mathrm{BC}}\right|}{k-1}, \forall k \neq 1.$ When $k = 1$, there is no physical upper bound from (4.18). Based on the objective of minimizing the magnitude of primary component, we can actually enforce an approximate upper bound for $k = 1$, for example, let $r_{ub} = \left|\overrightarrow{\mathrm{OB}}\right| + \left|\overrightarrow{\mathrm{OC}}\right|, \ \forall k = 1.$ Thus, the ambient magnitude is bounded, and the same numerical method DS (as used in APES) is employed to estimate $r$ in AMES. Consider an array of discrete ambient magnitude values

$\hat{r}(d)=\left(1-\frac{d-1}{D-1}\right)r_{lb}+\frac{d-1}{D-1}r_{ub},$ where $d\in\{1,2,\ldots,D\}$ with $D$ being the total number of ambient magnitude estimates considered. For each magnitude estimate $\hat{r}(d)$, we select the one $\left(\hat{\mathrm{P}}_{\mathrm{Re}},\hat{\mathrm{P}}_{\mathrm{Im}}\right)$ of two solutions from (4.14) which gives the smaller primary magnitude. First, we write the two solutions from (4.14) as:

$$\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{Re}}^{(1)}&=\frac{\mathrm{B}_{\mathrm{Re}}+\mathrm{C}_{\mathrm{Re}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)\left(k^2-1\right)\hat{r}^2+\left(\mathrm{B}_{\mathrm{Im}}-\mathrm{C}_{\mathrm{Im}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},\\
\hat{\mathrm{P}}_{\mathrm{Im}}^{(1)}&=\frac{\mathrm{B}_{\mathrm{Im}}+\mathrm{C}_{\mathrm{Im}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)\left(k^2-1\right)\hat{r}^2-\left(\mathrm{B}_{\mathrm{Re}}-\mathrm{C}_{\mathrm{Re}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},
\end{aligned}\tag{4.20}$$

$$\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{Re}}^{(2)}&=\frac{\mathrm{B}_{\mathrm{Re}}+\mathrm{C}_{\mathrm{Re}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)\left(k^2-1\right)\hat{r}^2-\left(\mathrm{B}_{\mathrm{Im}}-\mathrm{C}_{\mathrm{Im}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},\\
\hat{\mathrm{P}}_{\mathrm{Im}}^{(2)}&=\frac{\mathrm{B}_{\mathrm{Im}}+\mathrm{C}_{\mathrm{Im}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)\left(k^2-1\right)\hat{r}^2+\left(\mathrm{B}_{\mathrm{Re}}-\mathrm{C}_{\mathrm{Re}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},
\end{aligned}\tag{4.21}$$

We compare the power of the primary components given by these two solutions as:

$$\begin{aligned}
\left|\hat{\mathrm{P}}^{(1)}\right|^2\text{-}\left|\hat{\mathrm{P}}^{(2)}\right|^2&=\left[\hat{\mathrm{P}}_{\mathrm{Re}}^{(1)}\right]^2+\left[\hat{\mathrm{P}}_{\mathrm{Im}}^{(1)}\right]^2-\left[\hat{\mathrm{P}}_{\mathrm{Re}}^{(2)}\right]^2-\left[\hat{\mathrm{P}}_{\mathrm{Im}}^{(2)}\right]^2\\
&=4\left[\frac{\mathrm{B}_{\mathrm{Re}}+\mathrm{C}_{\mathrm{Re}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)\left(k^2-1\right)\hat{r}^2}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}\right]\frac{\left(\mathrm{B}_{\mathrm{Im}}-\mathrm{C}_{\mathrm{Im}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}\\
&\quad-4\left[\frac{\mathrm{B}_{\mathrm{Im}}+\mathrm{C}_{\mathrm{Im}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)\left(k^2-1\right)\hat{r}^2}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}\right]\frac{\left(\mathrm{B}_{\mathrm{Re}}-\mathrm{C}_{\mathrm{Re}}\right)\beta}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}\\
&=-\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}}\text{-}\mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)\frac{2\beta}{\left|\overrightarrow{\mathrm{BC}}\right|^2}.
\end{aligned}\tag{4.22}$$

By selecting the solution with smaller power, we arrive at

$$\hat{\mathrm{P}}_{\mathrm{Re}}, \hat{\mathrm{P}}_{\mathrm{Im}} = \begin{cases} \hat{\mathrm{P}}_{\mathrm{Re}}^{(1)}, \hat{\mathrm{P}}_{\mathrm{Im}}^{(1)}, & \left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}}\text{-}\mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right) \geq 0 \\ \hat{\mathrm{P}}_{\mathrm{Re}}^{(2)}, \hat{\mathrm{P}}_{\mathrm{Im}}^{(2)}, & \text{otherwise} \end{cases}. \tag{4.23}$$

Therefore, we can unify the solution for the selected $\left(\hat{\mathrm{P}}_{\mathrm{Re}}, \hat{\mathrm{P}}_{\mathrm{Im}}\right)$ from (4.14) based on the sign of $\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}}\text{-}\mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)$, that is

$$\begin{aligned} \hat{\mathrm{P}}_{\mathrm{Re}}(d) &= \frac{\mathrm{B}_{\mathrm{Re}} + \mathrm{C}_{\mathrm{Re}}}{2} + \frac{\left(\mathrm{C}_{\mathrm{Re}} - \mathrm{B}_{\mathrm{Re}}\right)\left(k^2 - 1\right)\hat{r}^2(d)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2} \\ &\quad + \frac{\left(\mathrm{B}_{\mathrm{Im}} - \mathrm{C}_{\mathrm{Im}}\right)\beta(d)\,\mathrm{sgn}\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}, \\ \hat{\mathrm{P}}_{\mathrm{Im}}(d) &= \frac{\mathrm{B}_{\mathrm{Im}} + \mathrm{C}_{\mathrm{Im}}}{2} + \frac{\left(\mathrm{C}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}}\right)\left(k^2 - 1\right)\hat{r}^2(d)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2} \\ &\quad + \frac{-\left(\mathrm{B}_{\mathrm{Re}} - \mathrm{C}_{\mathrm{Re}}\right)\beta(d)\,\mathrm{sgn}\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}, \end{aligned} \tag{4.24}$$

where sgn($x$) is the sign of $x$. The estimated magnitude of the primary component is obtained as

$$\left|\hat{P}_1(d)\right| = \sqrt{\hat{\mathrm{P}}_{\mathrm{Re}}^2(d) + \hat{\mathrm{P}}_{\mathrm{Im}}^2(d)}, \tag{4.25}$$

The estimated ambient magnitude then corresponds to the minimum of the primary component magnitude, i.e., $\hat{r}^* = \hat{r}\left(d^*\right)$, where $d^* = \arg\min_{d\in\{1,2,\ldots,D\}} \left|\hat{P}_1(d)\right|$. Finally, the extracted primary and ambient components are computed using (4.15).

### 4.2.3 Computational cost of APES and AMES

In this subsection, we compare the computational cost of APES and AMES, as shown in Table 4.2. In general, both AMES and APES are quite computational extensive. AMES requires more operations which include square root, addition,

Table 4.2 Computational cost of APES, AMES and APEX (for every time-frequency bin)

| Operation | Square root | Addition | Multiplication | Division | Comparison | **Trigonometric operation** |
|---|---|---|---|---|---|---|
| APES | *D* | 15*D*+18 | 15*D*+13 | 4*D*+6 | *D*-1 | **7*D*+6** |
| AMES | 2*D*+2 | 25*D*+35 | 24*D*+24 | 9*D*+13 | *D*-1 | **0** |
| APEX | 0 | 13 | 7 | 4 | 1 | **7** |

*D*: number of phase or magnitude estimates in discrete searching

multiplication, and division, but requires no trigonometric operations. By contrast, APES requires 7*D*+6 times of trigonometric operations for every time-frequency bin. The computational efficiency of these two approaches is affected by the implementation of these operations.

### 4.2.4 An approximate solution: APEX

To obtain a more efficient approach for ambient spectrum estimation, we derive an approximate solution in this subsection. For every time-frequency bin, we can rewrite (4.1) for the two channels as:

$$\begin{aligned}|X_0|^2 &= |P_0|^2 + |A_0|^2 + 2|P_0||A_0|\cos\theta_{PA0} = k^{-2}|P_1|^2 + |A|^2 + 2k^{-1}|P_1||A|\cos\theta_{PA0},\\ |X_1|^2 &= |P_1|^2 + |A_1|^2 + 2|P_1||A_1|\cos\theta_{PA1} = |P_1|^2 + |A|^2 + 2|P_1||A|\cos\theta_{PA1},\end{aligned} \tag{4.26}$$

where $\theta_{PA0}$, $\theta_{PA1}$ are the phase differences between the spectra of the primary and ambient components in channel 0 and 1, respectively. From (4.26), we can obtain that

$$\left(1-k^{-2}\right)|P_1|^2 + 2|A|\left(\cos\theta_{PA1} - k^{-1}\cos\theta_{PA0}\right)|P_1| - \left(|X_1|^2 - |X_0|^2\right) = 0. \tag{4.27}$$

Solving (4.27) for $|P_1|$, we arrive at

$$\begin{aligned}\forall k>1,\ |P_1| &= \frac{|A|\left(k^{-1}\cos\theta_{PA0}-\cos\theta_{PA1}\right)}{1-k^{-2}}\\ &\quad+\frac{\sqrt{|A|^2\left(k^{-1}\cos\theta_{PA0}-\cos\theta_{PA1}\right)^2+\left(1-k^{-2}\right)\left(|X_1|^2-|X_0|^2\right)}}{1-k^{-2}},\\ \forall k=1,\ |P_1| &= \frac{|X_1|^2-|X_0|^2}{2|A|\left(\cos\theta_{PA1}-\cos\theta_{PA0}\right)}\\ &= \frac{|X_1|^2-|X_0|^2}{-4|A|\left(\sin\frac{\theta_{PA0}+\theta_{PA1}}{2}\sin\frac{\theta_0-\theta_1}{2}\right)}.\end{aligned} \tag{4.28}$$

From (4.28), when $k > 1$, the minimization of $|P_1|$ can be approximately achieved by minimizing $k^{-1}\cos\theta_{PA0}-\cos\theta_{PA1}$ (considering that $|X_1|^2 \geq |X_0|^2$ in most cases since $k \geq 1$), which leads to $\theta_{PA0}=\pi,\ \theta_{PA1}=0.$ According to the relation between the two ambient phases in (4.9), we can infer that it is impossible to always achieve both $\theta_{PA0}=\pi$ and $\theta_{PA1}=0$ at the same time. Clearly, since $k>1,$ a better approximate solution would be taking $\theta_{PA1}=0.$ On the other hand, when $k$ = 1, one approximate solution to minimize $|P_1|$ would be letting $\theta_0-\theta_1=\pi.$ These constraints can be applied in either APE or AME framework. Here, applying the constraints in APE is more straightforward and we shall obtain the approximate phase estimation as:

$$\hat{\boldsymbol{\theta}}_1^* = \begin{cases}\angle\mathbf{X}_1, & \forall k>1\\ \angle\left(\mathbf{X}_1-\mathbf{X}_0\right), & \forall k=1\end{cases}. \tag{4.29}$$

As the phase (or the phase difference) of the input signals is employed in (4.29), we refer to this approximate solution as APEX. As shown in Table 4.2, APEX requires the lowest computational cost and is significantly more efficient than either APES or AMES. The performance of these approaches will be evaluated in the following sections.

## 4.3 Performance measures

An evaluation framework for PAE was initially proposed in [HTG14]. In general, we are concerned with the extraction accuracy and spatial accuracy in PAE. The overall extraction accuracy of PAE is quantified by error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components, where lower ESR indicates better extraction of these components. The ESR for the primary and ambient components are computed as

$$\begin{aligned}\mathrm{ESR}_{\mathrm{P}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{p}}_c-\mathbf{p}_c\right\|_2^2}{\left\|\mathbf{p}_c\right\|_2^2}\right\},\\ \mathrm{ESR}_{\mathrm{A}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{a}}_c-\mathbf{a}_c\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\},\end{aligned} \tag{4.30}$$

where $\mathbf{p}_c$, and $\mathbf{a}_c$ are the time-domain primary and ambient components of the whole signal, respectively. The extraction error can be further decomposed into three components, namely, the distortion, interference, and leakage (refer to Chapter 3 for the explanation of these three error components). Corresponding performance measures of these error components can be computed directly for PAE approaches with analytic solutions. As there is no analytic solution for these ASE approaches, we need to find alternative ways to compute these measures. In this section, we propose a novel optimization technique to estimate these performance measures.

We consider the extracted primary component in time domain $\hat{\mathbf{p}}_c$. Since the true primary components in two channels are completely correlated, no interference is incurred [HTG14]. Thus we can express $\hat{\mathbf{p}}_c$ as

$$\hat{\mathbf{p}}_c = \mathbf{p}_c + Leak_{\mathbf{p}_c} + Dist_{\mathbf{p}_c}, \tag{4.31}$$

where the leakage is $Leak_{\mathbf{p}_c} = \left( w_{Pc,0}\mathbf{a}_0 + w_{Pc,1}\mathbf{a}_1 \right)$, and the distortion is $Dist_{\mathbf{p}_c}$. To compute the measures, we need to estimate $w_{Pc,0}, w_{Pc,1}$ first. Considering that $\mathbf{p}_c$, $\mathbf{a}_0$, and $\mathbf{a}_1$ are inter-uncorrelated, we propose the following way to estimate $w_{Pc,0}, w_{Pc,1}$, with

$$\left( w^*_{Pc,0}, w^*_{Pc,1} \right) = \arg \min_{\left( w_{Pc,0}, w_{Pc,1} \right)} \left\| \hat{\mathbf{p}}_c - \mathbf{p}_c - \left( w_{Pc,0}\mathbf{a}_0 + w_{Pc,1}\mathbf{a}_1 \right) \right\|_2^2, \tag{4.32}$$

Thus, we can compute the measures, leakage-to-signal ratio (LSR) and distortion-to-signal ratio (DSR), for the primary components as

$$\begin{aligned} \mathrm{LSR_P} &= 10\log_{10}\left\{ \frac{1}{2}\sum_{c=0}^{1} \frac{\left\| w^*_{Pc,0}\mathbf{a}_0 + w^*_{Pc,1}\mathbf{a}_1 \right\|_2^2}{\left\| \mathbf{p}_c \right\|_2^2} \right\}, \\ \mathrm{DSR_P} &= 10\log_{10}\left\{ \frac{1}{2}\sum_{c=0}^{1} \frac{\left\| \hat{\mathbf{p}}_c - \mathbf{p}_c - \left( w^*_{Pc,0}\mathbf{a}_0 + w^*_{Pc,1}\mathbf{a}_1 \right) \right\|_2^2}{\left\| \mathbf{p}_c \right\|_2^2} \right\}. \end{aligned} \tag{4.33}$$

Next, we express $\hat{\mathbf{a}}_c$ in a similar way, as

$$\hat{\mathbf{a}}_c = \mathbf{a}_c + Leak_{\mathbf{a}_c} + Intf_{\mathbf{a}_c} + Dist_{\mathbf{a}_c}, \tag{4.34}$$

where the leakage is $Leak_{\mathbf{a}_c} = w_{Ac,c}\mathbf{p}_c$, and the interference $Intf_{\mathbf{a}_c} = w_{Ac,1-c}\mathbf{a}_{1-c}$ originates from the uncorrelated ambient component. The two weight parameters $w_{Ac,c}, w_{Ac,1-c}$ can be estimated as

$$\left( w^*_{Ac,c}, w^*_{Ac,1-c} \right) = \arg \min_{\left( w_{Ac,c}, w_{Ac,1-c} \right)} \left\| \hat{\mathbf{a}}_c - \mathbf{a}_c - \left( w_{Ac,c}\mathbf{p}_c + w_{Ac,1-c}\mathbf{a}_{1-c} \right) \right\|_2^2, \tag{4.35}$$

Thus, we compute the measures LSR, interference-to-signal ratio (ISR), and (DSR) for the ambient components as

$$
\begin{aligned}
\mathrm{LSR_A} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|w^*_{Ac,c}\mathbf{p}_c\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\},\\
\mathrm{ISR_A} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|w^*_{Ac,1-c}\mathbf{a}_{1-c}\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\},\\
\mathrm{DSR_A} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{a}}_c-\mathbf{a}_c-\left(w^*_{Ac,c}\mathbf{p}_c+w^*_{Ac,1-c}\mathbf{a}_{1-c}\right)\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}.
\end{aligned}
\tag{4.36}
$$

Previous experience on evaluating linear estimation based PAE approaches such as PCA and least-squares suggests that these parameters $w_{Pc,0}, w_{Pc,1}, w_{Ac,c}, w_{Ac,1-c}$ are bounded to [-1, 1], hence we can employ a simple numerical searching method similar to DS to determine the optimal estimates of these parameters using a certain precision [HTG14]. As audio signals from digital media are quite non-stationary, these measures shall be computed for every frame and can be averaged to obtain the overall performance for the whole track.

On the other hand, spatial accuracy is measured using the inter-channel cues. For primary components, the accuracy of the sound localization is mainly evaluated using inter-channel time and level differences (i.e., ICTD and ICLD). In this chapter, there is no ICTD involved in the basic mixing model for stereo input signals, and the ICLD is essentially determined by the estimation of *k*, which is common between the proposed approaches and the existing linear estimation based approaches such as PCA [HTG14]. For these two reasons, spatial accuracy is not evaluated for primary component extraction, but is focused on the extraction of ambient components. The spatial accuracy of the ambient component is evaluated in terms of its diffuseness, as quantified by inter-channel cross-correlation coefficient (ICC, from 0 to 1) and the ICLD (in dB). It is clear that a more diffuse ambient component requires both ICC and ICLD to be closer to 0.

## 4.4 Experiments and discussions

In this section, we present a comprehensive objective and subjective evaluation of the proposed ASE approaches and two existing PAE approaches, namely, PCA [GoJ07b], and time-frequency masking [MGJ07]. We present a preliminary experimental result on APES[1], followed by the detailed results on ASE approaches[2]. To examine the robustness of these PAE approaches, we evaluate the proposed approaches using synthesized mixed signal with unequal ambient magnitude in two channels. Lastly, subjective listening tests were conducted to examine the perceptual timbre and spatial quality of different PAE approaches.

### 4.4.1 Experimental results on APES

Experiments using synthesized mixed signals were carried out to evaluate the proposed approach. One frame (consists of 4096 samples) of speech signal is selected as the primary component, which is amplitude panned to channel 1 with a panning factor *k* = 4, 2, 1. A wave lapping sound recorded at the beach is selected as the ambient component, which is decorrelated using all-pass filters with random phase [Ken95b]. The stereo input signal is obtained by mixing the

[1] The source code and demo tracks are available: http://jhe007.wix.com/main#!ambient-phase-estimation/cied

[2] The source code and demo tracks are available: http://jhe007.wix.com/main#!ambient-spectrum-estimation/c6bk.

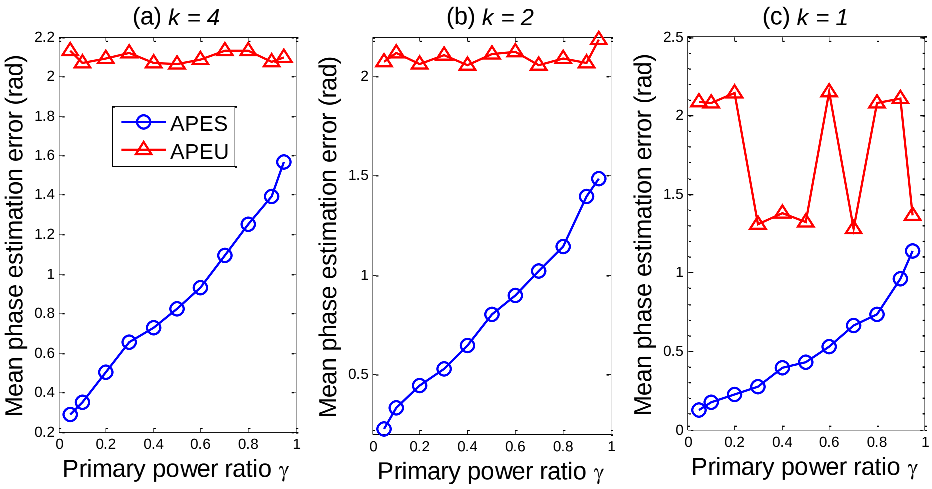


Figure 4.3 Comparison of ambient phase estimation error between APES and APEU with (a) *k* = 4; (b) *k* = 2; and (c) *k* = 1. Legend in (a) applies to all the plots.

primary and ambient components using different values of primary power ratio ranging from 0 to 1 with an interval of 0.1.

Our experiments compare the extraction performance of APES, APEU, PCA [GoJ07b], and two time-frequency masking approaches: Masking [MGJ07] and Masking_2 [AvJ04]. In the first three experiments, DS with *D* = 100 is used as the searching method of APES. Extraction performance is quantified by the error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components.

First, we examine the significance of ambient phase estimation by comparing the performance of APES with APEU. In Fig. 4.3, we show the mean phase estimation error and it is observed that compared to a random phase in APEU, the phase estimation error in APES is much lower. As a consequence, ESRs in APES are significantly lower than those in APEU, as shown in Fig. 4.4. This result indicates that obviously, close ambient phase estimation is

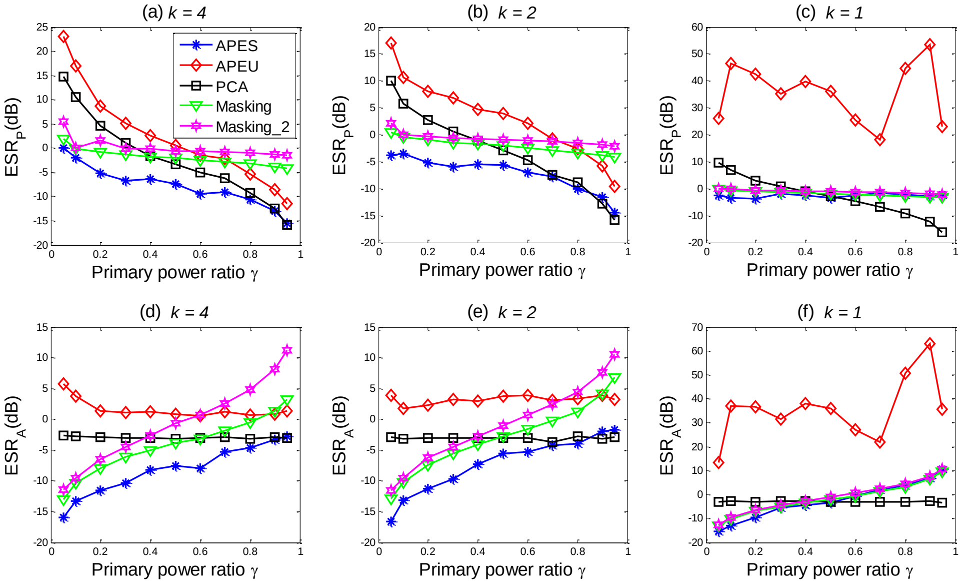


Figure 4.4 ESR of (a-c) extracted primary component and (d-f) extracted ambient component, with respect to 3 different values of primary panning factor (*k* = 4, 2, 1), using APES, APEU, PCA [GoJ07b], Masking [MGJ07], and Masking_2 [AvJ04]. Legend in (a) applies to all the plots.

necessary.

Second, we compare the APES with some other PAE approaches in the literature. From Fig. 4.4, it is clear that APES significantly outperforms other approaches in terms of ESR for $\gamma \leq 0.8$ and $k \neq 1$, suggesting that a better extraction of primary and ambient components is found with APES when primary components is panned and ambient power is strong. When *k* = 1, APES has comparable performance to the masking approaches, and performs slightly better than PCA for $\gamma \leq 0.5$. Referring to Fig. 4.3 that the ambient phase estimation error is similar for different *k* values, we can infer that the relatively poorer performance of APES for *k* = 1 is an inherent limitation of APES. Moreover, we compute the mean ESR across all tested $\gamma$ and *k* values and find

Table 4.3 Comparison of APES with different searching methods

| Method | Computation time (s) | $\mathrm{ESR_P}$ (dB) | $\mathrm{ESR_A}$ (dB) |
|---|---|---|---|
| DS (*D*=10) | 0.18 | -7.28 | -7.23 |
| DS (*D*=100) | 1.62 | -7.58 | -7.50 |
| SA | 426 | -7.59 | -7.51 |

that the average error reduction in APES over PCA and the two time-frequency masking approaches are 3.1, 3.5, and 5.2 dB, respectively. Clearly, the error reduction is even higher (up to 15 dB) for low $\gamma$ values.

Lastly, we compare the performance, as well as the computation time among different searching methods in APES: SA, DS with $D$ = 10 and 100. The results with $\gamma$ = 0.5 and $k$ = 4 are presented in Table 4.3. It is obvious that SA requires significantly longer computation time to achieve similar ESR when compared to DS. More interestingly, the performance of DS does not vary significantly as the precision of the search increases (i.e., $D$ is larger). However, the computation time of APES increases almost proportionally as $D$ increases. Hence, we infer that the proposed APES is not very sensitive to phase estimation errors and therefore the efficiency of APES can be improved by searching a limited number of phase values.

### 4.4.2 Experimental results on ASE approaches

In these experiments, the searching method of APES or AMES is DS with $D$ = 100. Based on the performance measures introduced in Section 4.3, we shall compare the overall extraction error performance, the specific error performance including leakage, distortion, and interference, as well as the spatial accuracy of the ambient components. Additionally, we will also compare the efficiency of these PAE approaches in terms of the computation time based

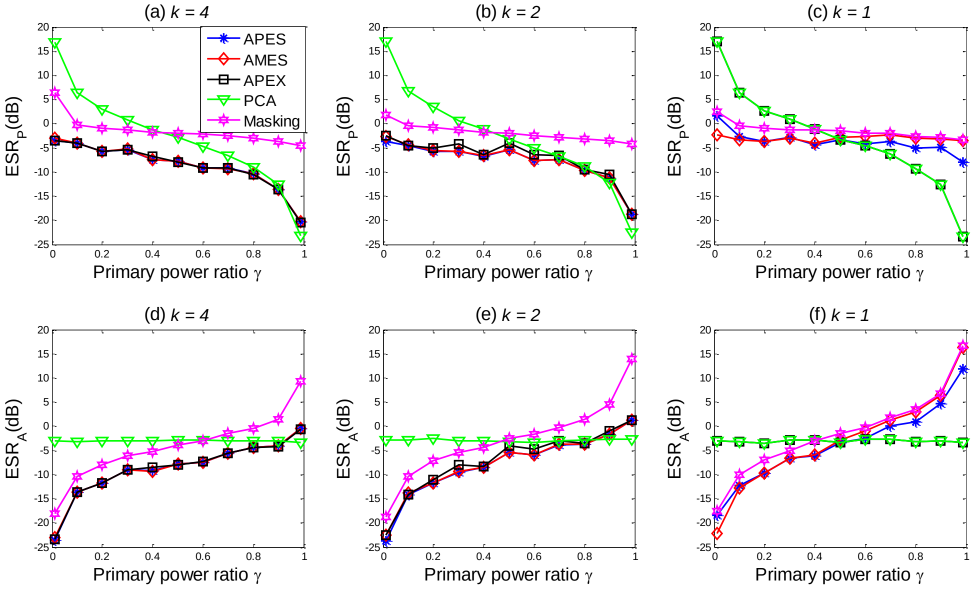


Figure 4.5 Comparison of the ESR of (a-c) extracted primary components and (d-f) extracted ambient components, with respect to different *k* values, using APES, AMES, APEX, PCA [GoJ07b], and Masking [MGJ07].

on our simulation. The stereo mixed signals employed in the experiments are synthesized in the following way. A frame (4096 samples, sampling rate: 44.1 kHz) of speech signal is selected as the primary component, which is amplitude panned to channel 1 with a panning factor $k \in \{1,\ 2,\ 4\}$. A wave lapping sound recorded at the beach is selected as the ambient component, which is decorrelated using all-pass filters with random phase [Ken95b]. The stereo signal is obtained by mixing the primary and ambient components based on different $\gamma$ values ranging from 0 to 1 with an interval of 0.1.

In the first experiment, we compare the overall performance of the three ASE approaches with two other PAE approaches in the literature, namely, PCA [GoJ07b] and Masking [MGJ07]. For the proposed ASE approaches, FFT size is set as 4096, whereas for Masking, the best setting for FFT size is found as 64.

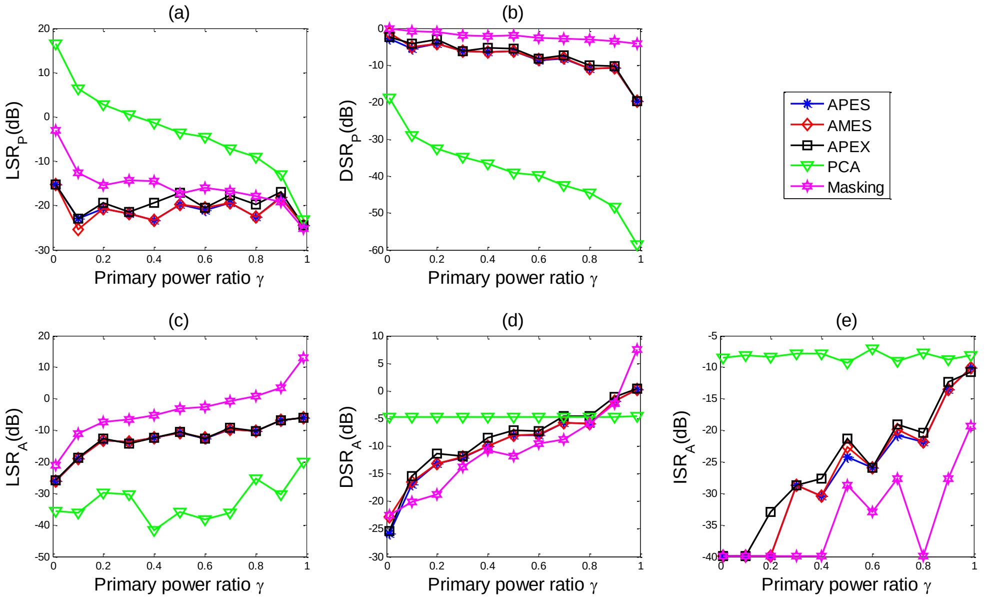


Figure 4.6 Comparison of the specific error performance of (a-b) LSR and DSR in the extracted primary components and (c-e) LSR, DSR, and ISR in the extracted ambient components using APES, AMES, APEX, PCA, and Masking.

The ESR of these approaches with respect to different values of $\gamma$ and *k* is illustrated in Fig. 4.5. Our observations of the ESR performance are as follows:

1) Generally, the performance of all these PAE approaches varies with $\gamma$. As $\gamma$ increases, ESR$_P$ decreases while ESR$_A$ increases (except ESR$_A$ of PCA). Considering primary components to be more important in most applications, it becomes apparent that the two representative existing approaches cannot perform well when $\gamma$ is low.

2) Primary panning factor *k* is the other factor that affects the ESR performance of these PAE approaches except PCA. For the Masking approach, the influence of *k* is insignificant for most cases except ESR$_P$ at very low $\gamma$ and ESR$_A$ at very high $\gamma$. By contrast, the ASE approaches are

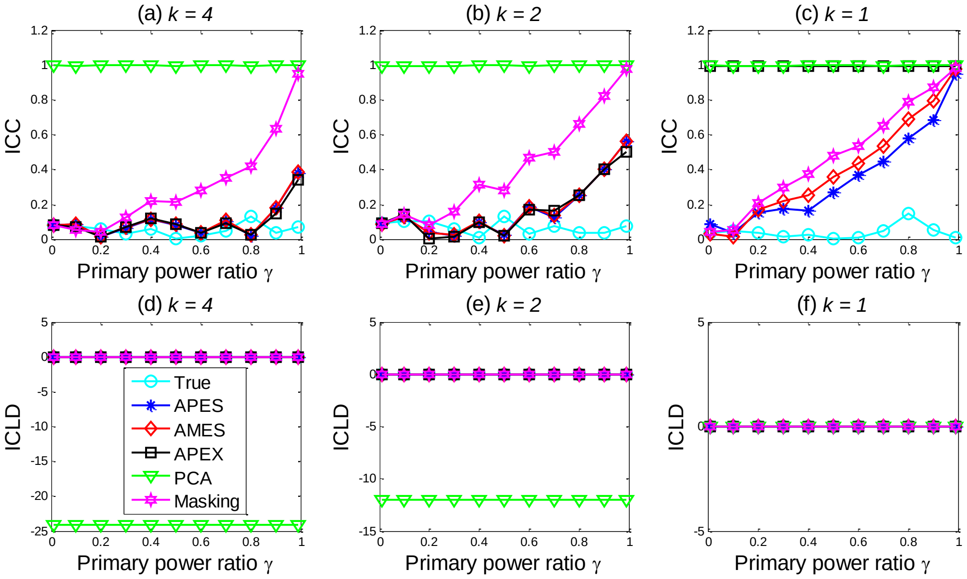


Figure 4.7 Comparison of the diffuseness of the extracted ambient components in terms of (a)-(c) ICC and (d)-(f) ICLD using APES, AMES, APEX, PCA, and Masking.

more sensitive to *k*. The ESR of APES and AMES are lower at higher *k*, especially when $\gamma$ is high. For APEX, the performance varies between *k* > 1, and *k* = 1, which was implied in (4.29).

3) Irrespective of $\gamma$ and *k*, APES and AMES perform quite similar. Both APES and AMES outperform existing approaches at lower $\gamma$, i.e., from $\gamma <$ 0.8 when *k* = {2, 4} to $\gamma <$ 0.5 when *k* = 1. APEX can be considered as an approximate solution to APES or AMES for *k* > 1, and when *k* = 1, it becomes identical to PCA (this can also be verified theoretically).

In the second experiment, we look into the specific error performance of ASE approaches at *k* = 2. Note that there are some slight variations in these error measures for close $\gamma$ values, which is due to the inaccuracy in the estimation of specific error components. Nevertheless, we can observe the following trends. As shown in Fig. 4.6(a) and 4.6(b), we found that the performance improvement of ASE approaches in extracting primary components lies in the reduction of the ambient leakage, though at the cost of introducing more distortion. For ambient component extraction, PCA and Masking yield the least amount of leakage and interference, respectively. Note that the little amount of leakage in PCA and interference in Masking are actually due to the estimation error, since none of them theoretically exist in the extracted ambient components. Nevertheless, the ASE approaches yields moderate amount of these errors, which results in a better overall performance.

In the third experiment, we examine the spatial accuracy of PAE in terms of the diffuseness of the extracted ambient components. As shown in Fig. 4.7(a)-(c), the lowest and highest ICC are achieved with true ambient components and ambient components extracted by PCA, respectively. The ASE approaches outperform the existing approaches, and are more effective in extracting diffuse ambient components at higher *k* and lower $\gamma$. For ICLD of the extracted ambient components as shown in Fig. 4.7(d)-(f), we observed that all approaches extract ambient components with equal level between the two channels, whereas PCA works only for $k = 1$.

In the fourth experiment, we compare the extraction performance as well as the computation time among these PAE approaches. The simulation was carried out on a PC with i5-2400 CPU, 8 GB RAM, 64-bit windows 7 operating system and 64-bit MATLAB 7.11.0. Though MATLAB simulations do not provide precise computation time measurement compared to the actual implementation, we could still obtain the relative computation performance among the PAE approaches. The results of computation time averaged across all the $\gamma$ and *k*

Table 4.4 Average ESR, ICC, and computation time of PAE approaches

| Method | APES | AMES | APEX | PCA [GoJ07b] | Masking [MGJ07] |
|---|---|---|---|---|---|
| $ESR_P$ (dB) | -6.73 | -6.31 | -6.25 | -3.02 | -1.57 |
| $ESR_A$ (dB) | -6.73 | -6.31 | -6.25 | -3.02 | -2.77 |
| ICC of ambient components | 0.19 | 0.22 | 0.42 | 1 | 0.40 |
| Computation time (ms) | 3921.8 | 217.1 | 4.8 | 0.06 | 5.0 |

values are summarized in Table 4.4. It is obvious that the three ASE approaches perform better than PCA and Masking on the average. But when we compare the computation time among APES, AMES, and APEX, we found that AMES is around 20x faster than APES, but is still far away from the computation time of the existing approaches. The APEX, which estimates the ambient phase directly using the phase of the input signals, is over 40x faster as compared to AMES and has similar computational performance as the Masking approach, and hence can be considered as a good alternative ASE approach for PAE. Furthermore, in order to achieve real-time performance (in frame-based processing), the processing time must be less than 4096/44.1 = 92.88 (ms). It is clear that APEX, together with PCA and Masking satisfies this real-time constraint.

### 4.4.3 Experimental results on robustness of ASE approaches

To investigate the robustness of the proposed ASE approaches, we conduct experiments with the input signals containing unequal ambient magnitudes in the two channels. To quantify the violation of the assumption of equal ambient

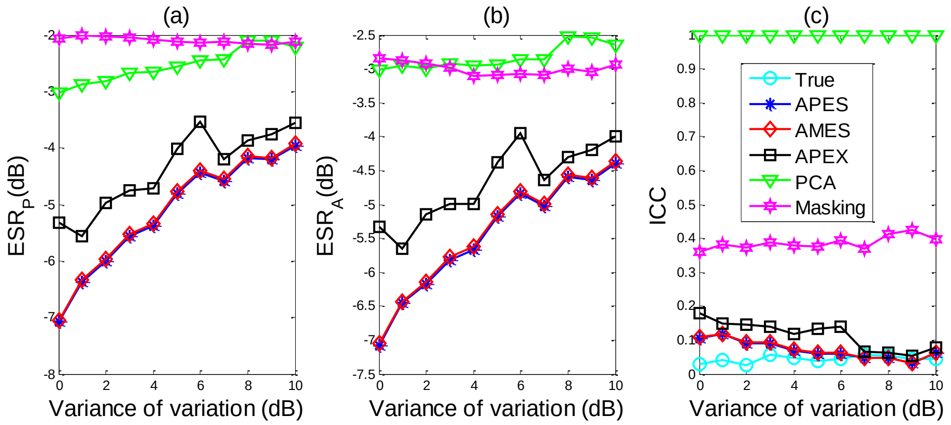


Figure 4.8 Comparison of the performance of PAE approaches in the presence of normally distributed variations in the ambient magnitudes in two channels (with $\gamma = 0.5$, $k$ = 2): (a) ESR$_P$, (b) ESR$_A$, (c) ICC of ambient components.

magnitude, we introduce an inter-channel variation factor $v$ that denotes the range of variation of the ambient magnitude in one channel as compared to the other channel. Let us denote the ambient magnitude in the two channels as $r_0$, $r_1$. The variation of ambient magnitude is expressed as $v = 10\log_{10}\left(r_1/r_0\right) \quad (\text{dB})$. In the ideal case, we always have $v = 0$. To allow variation, we consider $v$ as a random variable with mean equal to 0, and variance as $\sigma^2$. In this experiment, we consider two types of distributions for the variation, namely, normal distribution and uniform distribution, and examine the performance of these PAE approaches with respect to different variance of variations, i.e., $\sigma^2 \in [0,\ 10]$, at $\gamma = 0.5$, and $k$ = 2. We run the experiment 10 times and illustrate the averaged performance in terms of ESR and ICC in Figs. 4.8 and 4.9. We observed that as the variance of the variation increases, the ESR performance of proposed ASE approaches becomes worse, though ICC was not affected much. The ASE approaches are more robust to

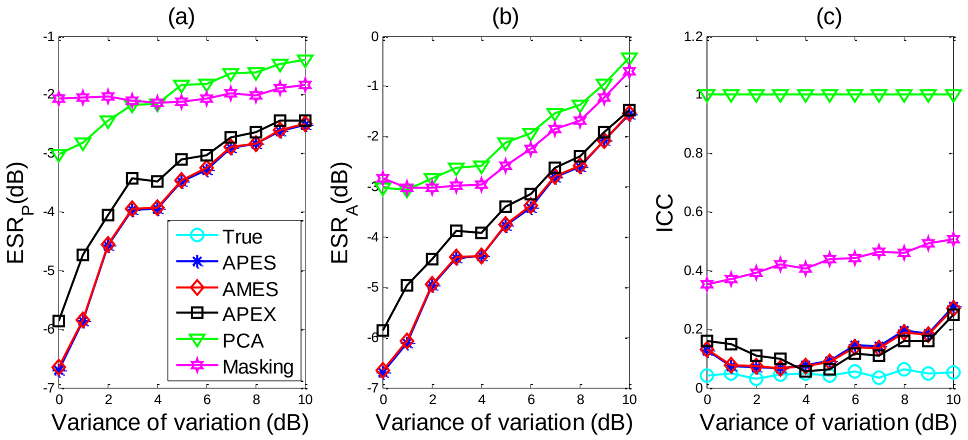


Figure 4.9 Comparison of the performance of PAE approaches in the presence of uniformly distributed variations in the ambient magnitudes in two channels (with $\gamma = 0.5$, $k$ = 2): (a) ESR$_P$, (b) ESR$_A$, (c) ICC of ambient components.

ambient magnitude variations under normal distribution compared to uniform distribution. Compared to PCA and Masking, the proposed approaches are still better with the variance of variation up to 10 dB. Therefore, we conclude that the three ASE approaches are in general robust to ambient magnitude variations.

### 4.4.4 Experimental results on subjective listening tests

Lastly, subjective tests were carried out to evaluate the perceptual performance of these PAE approaches. A total of 17 subjects (15 males and two females), who were all between 20-30 years old, participated in the listening tests. None of the subjects reported any hearing issues. The tests were conducted in a quiet listening room at Nanyang Technological University, Singapore. An Audio Technica MTH-A30 headphone was used. The stimuli used in this test were synthesized using amplitude panned ($k$ = 2) primary components (speech, music, and bee sound) and decorrelated ambient components (forest, canteen, and waterfall sound) based on two values of primary power ratio ($\gamma = 0.3,\ 0.7$) for the duration of 2 to 4 seconds. Both the extraction accuracy and spatial accuracy were examined. The testing procedure was based on MUSHRA [ITU03b], [LNZ14], where a more specific anchor (i.e., the mixture) is used instead of the low-passed anchor, according to recent revision of MUSHRA as discussed in [LNZ14]. The MATLAB GUI was modified based on the one used in [EVH11]. Subjects were asked to listen to the clean reference stimuli and processed stimuli obtained from different PAE approaches, and give a score of 0-100 as the response, where 0-20, 21-40, 41-60, 61-80, and 81-100 represent a bad, poor, fair, good, and excellent quality, respectively. Finally, we analyzed the subjects' responses for the hidden reference (clean primary or ambient components), mixture, and three PAE approaches, namely, Masking [MGJ07], PCA [GoJ07b], and APEX. Note that APEX is selected as the representative of ASE approaches because APES and AMES exhibit very similar extraction results. For each PAE approach, we combine the subjective scores of different test stimuli and different values of primary power ratio, so as to represent the overall performance of these PAE approaches. According to [ITU14], we conducted the post-screening to detect the outliers by excluding the scores of the subject who rates the hidden reference lower than 90. The mean subjective score with 95% confidence interval of the extraction and spatial accuracy for the tested PAE approaches are illustrated in Figs. 4.10. Despite the relatively large variations among the subjective scores that are probably due to the different scales employed by the subjects and the differences among the stimuli, we observe the following trends.

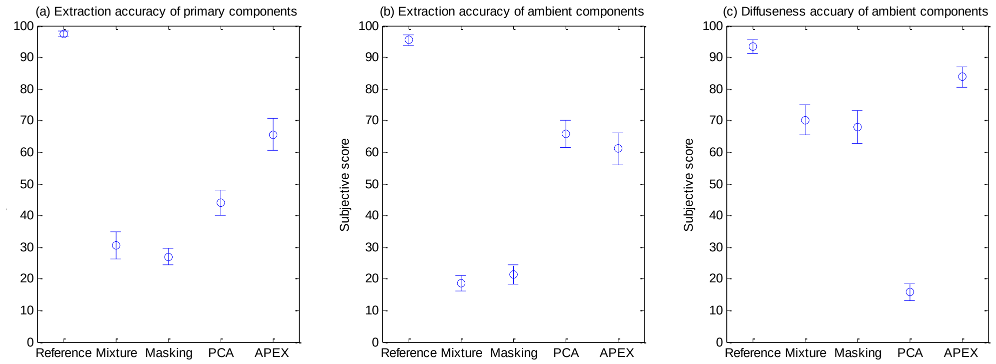


Figure 4.10 Subjective performance (mean with 95% confidence interval) for (a) the extraction accuracy of primary components, (b) the extraction accuracy of ambient components, and (c) diffuseness accuracy of ambient components.

On one hand, we observed that APEX outperforms the other PAE approaches in extracting accurate primary components, as shown in Fig. 4.10(a). In Fig. 4.10(b), APEX, though slightly worse off than PCA, still produces considerable accuracy in ambient extraction. The good perceptual performance of ambient components extracted from PCA lies in the very low amount of primary leakage, as shown in Fig. 4.6(c). On the other hand, we found that the spatial performance were also affected by the undesired leakage signals as compared to the clean reference, as found in the mixtures, which preserve the same spatial quality as the reference, but were rated lower than the reference. With respect to the diffuseness of the ambient components, APEX performs the best, whereas PCA performs poorly. We find that PCA sacrifices on the diffuseness of the extracted ambient components for the sake of a better perceptual extraction performance. A further analysis of the ANOVA results shows that the *p*-values between the APEX and Masking, PCA are extremely small, which reveals that the differences among the performance of these PAE approaches are significant. To sum up the subjective evaluation results, the proposed ASE approaches yield the best performance in terms of extraction and spatial accuracy, which is consistent with our objective evaluation results.

## 4.5 Conclusions

In this chapter, we presented a novel formulation of the PAE problem in the time-frequency domain. By taking advantage of equal magnitude of ambient component in two channels, the PAE problem is reformulated as an ambient spectrum estimation problem. The ASE framework can be considered in two ways, namely, ambient phase estimation, and ambient magnitude estimation. The novel ASE formulation provides a promising way to solve PAE in the sense that the optimal solution leads to perfect primary and ambient extraction, which is unachievable with existing PAE approaches. In this chapter, ASE is solved based on the sparsity of the primary components, resulting in two approaches, APES and AMES. To thoroughly evaluate the performance of extraction error, we proposed an optimization method to compute the leakage, distortion and interference of the extraction error for PAE approaches without analytical solutions.

Based on our experiments, we observed significant performance improvement of the proposed approaches over existing approaches. The improvement on error reduction is around 3-6 dB on average and up to 10-20 dB for lower $\gamma$, which is mainly due to the lower residual error from the uncorrelated ambient components. Moreover, the ASE approaches perform better for mixed signals having heavily panned primary components (e.g., $k$ = 4) than those having slightly panned primary components (e.g., $k$ = 1). In terms of the spatial accuracy, the ASE approaches extract more diffuse ambient components. With respect to the computational efficiency of APES and AMES, the value of *D* is an important factor, where the efficiency of these two ASE approaches can be improved by lowering the precision of the phase/magnitude estimation, without introducing significant degradation on the extraction performance. Furthermore, we found that AMES is an order of magnitude faster than APES under the same setting in MATLAB simulation, but is still not as efficient as existing approaches. For this purpose, we have also derived an approximate solution APEX and verified its effectiveness, as well as its efficiency in our simulation. Besides the ideal situation where the ambient magnitudes are equal in two channels, the robustness of these ASE approaches was also examined by introducing statistical variations to the ambient magnitudes in the two channels of the stereo signal. It was found that the proposed approaches can still yield better results with the variance of variations up to 10 dB. The objective performance of the proposed ASE approaches was also validated in our subjective tests. In the next two chapters, we will study PAE that deals with more complex signals.

# Chapter 5

# Time-Shifting based Primary Ambient Extraction

In practice, PAE is usually applied to the input signals without any prior information. To achieve better extraction of the primary and ambient components, PAE requires the signal model to match the input signal more closely. As presented in the previous two chapters, most work focus on the ideal case. To date, little work has been reported to deal with input signals that do not fulfill all the assumptions of the stereo signal model. In [UsB07], a normalized least-mean-square approach was proposed to address the problem in extracting the reverberation from stereo microphone recordings. Härmä [Har11] tried to improve the performance of PAE by classifying the time-frequency regions of the stereo signal into six classes. Thompson *et al.* [TSW12] introduced a primary extraction approach that estimates the magnitude and phase of the primary component from a multichannel signal by using a linear system of the pairwise correlations. The latter approach requires at least three channels of the input signal and is not applicable to stereo input signals.

This chapter[1] focuses on PAE that deals with real-world stereo input signals that may not fit the typical PAE signal model. In Section 5.1, we discuss the

[1] The work reported in this chapter is an extension from the author's conference paper [HTG13] presented at ICASSP 2013, and Journal paper [HGT15c] published in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, October 2015 issue.

complex cases of real-world signals and identified one of the most frequently occurring cases, known as the primary-complex case. The performance analysis of PCA based PAE in the primary-complex case is presented in Section 5.2. The proposed SPCA based PAE to address the problem in the primary-complex case is discussed in Section 5.3. Section 5.4 presents our comparative evaluation on the performance of PCA and SPCA based PAE using four experiments. Finally, we conclude this chapter in Section 5.5.

## 5.1 Complex cases in PAE

Referring to the stereo signal model discussed in Section 3.1, we shall recall that there are three key assumptions. In practice, none of these three assumptions can be satisfied completely. By relaxing any one of them, we can come up with one complex case. As seen in stereo microphone recordings, movies, and gaming tracks, the primary components in stereo signals can be amplitude panned and time-shifted. In addition, spectral differences can be found in the primary components that are obtained using binaural recording or binaural synthesis based on head-related transfer functions (HRTFs) [Beg00]. We shall classify this type of stereo signals as the primary-complex signals. The primary components in the primary-complex signals usually exhibit partial correlation at zero lag. Other types of complex stereo signals, such as those involving (partially) correlated ambient components, are less common, and hence are not considered in this chapter.

Therefore, we shall focus our study of PAE on two cases, namely, the ideal and primary-complex cases, where the primary components are completely correlated and partially correlated at zero lag, respectively. The performance of PAE is quantified by the measures of extraction accuracy and spatial accuracy. Performance degradation due to the mismatch of the input signal with the stereo signal model and the proposed solution to deal with this mismatch is studied extensively in this chapter. PCA is taken as a representative PAE approach in our study. More in-depth analysis on the performance of PCA based PAE is conducted for the extraction of both the primary and ambient components. In the primary-complex case, the performance degradation of PCA based PAE with respect to the value of primary correlation is discussed, and we find the main cause of low primary correlation and the consequent performance degradation to be the time difference of the primary component. Hence, we propose a time-shifting technique to deal with PAE in the primary-complex case. The time-shifting technique is incorporated into PCA based PAE, resulting in a new approach referred to as the time-shifted PCA (SPCA). A new overlapped output mapping method has also been proposed to avoid the switching artifacts caused by time-shifting. To validate the advantages of the proposed time-shifting technique and verify the improved performance of the proposed approach over conventional approaches more comprehensively, four experiments have been conducted using more realistic test signals. It shall be noted that the proposed time-shifting technique, though studied with PCA in this chapter, can be incorporated into any other PAE approaches that are derived based on the stereo signal model.

## 5.2 Performance of conventional PAE in the primary-complex case

In practice, it is unlikely for any stereo input signals to fulfill all the assumptions stated in Section 3.1. Several non-ideal cases can be defined by relaxing one or more of the assumptions of the stereo signal model. In this chapter, we focus our discussions on one commonly occurring non-ideal case, referred to as the primary-complex case, which defines a partially correlated primary component at zero lag. To investigate the performance of PCA based PAE in the primary-complex case, we shall examine the estimation of $k$ and $\gamma$ first, and then evaluate the performance in terms of extraction accuracy and spatial accuracy.

Considering a stereo signal having a partially correlated primary component at zero lag, the first assumption of the stereo signal model, as stated in (3.3), is allowed to be relaxed to

$$0 < \left[ \phi_{\mathrm{P}} = \frac{\mathbf{p}_0{}^T \mathbf{p}_1}{\sqrt{\left(\mathbf{p}_0{}^T \mathbf{p}_0\right)\left(\mathbf{p}_1{}^T \mathbf{p}_1\right)}} \right] < 1, \tag{5.1}$$

where $\phi_{\mathrm{P}}$ is the correlation coefficient of the primary component at zero lag (primary correlation for short), and the rest of the assumptions in (3.3) and (3.4) remain unchanged. Here, only the positive primary correlation is considered, since the negatively correlated primary component can be converted into positive by simply multiplying the primary component in either channel by -1. In the primary-complex case, the correlations of the input signals at zero lag are computed as:

$$r_{00} = N\left(P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{11} = N\left(k^2 P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{01} = N\phi_{\mathrm{P}} k P_{\mathbf{p}_0}. \tag{5.2}$$

Hence, the estimated $k$ and $\gamma$ are:

$$\hat{k}_{pc} = \phi_{\mathrm{P}} \frac{r_{11} - r_{00}}{2r_{01}} + \sqrt{\left( \phi_{p} \frac{r_{11} - r_{00}}{2r_{01}} \right)^{2} + 1}, \tag{5.3}$$

$$\hat{\gamma}_{pc} = \frac{2r_{01} + \phi_{\mathrm{P}} \left( r_{11} - r_{00} \right) \hat{k}_{pc}}{\phi_{\mathrm{P}} \left( r_{11} + r_{01} \right) \hat{k}_{pc}}, \tag{5.4}$$

where the subscript "*pc*" stands for "the primary-complex case". Clearly, accurate estimation of $k$ and $\gamma$ in the primary-complex case requires the additional knowledge about the primary correlation $\phi_{\mathrm{P}}$. However, this primary correlation is usually unavailable as only the mixed signal is given as input. In PCA based PAE, the estimates of $k$ and $\gamma$ for the ideal case, given in (3.8)-(3.9), are usually employed. In this section, these two solutions are re-expressed as $\hat{k}_{ic}$, $\hat{\gamma}_{ic}$, where the subscript "*ic*" stands for "ideal case". To see how accurate these ideal case estimates are, we substitute (5.2) into (3.8) and (3.9), and compute the ratio between the estimated $k$ and true $k$, and the ratio between estimated $\gamma$ and true $\gamma$ as

$$\Delta k = \frac{\hat{k}_{ic}}{k} = \frac{k^{2} - 1}{2\phi_{\mathrm{P}} k^{2}} + \sqrt{\left( \frac{k^{2} - 1}{2\phi_{\mathrm{P}} k^{2}} \right)^{2} + \frac{1}{k^{2}}}, \tag{5.5}$$

$$\Delta \gamma = \frac{\hat{\gamma}_{ic}}{\gamma} = \frac{k^{2} - 1 + 2\phi_{\mathrm{P}}}{k^{2} + 1}. \tag{5.6}$$

Using (5.5) and (5.6), the ratios of the $k$ and $\gamma$ in the primary-complex case with respect to the primary correlation are plotted in Fig. 5.1. It is clear that $k$ is only correctly estimated (i.e., $\Delta k = 0\ \mathrm{dB}$) when it equals one; and the estimation of $\gamma$ is more accurate (i.e., $\Delta\gamma$ closer to 1) as $k$ increases. The estimations of $k$ and $\gamma$ become less accurate as the primary correlation

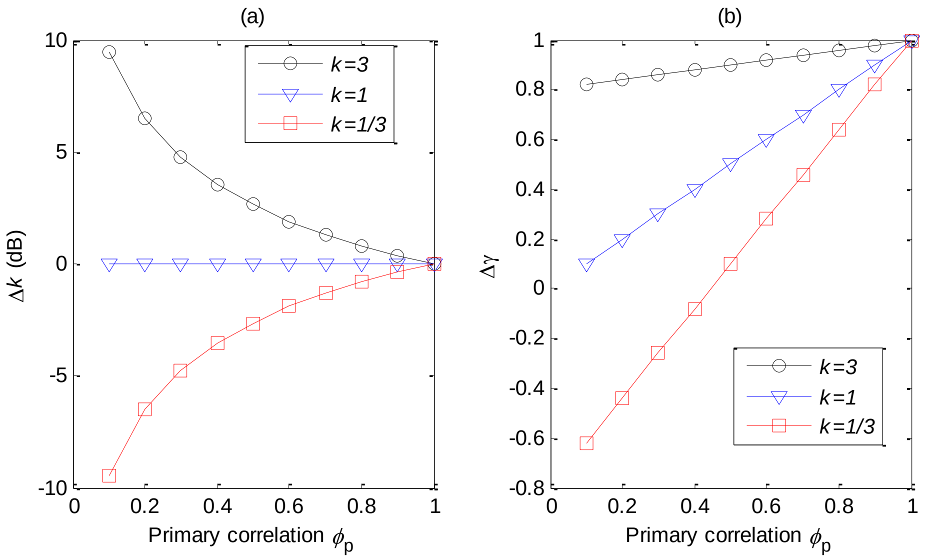


Figure 5.1 Estimation of (a) primary panning factor *k*, and (b) primary power ratio $\gamma$ in the primary-complex case with varying $\phi_{\mathrm{P}}$. The estimations are more accurate when $\Delta k$ and $\Delta\gamma$ are closer to 0 dB and 1, respectively.

decreases from one to zero. The inaccuracy in the estimates of *k* and $\gamma$ results in an incorrect ICLD of the extracted primary components and hence degrades the extraction performance.

Next, we analyze the extraction performance of PCA based PAE in the primary-complex case. First, we rewrite (3.28)-(3.29) using the true primary and ambient components:

$$\begin{aligned}\hat{\mathbf{p}}_{\mathrm{PCA},\,0} &= \mathbf{p}_0 - \mathbf{v} + \frac{1}{1+\hat{k}_{ic}^{\,2}}\left(\mathbf{a}_0 + \hat{k}_{ic}\mathbf{a}_1\right), \\ \hat{\mathbf{p}}_{\mathrm{PCA},\,1} &= \mathbf{p}_1 + \frac{1}{\hat{k}_{ic}}\mathbf{v} + \frac{\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\left(\mathbf{a}_0 + \hat{k}_{ic}\mathbf{a}_1\right),\end{aligned} \tag{5.7}$$

$$\begin{aligned}\hat{\mathbf{a}}_{\mathrm{PCA},\,0} &= \frac{\hat{k}_{ic}^{\,2}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_0 + \mathbf{v} + \frac{-\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_1, \\ \hat{\mathbf{a}}_{\mathrm{PCA},\,1} &= \frac{1}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_1 - \frac{1}{\hat{k}_{ic}}\mathbf{v} + \frac{-\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_0,\end{aligned} \tag{5.8}$$

where $\mathbf{v}=\frac{\hat{k}_{ic}}{1+\hat{k}_{ic}^{2}}\left(\hat{k}_{ic}\mathbf{p}_0-\mathbf{p}_1\right)$ is the interference signal decomposed from the input primary components $\mathbf{p}_0$, and $\mathbf{p}_1$. As compared to the ideal case (where $\mathbf{v}=\mathbf{0}$), this interference $\mathbf{v}$ introduces additional extraction error in the primary-complex case.

To evaluate the PAE performance, two groups of performance measures quantifying the extraction accuracy and spatial accuracy are introduced in Chapter 3. The extraction accuracy is usually quantified by the extraction error, which is given by the error-to-signal ratio (ESR) and is computed as:

$$\begin{aligned}\mathrm{ESR}_{\mathrm{P}}&=0.5\left(\frac{P_{\mathbf{p}_0-\hat{\mathbf{p}}_0}}{P_{\mathbf{p}_0}}+\frac{P_{\mathbf{p}_1-\hat{\mathbf{p}}_1}}{P_{\mathbf{p}_1}}\right),\\ \mathrm{ESR}_{\mathrm{A}}&=0.5\left(\frac{P_{\mathbf{a}_0-\hat{\mathbf{a}}_0}}{P_{\mathbf{a}_0}}+\frac{P_{\mathbf{a}_1-\hat{\mathbf{a}}_1}}{P_{\mathbf{a}_1}}\right).\end{aligned}\tag{5.9}$$

Smaller value of ESR indicates a better extraction.

In the second group of measures, we consider the spatial accuracy by comparing the inter-channel relations of the extracted primary and ambient components with those of the true components. Due to the differences in the spatial characteristics of the primary and ambient components, we shall evaluate these components separately. For the primary components, there are three widely used spatial cues, namely, ICC, ICTD, and ICLD. The accuracy of these cues can be used to evaluate the sound localization accuracy of the extracted primary components [Rum01], [RVE10]. There has been extensive research in ICTD estimation after the coincidence model proposed by Jeffress (see [Jef48], [Yos93], [JSY98], [KaN14] and references therein). Based on the Jeffress model [Jef48], the ICC of different time lags is calculated and the lag number that corresponds to the maximum ICC is determined as the estimated

Table 5.1 Performance of PCA based PAE in the primary-complex case.

| **Measures** | ESR | | |
|---|---|---|---|
| **Primary component** | $\frac{\hat{k}_{ic}{}^4 - 2\phi_{\mathrm{P}} k\hat{k}_{ic}{}^3 + \left(k^2 + k^{-2}\right)\hat{k}_{ic}{}^2 - 2\phi_{\mathrm{P}} k^{-1}\hat{k}_{ic} + 1}{2\left(\hat{k}_{ic}{}^2 + 1\right)^2} + \left(1 + k^{-2} + \frac{k^2 - k^{-2}}{\hat{k}_{ic}{}^2 + 1}\right)\frac{1-\gamma}{4\gamma}$ | | |
| **Ambient component** | $\frac{\hat{k}_{ic}{}^4 - 2\phi_{\mathrm{P}} k\hat{k}_{ic}{}^3 + \left(k^2 + 1\right)\hat{k}_{ic}{}^2 - 2\phi_{\mathrm{P}} k\hat{k}_{ic} + k^2}{\left(1+\hat{k}_{ic}{}^2\right)^2\left(1+k^2\right)}\frac{\gamma}{1-\gamma} + \frac{1}{2}$ | | |
| **Measures** | ICLD | ICC | ICTD |
| **Primary component** | $\hat{k}_{ic}{}^2$ | 1 | 0 |
| **Ambient component** | $\hat{k}_{ic}{}^{-2}$ | 1 | Not applicable |

ICTD. ICLD is obtained by taking the ratio of the signal power between the channels 1 and 0. For the extracted ambient components, we evaluate the diffuseness of these components using ICC and ICLD [SWB06]. Since the ambient component is uncorrelated and relatively balanced in the two channels of the stereo signal, a better extraction of the ambient component is achieved when ICC and ICLD of the ambient component is closer to zero and one, respectively.

In Table 5.1, we summarize the results of the performance measures for the extracted primary and ambient components when PCA based PAE is applied in the primary-complex (i.e., $\phi_{\mathrm{P}} \neq 1$) and ideal cases (i.e., $\phi_{\mathrm{P}} = 1$). To illustrate how the extraction accuracy is influenced by $\phi_{\mathrm{P}}$, the results of ESR using $\gamma \in \{0.2,\ 0.5,\ 0.8\}$ and $k$ = 3, are plotted in Fig. 5.2. It is clear that ESR is affected by the primary correlation $\phi_{\mathrm{P}}$. As shown in Fig. 5.2(a), the error of the extracted primary component decreases as $\phi_{\mathrm{P}}$ approaches one, except for $\gamma$ = 0.2. This exceptional case arises when $\gamma$ is low, and the ambient leakage in the extracted primary component becomes the main contributor for the extraction error. From Fig. 5.1(a), we notice that as $\phi_{\text{P}}$ increases, $\Delta k$ decreases, which leads to the decrease of $\hat{k}_{ic} = \Delta k \bullet k$ ; and hence the contributor from the ambient leakage in ESR$_{\text{P}}$ (i.e., $\left(1+k^{-2}+\frac{k^{2}-k^{-2}}{\hat{k}_{ic}^{\;2}+1}\right)\frac{1-\gamma}{4\gamma}$) increases, which finally leads to the increase of ESR$_{\text{P}}$ for $\gamma$ = 0.2. For the ESR of the extracted ambient component (ESR$_{\text{A}}$) as illustrated in Fig. 5.2(b), we observed that ESR$_{\text{A}}$ decreases gradually as $\phi_{\text{P}}$ increases, which leads to an extracted ambient component having less error. Based on these observations, we find that

1) In the ideal case, where $\phi_{\text{P}} = 1$, the primary and ambient components are extracted with relatively less error.
2) In the primary-complex case, the error of the primary and ambient components extracted in PCA based PAE generally increases for most values of $\gamma$ as $\phi_{\text{P}}$ decreases.
3) It is also found in Table 5.1 that ICC and ICTD in the primary component are always one and zero, respectively. These values imply that the ICTD of the primary component is completely lost after the extraction. The correct ICLD of the primary component can only be obtained when $k$ is accurately estimated.

From the above observations, it is concluded that the performance of PCA based PAE is degraded by the partially correlated primary components of the stereo signal in the primary-complex case. The degraded performance, as observed in PCA, actually originates from the inaccurate estimations of $k$ and $\gamma$. As found in Chapter 3, the linear estimation based PAE approaches are

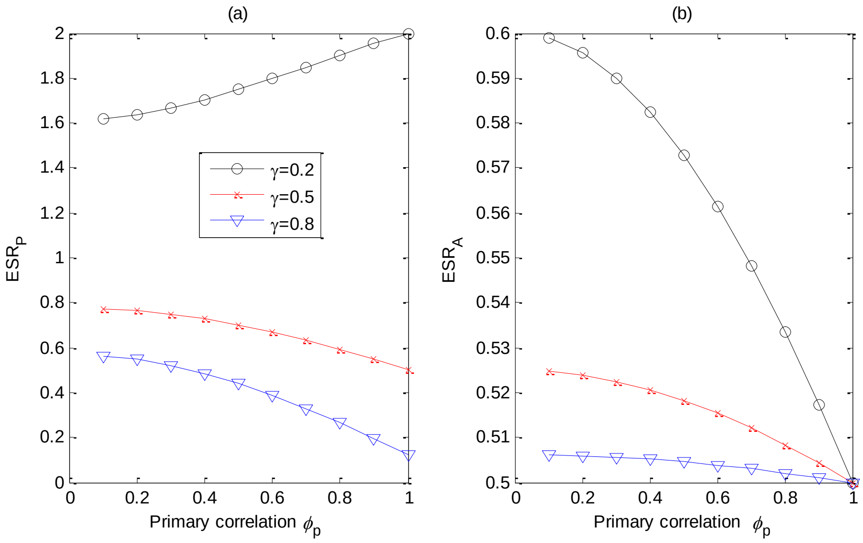


Figure 5.2 ESR of (a) primary extraction and (b) ambient extraction using PCA based PAE in the primary-complex case with varying $\phi_P$ according to the results in Table 5.1. Legend in (a) applies to both plots.

determined by these two parameters. Hence, it can be inferred that these linear estimation based PAE approaches as well as other PAE approaches that are derived based on the basic stereo signal model will encounter a similar performance degradation when dealing with stereo signals having partially correlated primary components.

## 5.3 Time-shifting technique applied in PAE

In the audio of moving pictures and video games, it is commonly observed that the primary components are amplitude panned and/or time-shifted [Wik13], [SWR13], where the latter leads to low correlation of the primary components at zero lag. As mentioned in the previous section, PCA based PAE dealing with such primary-complex signals leads to significant extraction error. Furthermore,

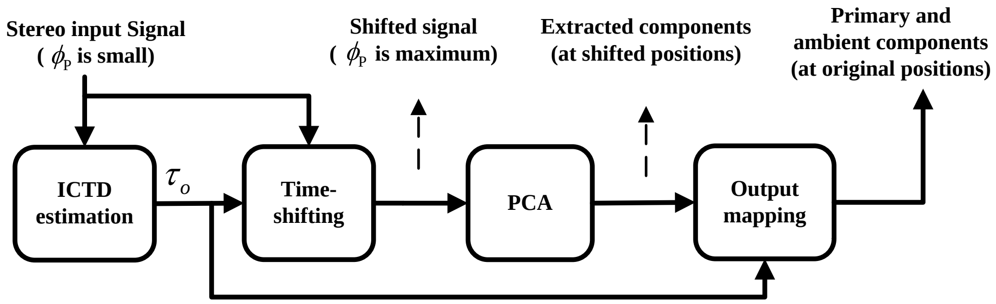


Figure 5.3 Block diagram of SPCA based PAE.

the ICTD of the primary component is completely lost after the extraction. To overcome these issues, we propose a time-shifting technique to be incorporated into PCA based PAE, which results in the proposed approach, namely, the time-shifted PCA (SPCA) based PAE. The proposed approach aims to retain the ICTD in the extracted primary component and time-shifts the primary components to increase the primary correlation, thereby enhancing the performance of PAE.

The block diagram of the proposed SPCA based PAE is shown in Fig. 5.3. In SPCA based PAE, the stereo input signal is first time-shifted according to the estimated ICTD of the primary component. Subsequently, PCA is applied to the shifted signal and extracts primary and ambient components at shifted positions. Finally, the time indices of extracted primary and ambient components are mapped to their original positions using the same ICTD. Let $\tau_o$ denotes the estimated ICTD, the final output for the *n*th sample in the extracted components can be expressed as

$$\begin{aligned}\hat{p}_{\text{SPCA, }0}(n) &= \frac{1}{1+\hat{k}_{ic}^{2}}\left[x_0(n)+\hat{k}_{ic}x_1(n-\tau_o)\right],\\ \hat{p}_{\text{SPCA, }1}(n) &= \frac{\hat{k}_{ic}}{1+\hat{k}_{ic}^{2}}\left[x_0(n+\tau_o)+\hat{k}_{ic}x_1(n)\right],\end{aligned} \tag{5.10}$$

$$\begin{aligned}\hat{a}_{\text{SPCA, }0}(n) &= \frac{\hat{k}_{ic}}{1+\hat{k}_{ic}^{2}}\left[\hat{k}_{ic}x_0(n)-x_1(n-\tau_o)\right],\\ \hat{a}_{\text{SPCA, }1}(n) &= -\frac{1}{1+\hat{k}_{ic}^{2}}\left[\hat{k}_{ic}x_0(n+\tau_o)-x_1(n)\right].\end{aligned} \tag{5.11}$$

It can be seen that the proposed approach is related to delayed-and-sum beamformer [VaB88] in the sense that each extracted component is a weighted sum of the input signals but with a delay or advance being applied in either channel. When ICTD $\tau_o = 0,$ the proposed SPCA based PAE reduces to the conventional PCA based PAE.

As mentioned in previous section, estimation of ICTD can be obtained using various approaches. In this chapter, we apply the Jeffress model [Jef48], which estimates the ICTD of the primary component using the maximum ICC of the primary component at various lags $\phi_{\text{P}}(\tau).$ When only the stereo signal is available, we cannot compute the ICC of the primary component directly. Instead, the ICC of the stereo input signal $\phi_x(\tau)$ is used to estimate the ICTD of the primary component. Due to the uncorrelated ambient component of the stereo signal, which remains uncorrelated after the stereo signal is time-shifted, we find that for each lag $\tau,$

$$\phi_x(\tau) = g\phi_{\text{P}}(\tau), \tag{5.12}$$

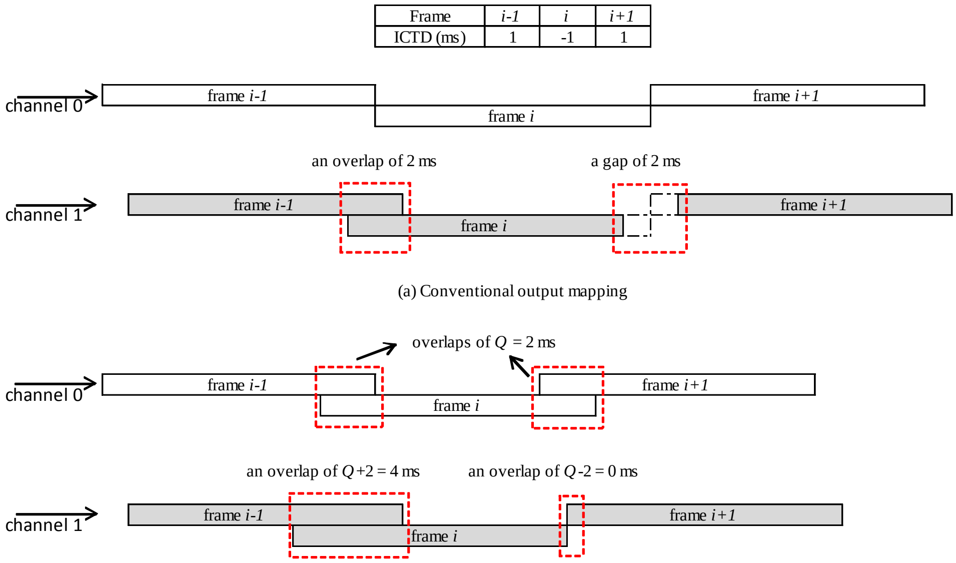
where $g = \sqrt{\dfrac{P_{\mathbf{p}_0}P_{\mathbf{p}_1}}{P_{\mathbf{x}_0}P_{\mathbf{x}_1}}}$ is lag-invariant. Therefore, the ICTD $\tau_o = \arg\max\limits_{\tau}\phi_{\text{P}}(\tau) = \arg\max\limits_{\tau}\phi_x(\tau).$ A detailed study on the estimation of ICTD based on ICC in complex situations is discussed in [FaM04]. Due to the effect of summing localization, the maximum number of lags considered for ICC and ICTD in spatial audio is usually limited to ±1 ms [Bla97]. The positive and negative values of ICTD account for the primary components that are panned to the directions of channel 0 and channel 1 in the auditory scene, respectively. As compared to the conventional PCA based PAE, the estimation of ICTD is one critical additional step, which inevitably incurs more calculations. More specifically, in the conventional PCA, the cross-correlations (i.e., $\phi_x(0)$) is only computed once. By contrast, the proposed SPCA requires a total of 89 times of cross-correlations (i.e., $\phi_x(\tau)$, $\forall\tau\in[-44,44]$, at a sampling rate fs = 44.1 kHz). One way to reduce the additional computation load is to increase the sample step size in ICTD estimation. For instance, computing only the cross-correlations with odd (or even) indices can reduce the additional computation load by half, at the cost of reducing the resolution of ICTD estimation.

The time-shifting operation is achieved by keeping the signal in channel 0 unchanged but delaying (or advancing) the signal in channel 1 by a duration equal to ICTD when ICTD ≤ 0 (or ICTD > 0). When the amounts of shifts in two successive frames are not the same, a proper mapping strategy is required to shift back the primary and ambient components that are extracted from the shifted signal to the original positions. To show how the change of ICTD affects the final output mapping, we consider two extreme cases, as illustrated in Fig. 5.4. The table in the top middle of Fig. 5.4 shows the ICTDs of three successive frames considered for these two cases. In the first case, we consider maximum ICTD decrease, i.e., the ICTD of frame *i-1* is 1 ms, which is decreased to -1 ms in frame *i*. In the second case, we consider maximum ICTD increase, that is, as compared to the frame *i*, the ICTD of frame *i+1* is increased to 1 ms. Consequently, the decrease and increase of ICTDs in these two cases lead to a 2 ms overlap and gap in channel 1 between these frames, respectively, as shown in Fig. 5.4(a). To generalize these two extreme cases, let us consider the change of ICTD in two successive frames as $\Delta\tau_o(i) = \tau_o(i) - \tau_o(i-1)$. Hence, we have

$$\text{Samples between the two frames of the extracted components in channel 1}$$
$$= \begin{cases} \text{overlap of } |\Delta\tau_o(i)|, & \Delta\tau_o(i) < 0 \\ \text{no overlap or gap}, & \Delta\tau_o(i) = 0. \\ \text{gap of } \Delta\tau_o(i), & \Delta\tau_o(i) > 0 \end{cases} \tag{5.13}$$

To retain the ICTD, a straightforward mapping method is to set the amplitude of the samples of the gap to zero and averaging the overlapped samples in a cross-fading manner. However, it can be easily understood and also revealed in our informal listening tests that perceivable switching artifacts are introduced by the gaps. This is because the gaps are not caused by the silence of the primary components, but are artificially created as a result of the increased ICTD.

To avoid the switching artifacts, all successive frames should be overlapped such that no gap between the frames can be found even when the ICTD increase reaches its maximum. The proposed overlapped output mapping strategy is depicted in Fig. 5.4(b). Let the duration of the overlapping samples in the stereo signals be *Q* ms. As compared to the conventional output mapping in Fig. 5.4(a), different amount of overlapping samples are found in both channels in Fig. 5.4(b). In channel 0, exact *Q* ms between each two frames is overlapped, while in channel 1, the duration of overlapping samples varies from frame to frame according to the change in the ICTDs. That is,



(a) Conventional output mapping

(b) Overlapped output mapping

Figure 5.4 An illustration of two output mapping strategies in the extreme cases: (a) conventional; (b) overlapped. The two channels 0 and 1 are depicted in white and grey, respectively. The table in the top middle shows the ICTDs for three successive frames. The value of *Q* in this example is selected as 2 ms.

$$\text{Samples between the two frames of the extracted components in channel 1} = \text{overlap of } \left[ Q*10^{-3}*\text{fs-}\Delta\tau_o(i) \right]. \quad (5.14)$$

To correspond to the two extreme cases, the duration of overlapping samples in channel 1 would be from *Q*-2 ms to *Q*+2 ms. In order to ensure no gap is found between any two successive frames, the duration of overlapping samples must be equal to or greater than 2 ms, i.e., $Q \geq 2$ ms. As shown in Fig. 5.4(b), where *Q* is chosen as the lowest value, i.e., *Q* = 2 ms, we find that even in the extreme case of maximum ICTD increase from frame *i* to frame *i+1*, there is no gap in channel 1. Therefore, no matter how much the ICTD changes, all frames can be handled appropriately without gap artifacts. Increasing *Q*

Table 5.2 Specifications of the four experiments

| # | Input signal | Primary component | Ambient component | Settings |
|---|---|---|---|---|
| 1 | Synthesized | Speech | Lapping wave | Fixed direction; different values of $\gamma$ |
| 2 | Synthesized | Shaking matchbox | Lapping wave | Panning directions with close $\gamma$ |
| 3 | Synthesized | Direct path of speech | Reverberation of speech | Varying directions with different $\gamma$ |
| 4 | Recorded | Speech | Canteen sound | Three directions with close $\gamma$ |

would also smoothen the extracted components, especially when the direction of the primary components changes rapidly. It is noted from (5.14) that the actual overlapping samples in different frames and channels can be varying. Thus, the cross-fading technique is required to adapt to these variations of the overlapping samples.

Based on the above discussions, we shall see that the proposed time-shifting and overlapped output mapping techniques work independently from PCA. Therefore, the same time-shifting and output mapping technique in the proposed SPCA can be applied seamlessly to improve the performance of many other existing PAE approaches, including time-frequency masking [AvJ04], PCA based approaches [God08], [JHS10], [BJP12], and other linear estimation based PAE approaches as discussed in Chapter 3, as well as ambient spectrum estimation based approaches as discussed in Chapter 4. However, it shall be noted that the ICTD estimation and time-shifting operations would incur additional computation and memory cost.

## 5.4 Experiments and discussions

To validate the performance of the proposed SPCA based PAE, a number of experiments[1] were conducted. As the focus of this chapter is to examine PAE with partially correlated primary components, we shall consider only one dominant source in the primary component of the stereo signal. Experimental results for PAE with time-shifting on multiple dominant sources can be found in Chapter 6. In this section, we present the results from four different experiments. To perform an accurate comparative analysis between PCA and SPCA, we manually synthesized directional signals and mixed them with ambient signals in the first two experiments. The first and second experiments considered static and moving primary component, respectively. In the first experiment, we compared the extraction performance of PCA and SPCA with respect to $\gamma$. While the direction of the primary component was fixed in the first experiment, the second experiment examined the estimation of the panning directions of the primary components using PCA and SPCA with $\gamma$ being close across the frames. The third experiment evaluated how PCA and SPCA perform when dealing with reverberation type of ambient components. To evaluate these two PAE approaches in a more realistic scenario, the fourth experiment was conducted using recorded signals of primary and ambient sound tracks that were played back over loudspeakers around a dummy head. Detailed specifications of the four experiments are given in Table 5.2.

[1] The source code and demo tracks are available: http://jhe007.wix.com/main#!research/c24xx

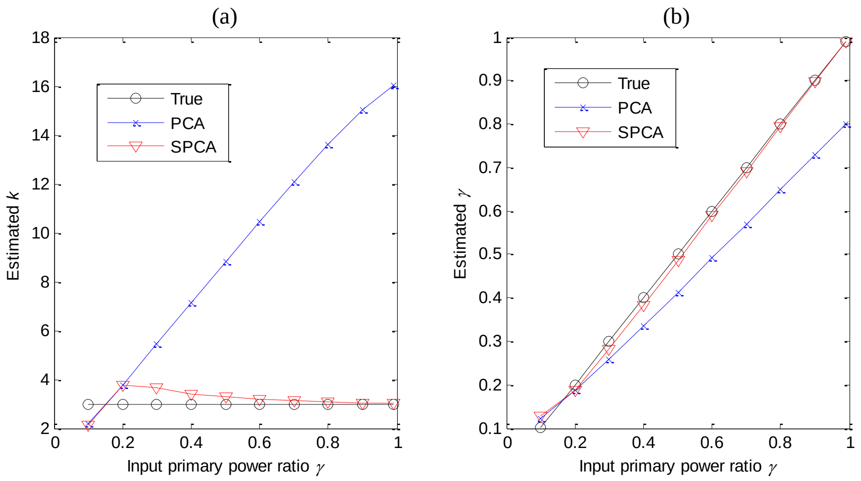


Figure 5.5 Comparison of the estimation of (a) *k* and (b) $\gamma$ between PCA and SPCA based PAE in the primary-complex case.

### 5.4.1 Experiment 1: fixed direction

In the first experiment, a speech clip was selected as the primary component, which is amplitude panned by *k* = 3 and time-shifted by $\tau_o = 40$ samples at a sampling rate of 44.1 kHz, both correspond to the direction of channel 1. The ambient component was taken from a stereo recording of lapping wave with low correlation (less than 0.1) and close to unity power ratio between the two channels. Subsequently, the primary and ambient components were linearly mixed based on the values of $\gamma$ ranging from 0 to 1. Finally, the extraction performance of PCA and SPCA was evaluated using the performance measures introduced in Section 5.2. Note that the correlation coefficient of the tested primary component at zero lag is 0.17, which is increased to one after time-shifting the synthesized signal by 40 samples according to the estimated ICTD. The unity correlation implies that the primary component is completely correlated in SPCA.

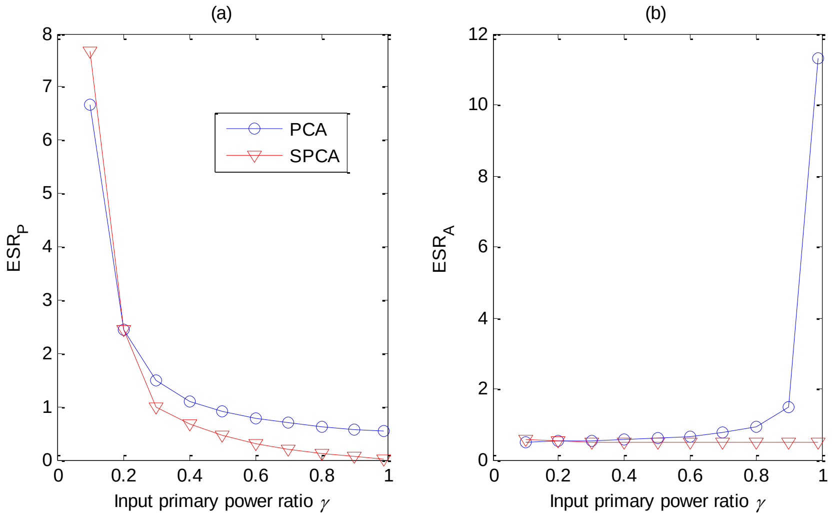


Figure 5.6 ESR of (a) primary extraction and (b) ambient extraction using PCA and SPCA in the primary-complex case. Legend in (a) applies to both plots.

The results of the performance measures of PCA and SPCA are shown in Figs. 5.5-5.7. In Fig. 5.5, there are significant errors in the estimations of *k* and $\gamma$ in PCA, which are estimated more accurately in SPCA. Fig. 5.6 summarizes the ESR of PAE using PCA and SPCA. For primary extraction as shown in Fig. 5.6(a), significant reduction (more than 50%) of ESR is obtained using SPCA when $\gamma \geq 0.5$. Based on Fig. 5.6(b), SPCA extracts the ambient components with smaller ESR than PCA, especially when $\gamma$ is high (more than 50% reduction for $\gamma \geq 0.8$). The significant improvement lies in the reduction of the leakage from the primary components in the extracted ambient component.

SPCA also outperforms PCA in terms of spatial accuracy of the extracted primary and ambient component. As shown in Fig. 5.7(a), the ICTD of the primary component extracted by SPCA is closer to the ICTD of the true primary component for $\gamma \geq 0.3$. When the primary components become too weak in the stereo signals, the estimation of ICTD in SPCA is less accurate. For

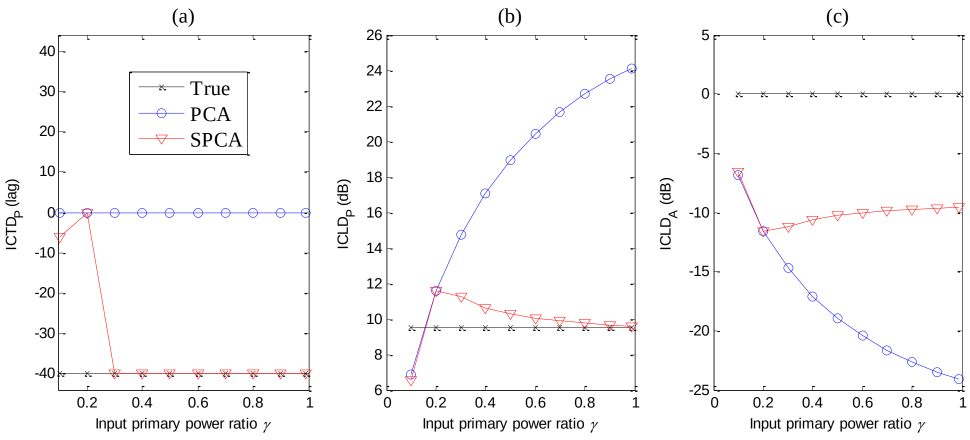


Figure 5.7 Comparison of spatial accuracy in PAE using PCA and SPCA in the primary-complex case. (a) ICTD in the extracted primary components; (b) ICLD in the extracted primary components; (c) ICLD in the extracted ambient components. Legend in (a) applies to all plots.

the ICLD whose just-noticeable difference (JND) is generally below 3 dB [Fal06b], we found that the ICLD of the primary component extracted by SPCA is significantly closer to the ICLD of the true primary component, as shown in Fig. 5.7(b). Therefore, the directions of the primary components extracted by SPCA would be more accurately reproduced and localized. For ambient extraction, we observed that the ICLD of the extracted ambient component for SPCA is closer to 0 dB as compared to PCA, as shown in Fig. 5.7(c). Even though neither approach can extract an uncorrelated and balanced ambient component, a relatively better ambient extraction is obtained with SPCA. Similar to the ideal case, this drawback of ambient extraction is an inherent limitation of PCA [HTG14]. Post-processing techniques like decorrelation [Fal06b] and post-scaling [Fal06], [BJP12] can be applied to further enhance ambient extraction. To sum up the first experiment, we can verify that when dealing with PAE having a directional primary component with time and level

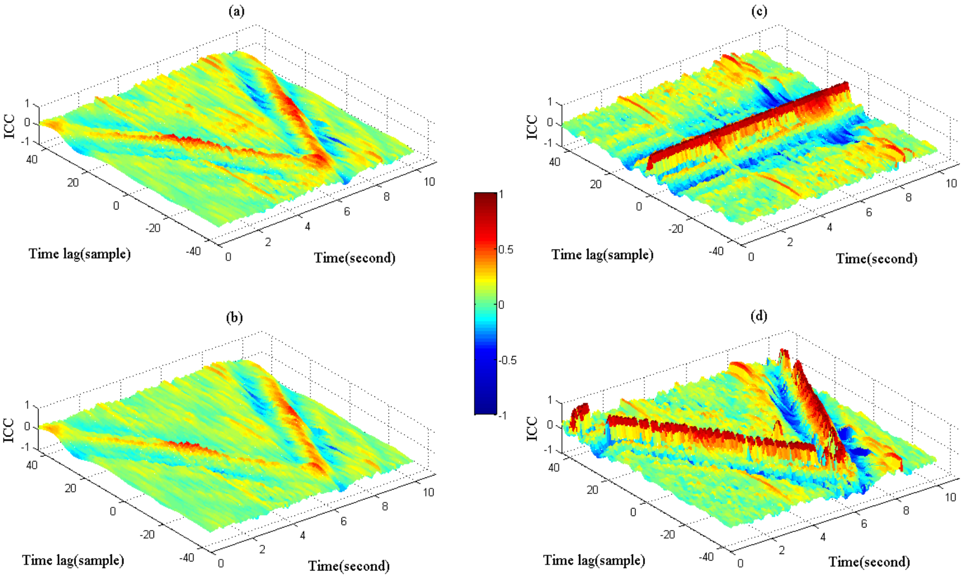


Figure 5.8 Short-time cross-correlation function of (a) true primary component; (b) stereo signal with mixed primary and ambient components; (c) primary component extracted using PCA; (d) primary component extracted using SPCA. Frame size is 4096 samples with 50% overlap.

differences, SPCA extracts the primary and ambient components more accurately than PCA.

### 5.4.2 Experiment 2: panning directions

In the second experiment, a binaural recording of a matchbox sound shaking around the dummy head in the anti-clockwise direction was taken as the primary component, and a wave lapping sound was used as the ambient component. The four plots in Fig. 5.8 illustrate the short-time cross-correlation of the true primary component, mixed signal, primary component extracted by PCA, and primary component extracted by SPCA. The positions of the peaks on the mesh of these plots represent the direction of the primary components, where the time lag at 40 represents extreme left and –40 represents extreme

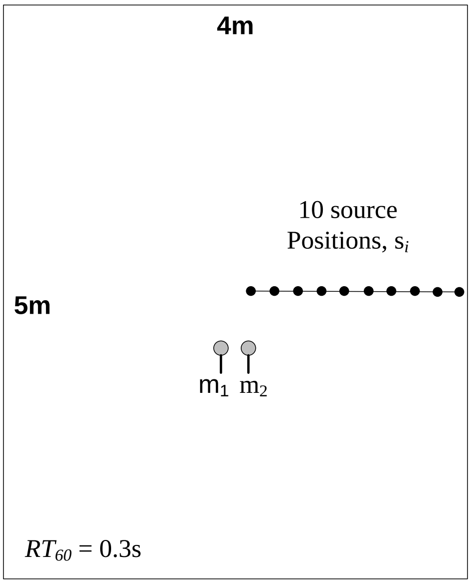


Figure 5.9. Specifications of room, microphone positions and source positions in the reverberation experiment.

right. The anti-clockwise panning of the primary component around the head, as shown in Fig. 5.8(a), becomes less obvious after mixing with the ambient component, as shown in Fig. 5.8(b). Comparing the correlation of the primary component extracted using PCA and SPCA, as shown in Fig. 5.8(c) and 5.8(d), respectively, we can easily verify that only SPCA based PAE preserves the spatial cues of the primary component from the mixed stereo signal. This experiment confirms that SPCA can correctly track the moving directions of the primary components and thus leads to an improved extraction performance with more accurate spatial cues, as compared to PCA.

### 5.4.3 Experiment 3: reverberation ambience

In the third experiment, we considered the extraction of a direct signal and its reverberation from a stereo recording in a reverberant room. For the purpose of a more accurate evaluation, simulated room impulse responses (RIRs) were

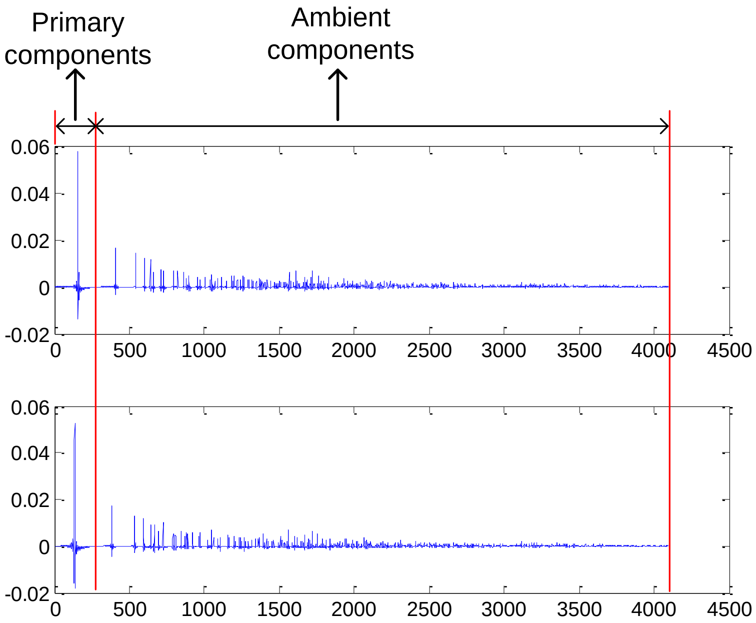


Figure 5.10 An example of the generated RIR and the division of the response for primary and ambient components.

used. The RIR was generated using the software from [Hab14], which is created using the image method [AlB79]. As specified in Fig. 5.9, the size of the room is $5\times4\times6$ m$^3$ with reverberation time $RT_{60}$ set as 0.3s. For the RIR generation, positions for two microphones were set as $m_1(2, 1.9, 2)$ and $m_2(2, 2.1, 2)$. The positions of a speech source varied in 10 locations (one at a time) in a straight line, as $(2.5, s_i, 2)$ with $s_i = 1.9+0.2*i$, $i = 1, 2, \ldots, 10$. The length of the RIR is 4096 samples with sampling frequency at 44.1 kHz. In either channel, the mixed signal was obtained by convolving the source with the generated RIR. The true primary components were synthesized by convolving only the direct paths with the source, while the remainder paths are used as the responses for the synthesis of the true ambient components, as shown in Fig. 5.10. It shall be noted that in this experiment, the primary and ambient components are correlated.

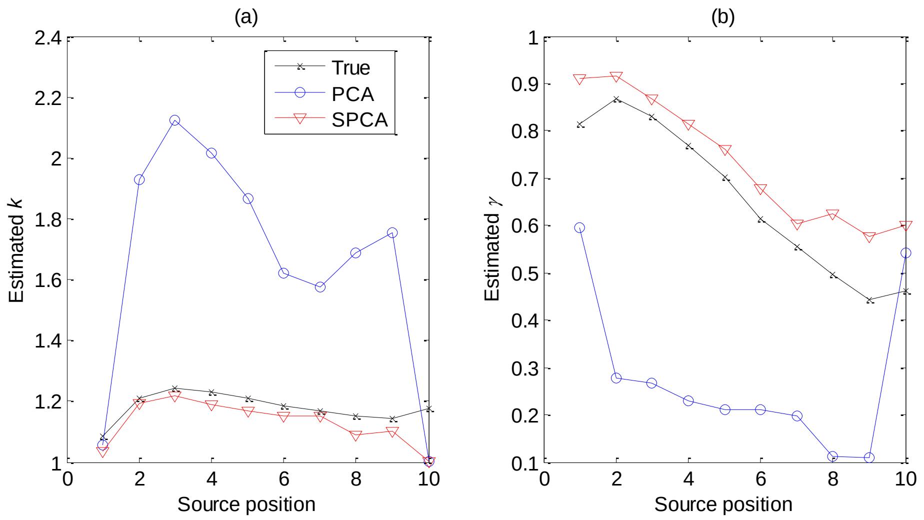


Figure 5.11 Comparison of the estimation of (a) $k$ and (b) $\gamma$ between PCA and SPCA based PAE in the reverberation experiment. Legend in (a) applies to both plots.

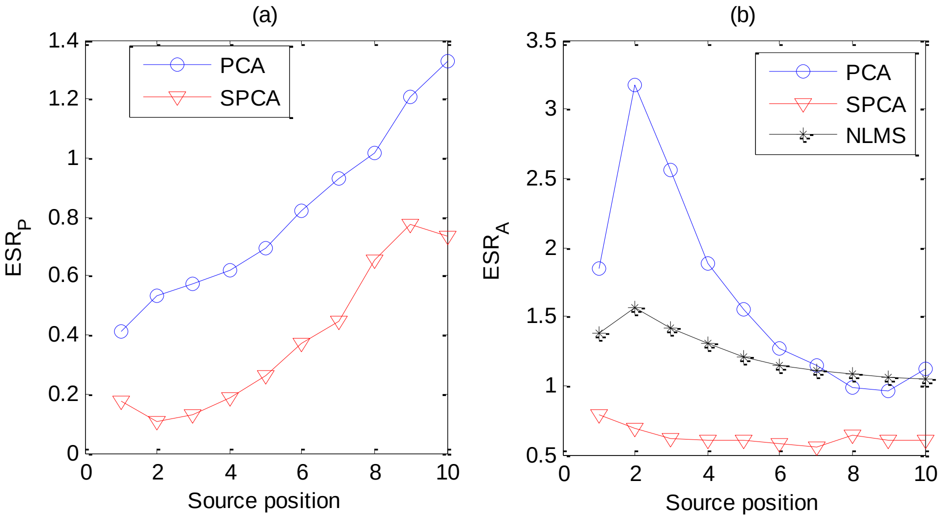


Figure 5.12 ESR of (a) primary extraction and (b) ambient extraction using PCA and SPCA in the reverberation experiment. The NLMS approach [UsB07] is included in (b) for comparison of ambient extraction performance.

Performance of PAE using PCA and SPCA is compared in Figs. 5.11-5.13. It can be observed clearly in these figures that as compared to PCA based PAE, SPCA based PAE can estimate $k$ and $\gamma$ much closer to their true values, thereby yielding a smaller ESR in both primary and ambient extraction, as well

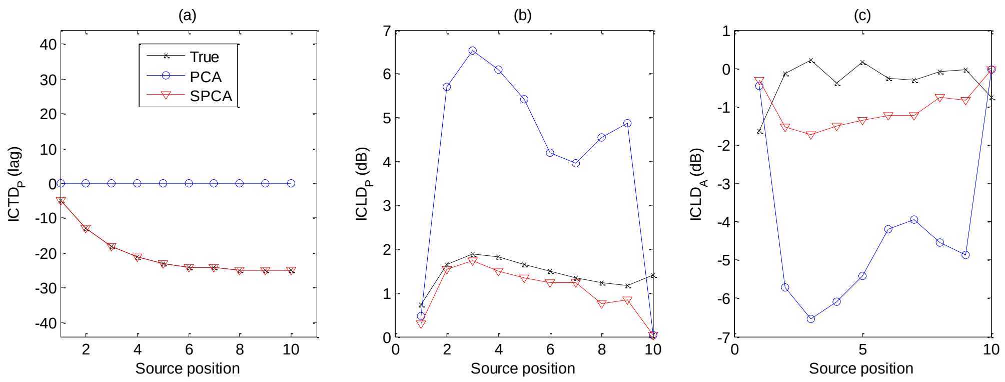


Figure 5.13 Comparison of spatial accuracy in PAE using PCA and SPCA in the reverberation experiment. (a) ICTD in the extracted primary components; (b) ICLD in the extracted primary components; (c) ICLD in the extracted ambient components. Legend in (a) applies to all plots.

as having spatial cues (i.e., ICTD, ICLD) closer to the true values. In particular, we have also applied the normalized least-mean-square (NLMS) approach proposed by Usher [UsB07] in the ambient extraction. As shown in Fig. 5.12 (b), the proposed SPCA approach also outperforms NLMS significantly.

### 5.4.4 Experiment 4: recordings

In the fourth experiment, we tested and compared these PAE approaches using recorded signals. The measurements were conducted in a recording room ($5.4 \times 3.18 \times 2.36$ m$^3$) with a reverberation time of 0.2s at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. The layout of the experiment setup is illustrated in Fig. 5.14. Four loudspeakers $A_1$ to $A_4$ were used to reproduce the ambient sound of a canteen. The primary component, a speech signal, was played back over loudspeaker P, which was placed at each of the three positions with 0°, 45°, and 90° azimuth in the horizontal plane. At the center of the room, a dummy head, which was fitted

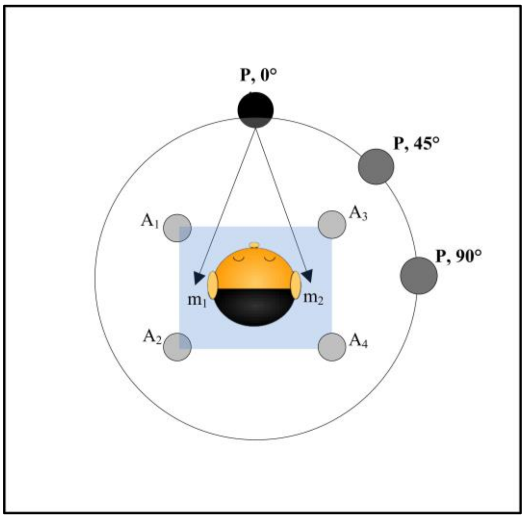


Figure 5.14. Layout of the fourth experiment setup. Four ambient loudspeakers are located at $A_1$-$A_4$. The primary loudspeaker P is positioned at one of the three directions 0 °, 45 °, 90 °in the horizontal plane with a radius of 1.5 meter. Two microphones $m_1$ and $m_2$ are mounted onto the two ears of the dummy head.

with a pair of microphones mounted on the two ears, was used to record the simulated sound scene. To evaluate the performance of the PAE approaches, the "ground truth" reference signals of this experiment (i.e., the true primary and ambient components) were recorded by muting either the one-channel primary loudspeaker or the four-channel ambient loudspeakers.

The performance of PCA and SPCA based PAE are summarized in Tables 5.3, 5.4 and 5.5. In Table 5.3 and Table 5.4, the performance of the two PAE approaches is examined by comparing $\gamma$, *k*, and the spatial cues with their true values, respectively. We observed that SPCA based PAE yields much closer results to the true values as compared to PCA based PAE for all directions of the primary component. From Table 5.5, we observed that the values of ESR in SPCA based PAE are lower (up to 50%) than those in PCA based PAE. These

Table 5.3 Comparison of $\gamma$, *k* in the fourth experiment.

| | $\gamma$ | | | *k* | | |
|---|---|---|---|---|---|---|
| θ | 0° | 45° | 90° | 0° | 45° | 90° |
| True | 0.81 | 0.79 | 0.86 | 0.95 | 1.47 | 1.81 |
| PCA | 0.66 | 0.31 | 0.57 | 0.93 | 6.06 | 3.18 |
| **SPCA** | **0.76** | **0.73** | **0.72** | **0.94** | **1.54** | **2.18** |

Table 5.4 Comparison of spatial cues in the fourth experiment.

| | ICTD$_P$ | | | ICLD$_P$ (dB) | | | ICLD$_A$(dB) | | |
|---|---|---|---|---|---|---|---|---|---|
| θ | 0° | 45° | 90° | 0° | 45° | 90° | 0° | 45° | 90° |
| True | 1 | -17 | -31 | -1.02 | 7.74 | 11.90 | 1.03 | 1.18 | 1.03 |
| PCA | 0 | 0 | 0 | -1.46 | 36.03 | 23.11 | 1.46 | -36.03 | -23.11 |
| **SPCA** | **1** | **-17** | **-31** | **-1.26** | **8.65** | **15.60** | **1.26** | **-8.65** | **-15.60** |

Table 5.5 Comparison of ESR in the fourth experiment.

| | Primary component | | | Ambient component | | |
|---|---|---|---|---|---|---|
| θ | 0° | 45° | 90° | 0° | 45° | 90° |
| PCA | 0.27 | 0.64 | 0.88 | 1.08 | 1.89 | 2.49 |
| **SPCA** | **0.21** | **0.31** | **0.34** | **0.81** | **1.02** | **1.39** |

observations from the fourth experiment indicate clearly that SPCA based PAE outperforms PCA based PAE in more practical situations.

## 5.5 Conclusions

In this chapter, we investigated the performance of PCA based PAE in the ideal and primary-complex cases. The performance of PAE was evaluated on extraction accuracy and spatial accuracy. In practice, the conventional PCA based PAE exhibits severe performance degradation when dealing with the input signals under the primary-complex case, where the primary component is partially correlated at zero lag. Without the knowledge of the correlation of the primary component, the two important parameters primary panning factor and primary power ratio of the stereo signal cannot be estimated accurately. Furthermore, it was found that as the primary correlation decreases, the error in the primary and ambient components extracted by PCA based PAE generally increases. Based on this finding, the proposed SPCA based PAE approach maximizes the primary correlation by appropriately time-shifting the input signals prior to the extraction process. Overlapped output mapping method with a minimum duration of 2 ms overlapping is required to avoid the switching artifacts introduced by time-shifting. As compared to the conventional PCA based PAE, the proposed approach retains the ICTD and corrects the ICLD of the extracted primary component, as well as reduces the extraction error by as much as 50%. With the improved performance of the proposed approach validated using synthesized signals and real-world recordings in our experiments, we conclude that the proposed time-shifting technique can be employed in PAE to handle more generic cases of stereo signals that contain partially correlated primary components. In the following chapter, we will discuss some ideas for PAE to handle an even more complex case, i.e., primary components with multiple sources coming from different directions.

# Chapter 6

# Multiple Source based Primary Ambient Extraction

In this chapter[1], we investigate an even more complex case in PAE. The basic stereo signal model introduced in Chapter 3 limits the number of the dominant source in the primary components to be only one. This assumption generally holds considering that each signal frame is quite short. However, it is still very likely to encounter the exceptional case where there are multiple dominant sources in the primary components. Conventional approaches that ignore this difference will not work well and a robust PAE approach must be devised to handle such cases. For this purpose, we will discuss two approaches to improve the performance of PAE under the case of multiple dominant concurrent sources. The first approach, known as the subband technique, is studied in Section 6.1, and Section 6.2 details the second approach referred to as the multi-shift technique. Similar to Chapter 5, PCA based PAE approaches are selected for our testing. Since it is the primary components that incur the challenge, we shall focus on the extraction of primary components, and the ambient components can be obtained by subtracting the extracted primary components from the mixed signal. Further discussions and conclusions are presented in Section 6.3.

[1] The work reported in this chapter is an extension from the author's conference papers [HGT14] presented in ICASSP 2014 and [HeG15] presented in ICASSP 2015.

## 6.1 Subband technique and frequency bin partitioning

In this section, we focus on the study of subband PAE in the case of multiple sources. First, we transform the time domain time-shifted PCA based PAE into frequency domain. Next, we discuss in detail the most important step of frequency-domain PAE, i.e., partitioning of the frequency bins. Subsequently, a series of simulations are presented to validate the PAE approaches.

### 6.1.1 Time-shifted PCA in frequency domain

First, we consider PAE with one dominant source in primary components in the frequency domain by converting the previous time-domain analysis into frequency domain. From (5.10)-(5.11), only parameters primary panning factor *k* and ICTD $\tau_o$ are relevant to the extracted primary components in PCA and SPCA, and both parameters are computed using the correlations. Therefore, we shall see how correlations are computed in frequency domain. As discussed in [WSL06], the correlation of different lag $\tau$ (in samples) between two signals $\mathbf{x}_i$ and $\mathbf{x}_j$ can be computed by

$$r_{ij}(\tau)=\begin{cases} IDFT\left(\mathbf{X}_i^*(l)\mathbf{X}_j(l)\right), \tau \geq 0 \\ IDFT\left(\mathbf{X}_i(l)\mathbf{X}_j^*(l)\right), \tau < 0 \end{cases}, \tag{6.1}$$

where $\mathbf{X}_i(l)$ is the *l*th bin of the DFT of $\mathbf{x}_i$ and * denotes complex conjugate. The ICTD is determined based on the maximum of the cross-correlation

$$\tau_o = \arg\max_{\tau}\left\{r_{01}(\tau)\right\}. \tag{6.2}$$

Time-shifting in time domain is equivalent to phase-shifting in frequency domain [Mit06], that is,

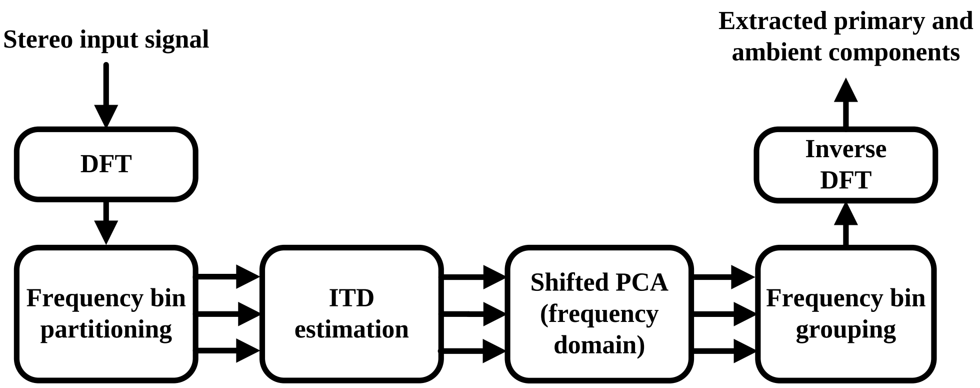

Stereo input signal
Extracted primary and ambient components
DFT
Inverse DFT
Frequency bin partitioning
ITD estimation
Shifted PCA (frequency domain)
Frequency bin grouping

Figure 6.1 Block diagram of frequency bin partitioning based PAE in frequency domain

$$\mathbf{x}_i\left[\left(n-\tau_o\right)_N\right]\overset{DFT}{\leftrightarrow}\mathbf{X}_i\left(l\right)e^{-j2\pi l\tau_o/N}. \tag{6.3}$$

Thus, we can rewrite (5.10) in the frequency domain as

$$\begin{aligned}\hat{\mathbf{P}}_{\text{SPCA},0}\left(l\right)&=\frac{1}{1+k^2}\left[\mathbf{X}_0\left(l\right)+k\mathbf{X}_1\left(l\right)e^{-j2\pi l\tau_o/N}\right],\\ \hat{\mathbf{P}}_{\text{SPCA},1}\left(l\right)&=\frac{k}{1+k^2}\left[\mathbf{X}_0\left(l\right)e^{j2\pi l\tau_o/N}+k\mathbf{X}_1\left(l\right)\right].\end{aligned} \tag{6.4}$$

### 6.1.2 Frequency bin partitioning

To effectively handle multiple sources in the primary components, frequency bins of the input signal are grouped into several partitions, as shown in Fig. 6.1. In each partition, there is only one dominant source and hence one corresponding value of *k* and $\tau_o$ is computed. Ideally, the number of partitions should be the same as the number of sources, and the frequency bins should be grouped in a way such that the magnitude of one source in each partition is significantly higher than the magnitude of other sources. However, the number and spectra of the sources in any given input signals are usually unknown. Hence, the ideal partitioning is difficult or impossible to achieve.

Alternatively, we consider two types of feasible partitioning methods, namely, fixed partitioning and adaptive partitioning. Regardless of the input signal, the fixed partitioning classifies the frequency bins into a certain number of partitions uniformly [AvJ04], [Fal06] or non-uniformly, such as equivalent rectangular bandwidth (ERB) [FaB03]. By contrast, adaptive partitioning takes into account of the input signal via the top-down (TD) or bottom-up (BU) method. BU method starts with every bin as one partition and then gradually reduces the number of partitions by combining the bins. Conversely, TD starts from one partition containing all frequency bins and iteratively divides each partition into two sub-partitions, according to certain conditions. As the number of partitions is usually limited, TD is more efficient than BU, and hence preferred.

To determine whether one partition requires further division, ICC-based criteria are proposed in TD partitioning. First, if the ICC of the current partition is already high enough, we consider only one source is dominant in the current partition and cease further division of the partition. Otherwise, the ICCs of the two divided sub-partitions are examined. The partitioning is continued only when at least one of two ICCs of the sub-partitions becomes higher, and neither ICC of the sub-partitions becomes too small, which indicates that no source is dominant. Suppose the ICCs of the current partition, and two uniformly divided sub-partitions are $\phi_0$, $\phi_1$, $\phi_2$, as shown in Fig. 6.2. For generality, a higher threshold of ICC $\phi_H$ and a lower threshold $\phi_L$ are introduced. Thus, we propose the following three criteria for the continuation of partitioning in TD:

a) $\phi_0 < \phi_H$, and

b) $\mathrm{Max}(\phi_1, \phi_2) > \phi_0$, and

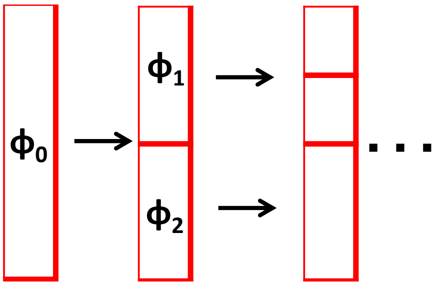


Figure 6.2 An illustration of top-down partitioning

c) Min($\phi_1$, $\phi_2$) > $\phi_L$.

The partitioning is stopped when any of the three criteria is unsatisfied.

### 6.1.3 Experimental results and discussions

To evaluate the performance of frequency-domain PAE approaches, a number of simulations are conducted. In these simulations, speech and music signals are selected as two sources in the primary components, which are amplitude panned and time-shifted separately to simulate different directions. To fulfill the assumptions of the stereo signal model, uncorrelated white Gaussian noise is used as the ambient component. Subsequently, the primary and ambient components are linearly mixed by letting PPR=0.9. DFT of size $N$=4096 (sampling frequency at 44.1 kHz), and Hanning window with 50% overlapping is applied. Both PCA and SPCA are employed in the testing, and their settings are listed as follows:

a) Full-band, without partitioning (denoted by F);

b-e) Fixed partitioning, with 2, 8, 32 uniform (U) partitions or 20 non-uniform (N) partitions based on ERB [FaB03], (denoted by 2U, 8U, 32U, and 20N, respectively);

f) TD adaptive partitioning, with $\phi_H$ =0.7, and $\phi_L$=0.05.

The performance of PAE is determined by the error-to-signal ratio (ESR) as in previous chapters, which can be computed as

$$\mathrm{ESR(dB)} = 10\log_{10}\left[0.5\left(\frac{\left\|\hat{\mathbf{p}}_0 - \mathbf{p}_0\right\|_2^2}{\left\|\mathbf{p}_0\right\|_2^2} + \frac{\left\|\hat{\mathbf{p}}_1 - \mathbf{p}_1\right\|_2^2}{\left\|\mathbf{p}_1\right\|_2^2}\right)\right]. \tag{6.5}$$

A better performance is achieved when ESR is smaller.

First, we test these PAE approaches with signals containing one source (a speech) in the primary components and the ESR results are presented in Table 6.1 (column “1S”). SPCA is better than PCA since it takes the time difference of the primary component into consideration. Comparing the results of SPCA in fixed partitioning with those in the full-band, we observed that the PAE performance degrades as the number of partitions increases. This observation indicates that the partitioning is not required and should be avoided for the single source case. Nevertheless, the performance of TD is quite close to the full-band approach.

Next, we test the performance of PAE when there are two sources in the primary components. Basically, three cases for the directions of two sources are specified as follows:

a) DS: in different sides, i.e., one in the left, the other in the right;

b) C: one in the center, the other in the left or right;

c) SS: in the same side, i.e., both are in the left or right.

Table 6.1 ESR of PAE for two sources

| Approach | Setting | 1S | SS | C | DS |
|---|---|---|---|---|---|
| PCA | F | -3.69 | -4.18 | -8.06 | -4.74 |
| | 2U | -3.38 | -3.95 | -8.19 | -5.04 |
| | 8U | -3.34 | -3.91 | -8.34 | -5.22 |
| | 32U | -3.16 | -3.89 | -8.44 | -5.48 |
| | 20N | -3.33 | -3.98 | -9.55 | -6.85 |
| | TD | -3.72 | -4.19 | -8.44 | -5.03 |
| SPCA | F | -14.78 | -10.16 | -8.07 | -6.45 |
| | 2U | -12.34 | -9.89 | -8.38 | -6.85 |
| | 8U | -11.52 | -9.8 | -8.57 | -7.11 |
| | 32U | -10.63 | -9.07 | -8.44 | -7.25 |
| | 20N | -10.34 | -7.29 | -9.07 | -7.73 |
| | TD | -14.13 | -10.41 | -8.58 | -7.93 |

The ESR results are shown in Table 6.1. First, we found that the performance of PCA is worse than that of SPCA, especially when no sources are in the center. Second, not all SPCA approaches with partitioning can yield a better performance than SPCA in full-band, especially when the directions of the two sources are closer (e.g., SS), as shown in Fig. 6.3. Generally, TD performs better than the fixed partitioning approaches, as well as the full-band approach. As the directions of the two sources get closer (i.e., from DS to SS), better performance with TD is usually achieved.

## 6.2 Multi-shift technique

In Chapter 5, we introduced a time-shifting techqnique to improve the performance of PAE when dealing with partically correlated primary components. The input signal is time shifted according to the estimated ICTD that corresponds to the direction of the dominant source. However, one single shift only accounts for one direction, which is ineffective for primary

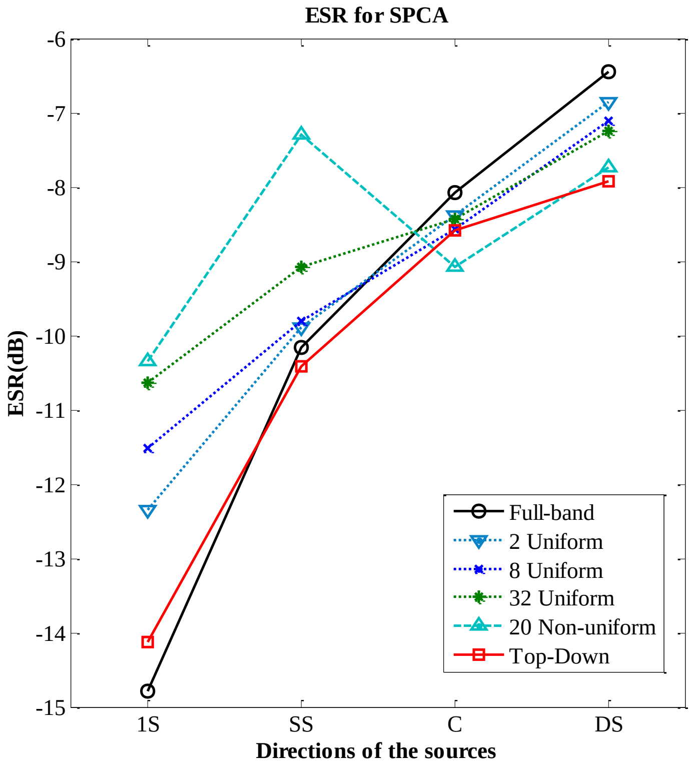


Figure 6.3 Comparison of ESR for SPCA with different partitioning settings

components that consist of sound sources from multiple directions. Thus, a common approach is to decompose the signal into subband before the extraction, assuming that only one source is dominant in each subband [Fal06], [HGT14]. Moreover, the directions of multiple sources can be tracked [RoW08] and localized [WoW12] in the presence of ambient noise. Nevertheless, subband PAE approaches become problematic when the spectra of the sources in the primary components overlap in certain subbands. Meanwhile, timbre change is an inevitable problem in subband PAE.

In this section, we investigate the primary component extraction (or primary extraction for short) with multiple directions by extending the single shift SPCA to multiple shifts. These shifts are performed based on the ICTD estimation. While in the output, the extracted primary components are correspondingly shifted back, weighted and summed to obtain the final results of the extracted primary components. We refer to this method as multi-shift PCA (MSPCA) in this section. The typical structure of MSPCA is shown in Fig. 6.4.

### 6.2.1 Multi-shift PCA

In many applications of spatial audio, concurrent sound sources from different directions and even the reflections of these sound sources (image sources) are frequently encountered in the stereo mix. These directions of the sources and reflections imply multiple different ICTDs. In such cases, SPCA with one single shift that corresponds to one single direction becomes problematic. Therefore, to account for multiple directions in the primary components of the stereo signal, we extend SPCA from one single shift to multiple shifts, and develop MSPCA for primary extraction. The typical structure of the MSPCA (MSPCA-T) is shown in Fig. 6.4. First, several ICTDs are estimated from the stereo input signal by finding the peaks in the short time cross correlation function [Mat13]. Next, the input signal is time shifted according to the estimated ICTDs [HTG13]. For every shifted version, PCA is applied to obtain the extracted primary components. Finally, the extracted primary components of all shifted versions are properly mapped, weighted and linearly summed to obtain the final output of the extracted primary components. Note that the weights are computed according to the significance of each shifted version.

Combining the selective time-shifting with the significance based weighting method, a consecutive structure for MSPCA can also be employed, as shown in Fig. 6.5. Instead of shifting the input signal according to a few selected ICTDs,

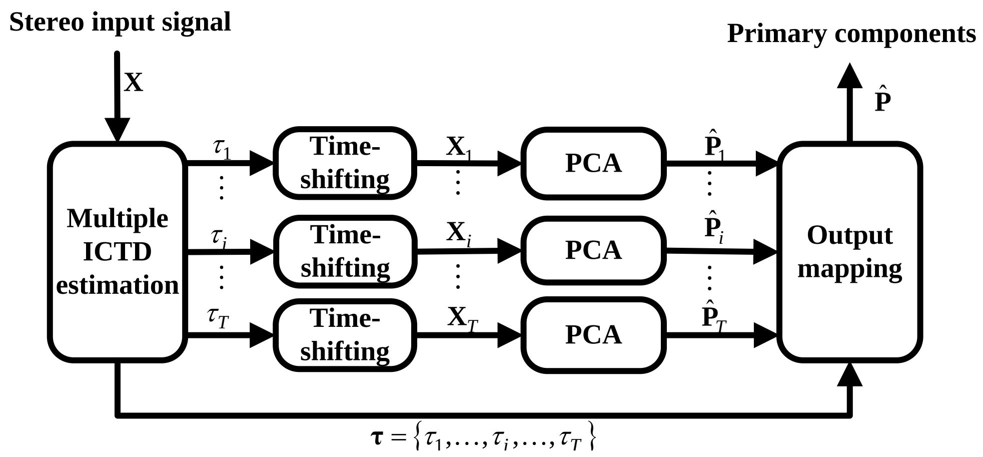


Figure 6.4 Typical structure of MSPCA (MSPCA-T). Stereo input signal $\mathbf{X}=\{\mathbf{x}_0,\mathbf{x}_1\}$; $\tau_i$ is the *i*th estimated ICTD (T is the total number of ICTDs); $\mathbf{X}_i$ and $\hat{\mathbf{P}}_i$ are the corresponding shifted signal and extracted primary component, respectively. The final output of the extracted primary components is denoted by $\hat{\mathbf{P}}$.

we perform the shifting consecutively lag by lag. Subsequently, PCA based primary extraction is employed for each shifted version. Before reversing the one-lag shifting and adding to the final output, the extracted primary components of each shifted version are weighted based on the significance of each shifted version. By assuming that those shifted versions having higher ICC are more significant, the weights are set higher for the shifted version with higher ICC. Via this ICC based weighting method, we can unify the consecutive MSPCA and MSPCA-T.

Let the stereo input signal be $\mathbf{X}=\{\mathbf{x}_0,\mathbf{x}_1\}$. The shifted signal is $\mathbf{X}_l=\{\mathbf{x}_0,\vec{\mathbf{x}}_1^l\}$ with *n*th sample of $\vec{\mathbf{x}}_1^l$ shifted by *l* lags, as $\vec{x}_1^l(n)=x_1(n-l)$, where $l\in[-L,L]$. The extracted primary components at the *l*th shifted version $\hat{\mathbf{P}}_l$ are computed using PCA. The final output of the extracted primary

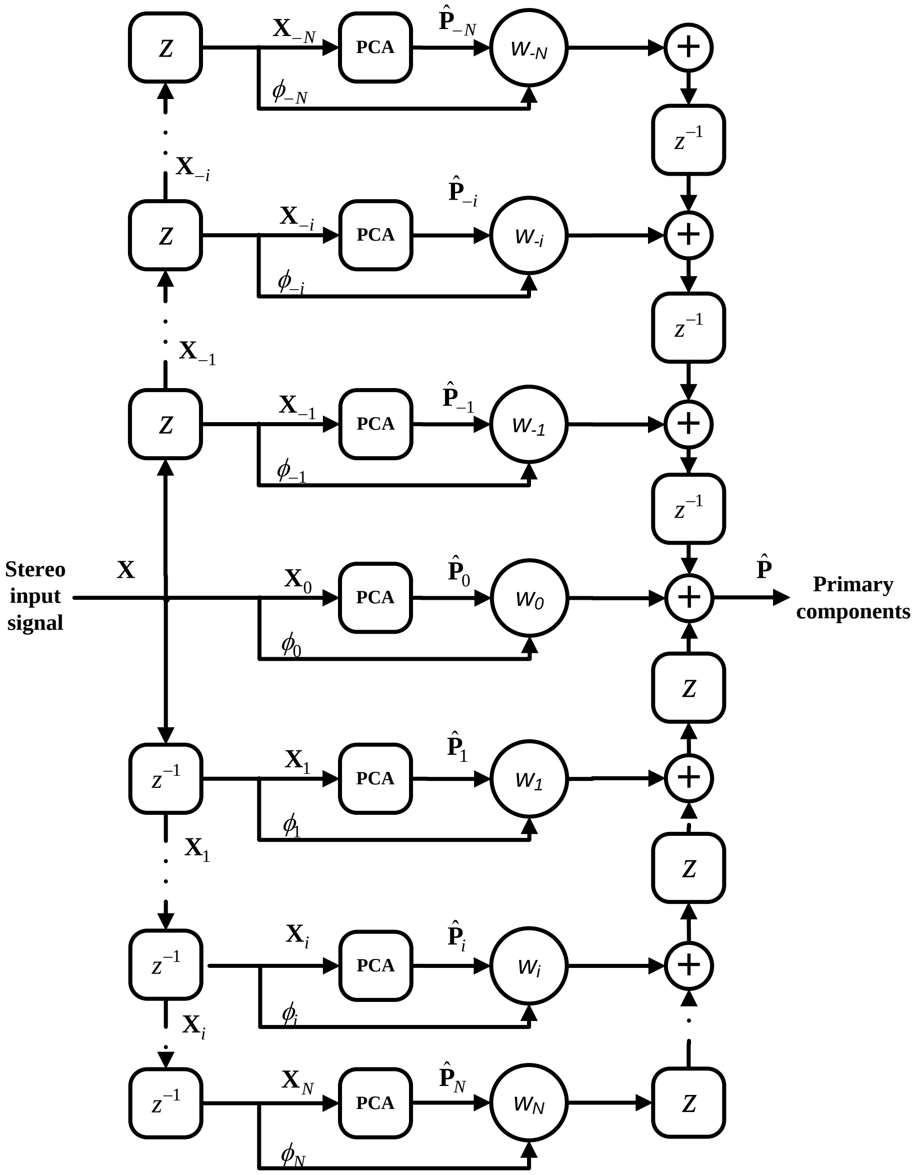


Figure 6.5 Block diagram of MSPCA with consecutive structure. The shift and reverse shift of a stereo signal is realized by delaying and advancing one channel of the stereo signal with the other channel kept unchanged.

components $\hat{\mathbf{P}}$ can be expressed as a weighted sum of the shifted back version of $\hat{\mathbf{P}}_l$. The *n*th sample of $\hat{\mathbf{P}}$ (either $\hat{\mathbf{p}}_0$ or $\hat{\mathbf{p}}_1$) is hence obtained by

$$\hat{P}(n) = \sum_{l=-L}^{L} w_l \hat{P}_l(n+l), \tag{6.6}$$

where $w_l \geq 0$ is the weight applied on $\hat{\mathbf{P}}_l$. To retain the overall signal power, the weights shall sum up to one, i.e., $\sum_{l=-L}^{L} w_l = 1$. Since the weights in consecutive MSPCA are proportional to the ICC of each lag, a straightforward way to obtain the weights is to employ the exponent of the ICC, i.e., $w_l = \phi_l^a \Big/ \sum_{l=-L}^{L} \phi_l^a$, where $a$ is the exponent and $\phi_l$ is the ICC of lag $l$. Larger values of $a$ lead to sparser weights. Examples of the exponent selection for the weighting methods are shown in the following section.

### 6.2.2 Experimental results and discussions

To evaluate the performance of the proposed MSPCA based primary extraction, a number of simulations and subjective listening tests are conducted. In our experiments, primary components consist of a speech signal and a music signal, which are amplitude panned by a factor of three and time shifted by 20 lags, towards the channel 1 and channel 0, respectively; and uncorrelated white Gaussian noise is used as the ambient component. Subsequently, the primary and ambient components are linearly mixed by setting the root-mean-square power of the speech, music and ambient component to be equal. This setting constraints the primary power ratio to 0.67. Next, PCA, SPCA and MSPCA with different settings are employed to extract primary components from the synthesized stereo signals. The searching range for ICTD is ±50 lags, which is around 2ms for sampling frequency at 44.1 kHz. Finally, the performance of primary extraction using these approaches is compared using objective metrics and subjective testing.

It can be found that PCA and SPCA can be considered as special cases of MSPCA by specifically setting the weights. Both PCA and SPCA have only one nonzero weights, but at different lags. While the corresponding lag for the unit weight in PCA is always zero, SPCA places the unit weight at the lag corresponding to maximum ICC. Since all weights shall sum up to one, this maximum weight for PCA and SPCA will be exactly equal to one. MSPCA-T can detect the two ICTDs by peak finding. After normalization, we can consider it having two nonzero weights at the two corresponding lags. For consecutive MSPCA, we examine two exponent values, namely, $a = 2$ and 10. Summarizing all different settings for these approaches, the weighting methods are compared in Fig. 6.6. As discussed, PCA and SPCA have only one nonzero weight at zero lag and -20 lag, respectively. For MSPCA-T, two weights are applied at two distinct lag positions, though the positive ICTD for the music is not as accurate as the negative ICTD for the speech. For consecutive MSPCA with different exponent values, the non-zero weights are found for all the lags, and apparently higher weights are given to those lags that are closer to the directions of the primary components. As the exponent value $a$ increases, the differences among the weights at various lags become more significant. When $a$ is high (e.g., $a$=10), the weighting method in consecutive MSPCA becomes similar to SPCA, as seen from Fig. 6.6(b) and Fig. 6.6(e).

After applying these approaches, the objective performance on the extraction accuracy of the primary component is determined by ESR, as defined in (6.5). The ESR results for these approaches are illustrated in Fig. 6.7. It is obvious that MSPCAs generally perform better than PCA or SPCA by having smaller ESR. It is also quite interesting to observe that consecutive MSPCA approaches

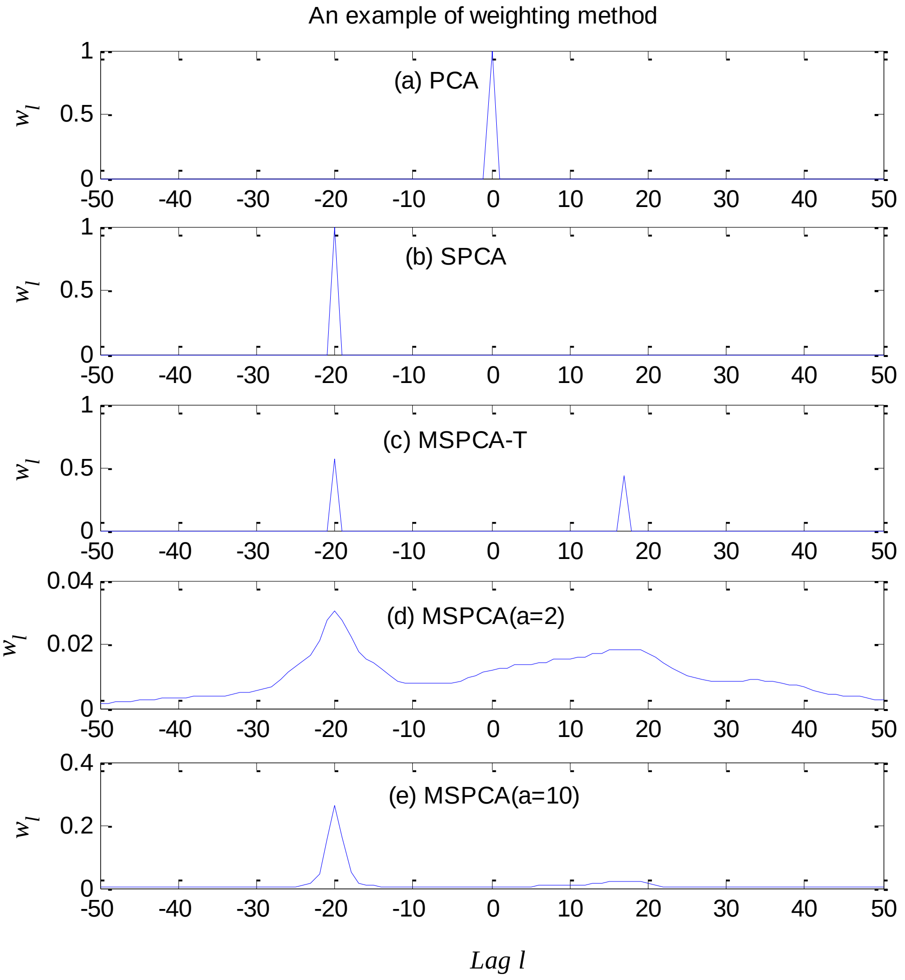


Figure 6.6 An illustration of the weighting methods in PCA, SPCA and MSPCAs. Negative and positive lags correspond to the direction towards the channel 1 and channel 0, respectively.

outperform MSPCA-T. This implies that the accuracy in the estimation of the number of the directions and the associated ICTDs are extremely critical for MSPCA-T. Failure to accurately estimate any ICTDs will degrade the overall extraction performance, as observed here. By contrast, consecutive MSPCA mitigates this problem by applying weights at all lags. Furthermore, the averaging of the ambient components across various shifted versions could also reduce ambient leakage in the extracted primary components. Between the two consecutive MSPCA approaches, MSPCA($a$=2) performs better than

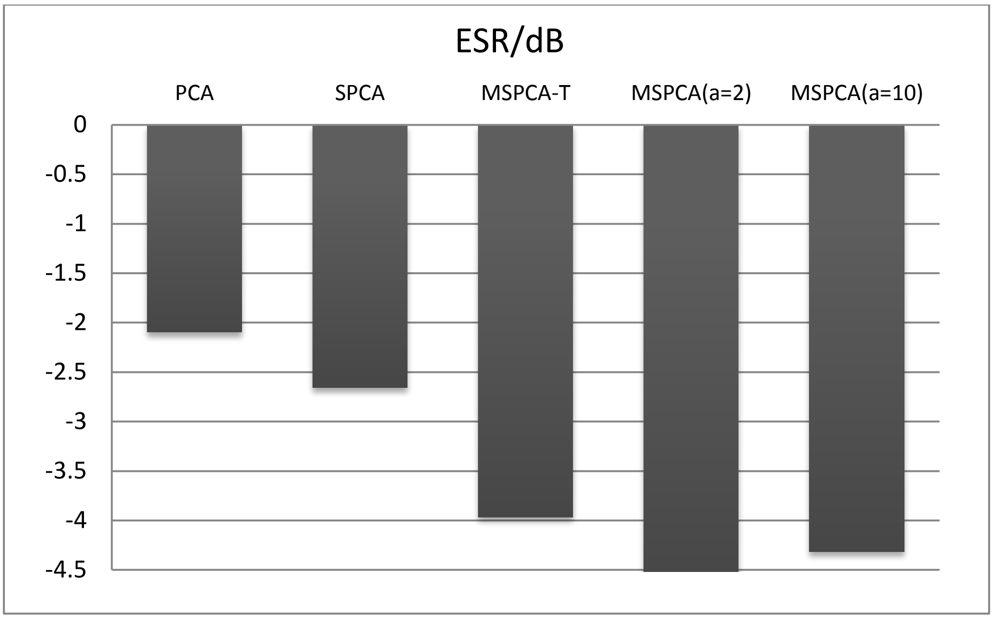


Figure 6.7 Objective performance on extraction accuracy measured by ESR for PCA, SPCA, MSPCAs.

MSPCA($a$=10). Therefore, the exponent applied on the ICC for the weights in consecutive MSPCA cannot be too large.

In addition to the objective assessment on the error performance, subjective testing of localization accuracy of the primary extraction was also conducted. The testing method was based on MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [ITU03b], [Vin13]. Nine signals, including primary components extracted using the five methods, one known reference, one hidden reference and two anchors, were tested. The subjects were asked to rate a score of 0-10, where a score of 0 denotes the worst localization (i.e., the two directions are reversed), and a score of 10 denotes the same directions perceived as the reference. When at least one direction is accurate, a score of no less than 5 shall be given, and a score of 3-7 shall be appropriate for those signals with perceived directions neither too close nor too bad. Finally, 12 subjects participated in the experiment and the results are shown in Fig. 6.8. Generally, MSPCAs produce

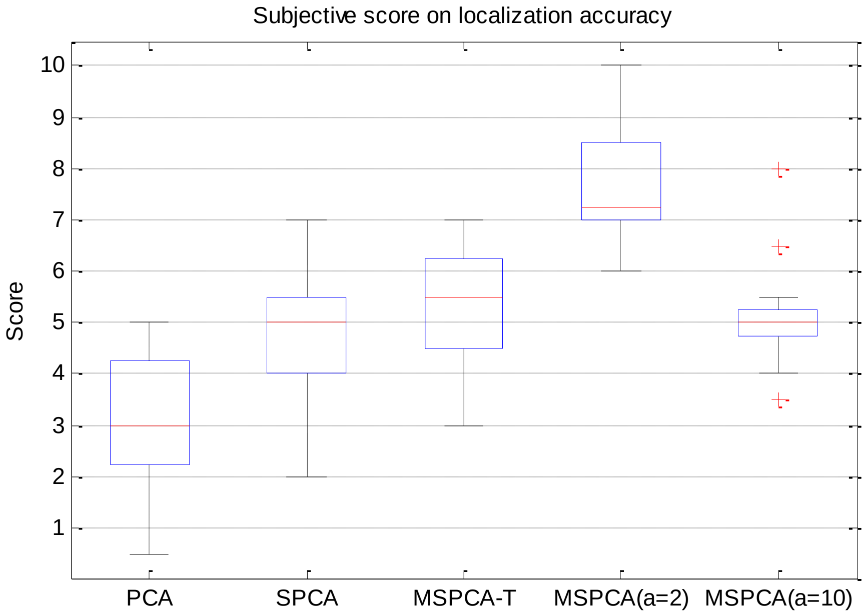


Figure 6.8 Subjective performance on localization accuracy for PCA, SPCA, MSPCAs.

more accurate localization of the primary components among these testing methods. Similar to the observation in ESR, MSPCA(*a*=2) performs the best and MSPCA(*a*=10) degrades the localization significantly. Therefore, it can be concluded that consecutive MSPCA with proper weighting can help improve both the extraction accuracy and localization accuracy of the primary components when there are multiple directions.

## 6.3 Further discussions and conclusions

In this chapter, two techniques to improve the performance of PAE in the case of multiple concurrent sources are discussed. PCA and SPCA based PAE approaches are employed in this study.

The subband technique was derived in frequency-domain. We found that frequency bin partitioning is unnecessary for one source, but this partitioning plays an essential role for multiple sources, especially when the spectra of the sources overlap. Conventional fixed partitioning and proposed top-down adaptive partitioning methods were compared for both PCA and SPCA in our simulations. Generally, SPCA outperforms PCA regardless of the partitioning methods. As for the influence of different partitioning methods in SPCA, we found that not all partitioning methods yield better performance than the full-band approach, whereas the best performance is obtained with the proposed ICC-based TD partitioning method.

On the other hand, multi-shift technique takes a time-domain approach that extends the single time-shifting into multiple shifts to account for the multiple directions. Two different structures of MSPCA are examined. While MSPCA with typical structure is simpler, its performance relies heavily on the correct estimation of the ICTDs. By contrast, consecutive MSPCA is more robust by applying weights on all shifted versions. The weighting method for different shifted versions is found to be critical to the extraction performance. In general, applying the exponential function of ICC with proper exponent value as the weightings yields a good performance in terms of the extraction accuracy as well as localization accuracy.

Comparing the subband and multi-shift techniques in PAE, we found that both approaches can be considered as preprocessing before applying the conventional PAE approaches. In both techniques, the input singal is processed signals to match the assumptions of the signal model as close as possible. ICC is critical in both approaches, in either the partitioning of frequency bins or output weighting of the shifted versions, to obtain the extracted primary components. Both approaches can be considered as a filtering process. The subband technique can be viewed as frequency-domain filtering where the spectra of the mixed signal is multiplied by a complex weight in each frequency bin. The multi-shift technique could be used as a filter with coefficients derived using ICC. Under the filtering framework, the subband technique can be combined directly with the multi-shift technique in PAE for practical spatial audio reproductions. In the next chapter, we will discuss how PAE is applied in the headphone based spatial audio reproduction system to achieve a natural listening experience.

# Chapter 7

# Natural Sound Rendering for Headphones

In previous chapters, various PAE approaches are studied in ideal case, primary-complex case, and multiple source case. Yet, it is not clearly addressed how PAE is seamlessly incorporated in the spatial audio reproduction system. With the strong growth of assistive and personal listening devices, natural sound rendering over headphones is becoming a necessity for prolonged listening in multimedia and virtual reality applications. Thus, in this chapter, we focus on the spatial audio reproduction for headphone playback. The aim of natural sound rendering is to recreate the sound scenes with the spatial and timbral quality as natural as possible, so as to achieve a truly immersive listening experience. However, rendering natural sound over headphones encounters many challenges. Therefore, PAE based signal processing techniques are presented in this chapter[1] to tackle these challenges and assist human listening.

## 7.1 Introduction

Sound plays an important role in our daily lives for communication,

[1] The work reported in this chapter is an extension from the author's Journal paper [SHT15] published in *IEEE Signal Processing Magazine*, March 2015 issue, and conference paper [HGT15d] presented in ICASSP 2015.

information, and entertainment. In most of these applications, listening is seldom from the physical sound sources but instead from playback devices, such as headphones or loudspeakers. Headphones, by virtue of their convenience and portability, are typically chosen as the preferred playback device, especially for personal listening. Therefore, to assist headphone listening, it is critical for the sound to be rendered in a way that listeners can perceive it as natural as possible. In this context, natural sound rendering essentially refers to rendering of the original sound scene using headphones to create an immersive listening experience and the sensation of "being there" at the venue of the acoustic event. To achieve natural sound rendering, the virtual sound rendered should exactly emulate all the spatial cues of the original sound scene, as well as the individual spectral characteristics of the listener's ears. In this chapter, we mainly consider the most widely used channel-based audio as the input signals for the natural sound rendering system, though some of the signal processing techniques discussed could also be used in other audio formats, such as object-based format and ambisonics [SWR13], [Pul07].

In recent years, the design criteria for commercial headphones have undergone significant development. At Harman, Olive *et al.* investigated the best target responses for designing headphones based on the listener's preference for the most natural sound [OWM13]. Creating realistic surround sound in headphones has become a common pursuit of many headphone technologies from Dolby, DTS, etc. Furthermore, personalized listening experience and incorporating the information of listening environment has also been the trends in headphone industry. These trends in headphones share one common objective: ***To render natural sound in headphones.***

## 7.2 Challenges and signal processing techniques

The listening process of humans can generally be considered as a source-medium-receiver model, as stated by Begault [Beg00]. This model is used in this chapter to highlight the differences between natural listening in real environment and listening over headphones. In natural listening, we listen to the physical sound sources in a particular acoustic space, with the sound waves undergoing diffraction, interference with different parts of our morphology (torso, head and pinna) before reaching the eardrum. This information of sound wave propagation can be encapsulated in spatial digital filters termed as head-related transfer functions (HRTFs) [Beg00]. Listeners also get valuable interaural cues for sound localization with head movements. However, headphone listening is inherently different from natural listening as the sources we are listening to are no longer physical sound sources but are recorded and edited sound materials. These differences between natural and headphone listening lead to various challenges in rendering natural sound over headphones, which can be broadly classified into the following three categories:

1) **From the perspective of source,** the sound scenes rendered for headphone listening should comprise not only the individual sound sources but also the features of the sound environment. Listeners usually perceive these sound sources to be directional, i.e., coming from certain directions. Moreover, in most of the digital media content, the sound environment is usually perceived by the listener to be diffuse (partially). This perceptual difference between the sound sources and the sound environment requires them to be considered separately in natural sound rendering [SWR13]. Though there are other formats that can represent the sound scenes (e.g., object-based, ambisonics), the

convention for today’s digital media is still primarily channel-based format. Hence, the focus of this chapter lies in the rendering of channel-based audio, where sound source and environment signals are mixed in each channel [SWR13]. In channel-based signals, where only the sound mixtures are available (assuming one mixture in every channel), it is necessary to extract the source signals and environment signals, which can be quite challenging. Furthermore, most of the traditional recordings are processed, and mixed for optimal playback over loudspeakers, rather than headphones. Direct playback of such recordings over headphones results in an unnatural listening experience, which is mainly due to the loss of crosstalk and localization issues.

2) **From the perspective of medium**, headphone listening does not satisfy free-air listening conditions as in natural listening. Since the headphone transfer function (HPTF) is not flat, equalization of the headphone is necessary. However, this equalization is tedious and challenging as the headphone response is highly dependent on the individual anthropometrical features and also varies with repositioning.

3) **From the perspective of receiver**, the omission of listener’s individualized filtering with the outer ear in headphone listening often leads to coloration and localization inaccuracies. These individualized characteristics of the listener are lost when the sound content is recorded or synthesized non-individually, i.e., the subject in the listening is different from the subject in the recording or synthesis. Furthermore, the sound in headphone listening is not adapted to the listener’s head movements, which departs from a natural listening experience.

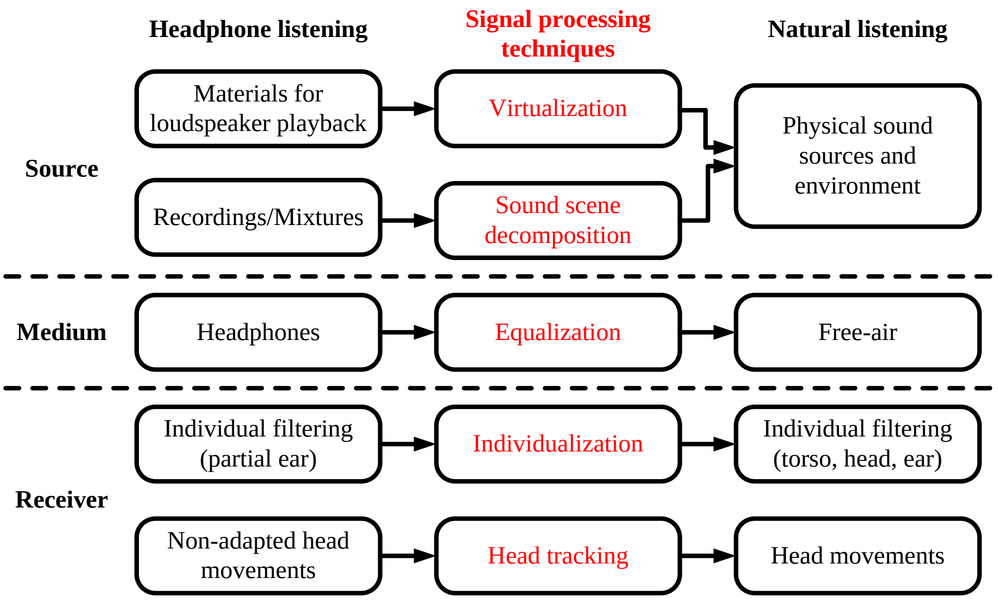

Headphone listening
Signal processing techniques
Natural listening
Source
Materials for loudspeaker playback
Virtualization
Physical sound sources and environment
Recordings/Mixtures
Sound scene decomposition
Medium
Headphones
Equalization
Free-air
Receiver
Individual filtering (partial ear)
Individualization
Individual filtering (torso, head, ear)
Non-adapted head movements
Head tracking
Head movements

Figure 7.1 A summary of the differences between natural listening and headphone listening and the corresponding signal processing techniques to solve these challenges for natural sound rendering. The main challenges and their corresponding signal processing techniques in each category (source, medium, and receiver) are highlighted and their interactions (not shown here) are further discussed in the chapter.

To tackle the above challenges and enhance natural sound rendering over headphones, digital signal processing techniques are commonly used. In Fig. 7.1, we summarize the differences between natural listening and headphone listening, and introduce the corresponding signal processing techniques to tackle these challenges, which are:

1) Virtualization: to match the desired playback for the digital media content;
2) Sound scene decomposition using blind source separation (BSS) and primary ambient extraction (PAE): to optimally facilitate the separate rendering of sound sources and sound environment;

3) Individualization of HRTF: to compensate for the lost or altered individual filtering of the sound in headphone listening;
4) Equalization: to preserve the original timbral quality of the source and alleviate the adverse effect of the inherent headphone response;
5) Head tracking: to adapt to the dynamic head movements of the listener.

The remainder of this Chapter is structured as follows. Virtualization and head tracking, due to their high interactions, are explained together in Section 7.3, followed by the decomposition of sound scenes in Section 7.4. Sections 7.5, and 7.6 describe individualization and equalization, respectively. These signal processing techniques are integrated and evaluated using subjective tests in Sections 7.7 and 7.8, respectively. Finally, the conclusions and future trends are presented in Section 7.9.

## 7.3 Virtualization

In digital media, sound is typically mixed for loudspeaker playback rather than headphone playback. The spatial sound to be rendered naturally over headphones should emulate the natural propagation of the acoustic waves emanating from the loudspeaker to the eardrum of the listener. To emulate stereo or surround sound loudspeaker rendering over headphones, virtualization techniques based on HRTFs corresponding to the loudspeaker positions are commonly used. Given these acoustic transfer functions (i.e., HRTFs), the virtualization technique is applicable to any multichannel loudspeaker setup, be it stereo, 5.1, 7.1, 22.2, or even loudspeaker arrays in wave-field synthesis. As shown in Fig. 7.2, for every desired loudspeaker position, the signal in the *m*th

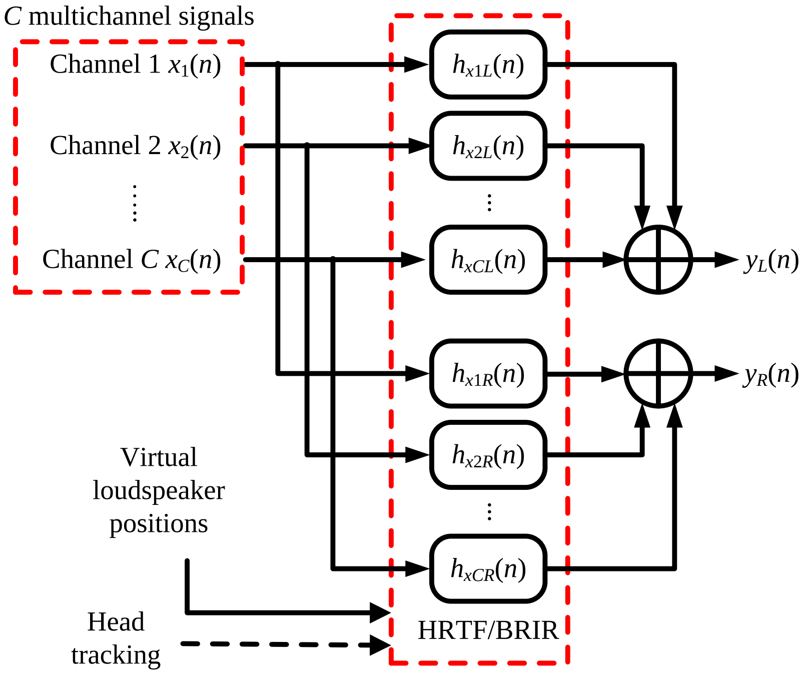


Figure 7.2 Virtualization of (a) multichannel loudspeaker signals $x_c(n)$ [GoJ07a]. Note that head tracking can be used to update the selected directions of HRTFs/BRIRs.

channel $x_c(n)$ is filtered with the corresponding HRTF $h_{xcL}(n)$, $h_{xcR}(n)$, and summed before being routed to the left and right ears [Beg00], [GoJ07a], respectively, as:

$$
\begin{aligned}
y_L(n) &= \sum_{c=1}^{C} h_{xcL}(n) * x_c(n), \\
y_R(n) &= \sum_{c=1}^{C} h_{xcR}(n) * x_c(n),
\end{aligned}
\qquad (7.1)
$$

where * denotes convolution and *M* is the total number of channels. When the HRTFs are directly applied to multichannel loudspeaker signals, the rendered sound scenes in headphone playback suffer from inaccurate virtual source directions, lack of depth, and reduced image width [GoJ07a], [BrS08].

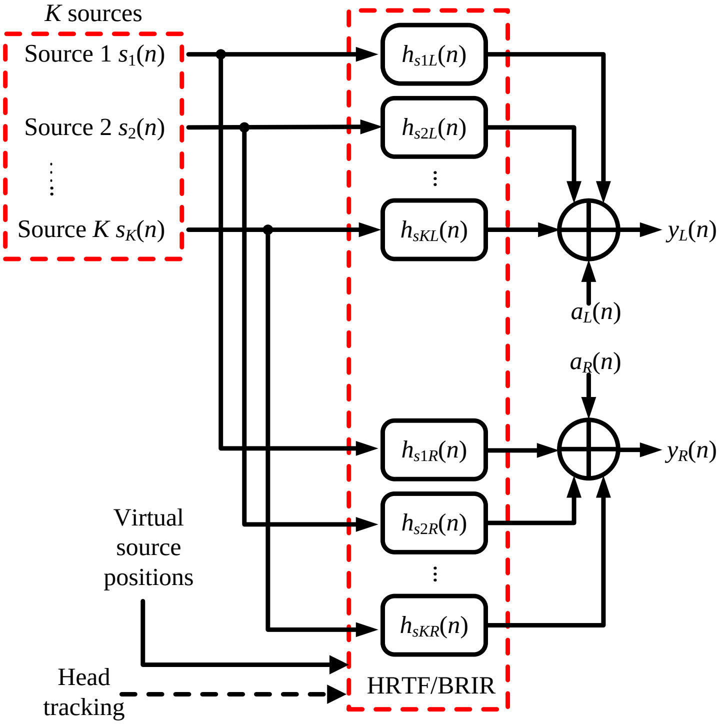


Figure 7.3 Virtualization of multiple sources $s_k(n)$ and environment signals $a_L(n), a_R(n)$. Signals $y_L(n), y_R(n)$ are sent to the left and right ear, respectively. Note that head tracking can be used to update the selected directions of HRTFs/BRIRs.

To solve these problems in virtualization of multichannel loudspeaker signals and achieve a faithful reproduction of the sound scenes, the HRTFs should be applied to the individual source signals that are usually extracted (using BSS, PAE) from the loudspeaker signals (i.e., mixtures). In this virtualization as shown in Fig. 7.3, the sources are rendered directly using the HRTFs of the corresponding source directions $h_{skL}(n)$, $h_{skR}(n)$:

$$\begin{aligned} y_L(n) &= \sum_{k=1}^{K} h_{skL}(n) * s_k(n) + a_L(n), \\ y_R(n) &= \sum_{k=1}^{K} h_{skR}(n) * s_k(n) + a_R(n), \end{aligned} \tag{7.2}$$

where $K$ is the total number of sources, $s_k(n)$ is the $k$th source in the multichannel signal, and the environment signals $a_L(n)$, $a_R(n)$ are the rendered signals representing the sound environment perceived at two ears. To render the acoustics of the environment, the environment signals can be either synthesized according to the sound environment [AvJ04] or extracted from the mixtures. Techniques like decorrelation [GoJ07a], [Fal06] and artificial reverberation [MeF10] are commonly employed to render the environment signals in order to create a more diffuse and natural sound environment.

Furthermore, adding the reverberation of sources (or the loudspeaker signals in virtualization of multichannel loudspeaker signals) can also improve the realism of the reproduced sound scene [FaB03]. Therefore, in virtualization, it is quite common to use binaural room impulse response (BRIR) [Beg00], [GoJ07a] that encapsulates HRTFs and reverberation. In this case, selecting the correct amount of early reflections as well as late reverberation is critical to recreate a faithful sound environment [Beg00]. In general, the BRIR that matches the sound environment of the scene or BRIR of a mixing studio is considered to be more suitable [OWM13]. As discussed in Section 7.2, natural sound rendering requires the accurate reproduction of both the sound sources and the sound environment. Compared to the virtualization of multichannel loudspeaker signals (Fig. 7.2), the latter technique of virtualizing the source and environment signals (Fig. 7.3) is more desirable as it is closer to natural listening [BrS08], [Fal06], [MeF10]. These virtualization techniques can also be incorporated into spatial audio coding systems, such as binaural cue coding [FaB03], spatial audio scene coding [GoJ07a], and directional audio coding [Pul07].

In virtualization, the directions of the sources (or the loudspeakers in virtualization of multichannel loudspeaker signals as in Fig. 7.2) have to be calibrated according to the head movements (as in natural listening). To fulfill this need, the HRTFs/BRIRs in the virtualization are updated on the fly based on these head movements that are often tracked by a sensor (e.g., accelerometer, gyroscope, camera, etc.). The latency between the head tracking and sound rendering should be such that the localization accuracy is not affected [AlD11]. When incorporated in the virtualization process, such a head tracking system can provide useful dynamic cues to resolve the localization conflicts [Beg00] and enhance natural sound rendering [BWA01], [AlD11]. It shall be noted that head tracking is more critical for the directional sources but less important for the diffuse signals like environment signals and late reverberation [AlD11]. This is because the perception of diffuse signals is less affected by head movements.

## 7.4 Sound scene decomposition

To achieve natural sound rendering in headphones, two important constituents of the sound scenes are required in the virtualization, namely, the individual sound sources and characteristics of the sound environment. However, this information is usually not directly available to the end user. One has to work with the existing digital media content that is available, i.e., the mastered mix distributed in channel-based formats (e.g., stereo, 5.1). Therefore, to facilitate natural sound rendering, it is necessary to extract the sound sources and/or sound environment from their mixtures. In this section, we discuss two types of techniques applied in sound scene decomposition, namely, BSS and PAE.

### 7.4.1 Decomposition using BSS

Extracting the sound sources from the mixtures, often referred to as BSS, has been extensively studied in the last few decades. In BSS, the sound scene is considered to be the sum of distributed sound sources. The basic mixing model in BSS can be considered as anechoic mixing, where the sources $s_k(n)$ in each mixture $x_c(n)$ have different gains $g_{ck}$ and delays $\tau_{ck}$. Hence, the anechoic mixing is formulated as follows:

$$x_c(n) = \sum_{k=1}^{K} g_{ck} s_k(n - \tau_{ck}) + e_c(n), \quad \forall c \in \{1, 2, \ldots, C\}, \tag{7.3}$$

where $e_c(n)$ is the noise in each mixture, which is usually neglected for most cases. Note that estimating the number of sources is quite challenging and it is usually assumed to be known in advance [HKO04]. This formulation can be simplified to represent instantaneous mixing by ignoring the delays, or can be extended to reverberant mixing by including multiple paths between each source and mixture. An overview of the typical techniques applied in BSS is listed in Table 7.1.

Based on the statistical independence and non-Gaussianity of the sources, independent component analysis (ICA) algorithms have been the most widely used techniques in BSS to separate the sources from mixtures in the determined case, where the numbers of mixtures and sources are equal [HKO04]. In the over-determined case, where there are more mixtures than sources, ICA is combined with principal component analysis (PCA) to reduce the dimension of

Table 7.1 Overview of typical techniques in BSS

| **Objective**: To extract $K$ ($K > 2$) sources from $C$ mixtures | | |
|---|---|---|
| **Case** | | **Typical techniques** |
| Determined: $K = C$ | | ICA [HKO04] |
| Over-determined: $K < C$ | | ICA with PCA or LS [HKO04] |
| Under-determined: $K > C$ | $C > 2$ | ICA with sparse solutions [HKO04], [PBD10] |
| | $C = 2$ | Time-frequency masking [YiR04] |
| | $C = 1$ | NMF [Vir06], [VBG14]; CASA [WaB06] |

the mixtures, or combined with least-squares (LS) to minimize the overall mean-square-error (MSE) [HKO04]. In practice, the under-determined case is the most common, where there are fewer mixtures than sources. For the under-determined BSS, sparse representations of the sources are usually employed to increase the likelihood of sources to be disjoint [PBD10]. The most challenging under-determined BSS is when the number of mixtures is two or lesser, i.e., in stereo and mono signals.

Stereo signals (i.e., $C = 2$), being one of the most widely used audio format, have been the focus in BSS. Many of these BSS techniques can be considered as time-frequency masking and usually assume one dominant source in one time-frequency bin of the stereo signal [YiR04]. In these time-frequency masking based approaches, a histogram for all possible directions of the sources is constructed, based on the range of the bin-wise amplitude and phase differences between the two channels. The directions, which appear as peaks in the histogram, are selected as source directions. These selected source directions are then used to classify the time-frequency bins, and to construct the mask. For every time-frequency bin $(m,l)$, the $k$th source at $c$th channel $\hat{S}_{ck}(n,l)$ is estimated as:

$$\hat{S}_{ck}(m,l) = \Psi_{ck}(m,l)\,X_c(m,l), \tag{7.4}$$

where the mask and the *m*th mixture are represented by $\Psi_{ck}(m,l)$ and $X_c(m,l)$, respectively.

In the case of single-channel (or mono) signals, the separation is even more challenging since there is no inter-channel information. Hence, there is a need to look into the inherent physical or perceptual properties of the sound sources. Non-negative matrix factorization (NMF) based approaches are extensively studied and applied in single-channel BSS in recent years. The key idea of NMF is to formulate an atom-based representation of the sound scene [Vir06], where the atoms have repetitive and non-destructive spectral structures. NMF usually expresses the magnitude (or power) spectrogram of the mixture as a product of the atoms and time varying non-negative weights in an unsupervised manner. These atoms, after being multiplied with their corresponding weights, can be considered as potential components of sources [VBG14]. Another technique applied in single-channel BSS is the computational auditory scene analysis (CASA) that simulates the segregation and grouping mechanism of human auditory system [WaB06] on the model-based representation (monaural case) of the auditory scenes. An important aspect worth considering is the directions of the extracted sources, which can usually come as a by-product in multichannel BSS. In single-channel BSS, this information of source directions has to be provided separately.

### 7.4.2 Decomposition using PAE

In most sound scenes, the mixture comprises not only the dry sources but also the reverberation and ambient sound, which are contributed by the acoustics of the surrounding space. Therefore, the mixing model of the sources in BSS usually does not match with the actual sound scenes. In this chapter, we refer to the dominant sources as primary (or direct) components, and the signals contributed by the sound environment as ambient (or diffuse) components. The primary and ambient components are perceived to be directional and diffuse, respectively. Different rendering methods should be applied to the primary and ambient components [BrS08], [AvJ04] due to their perceptual differences. Therefore, rendering of natural sound scenes requires the decomposition of the mixtures into primary and ambient components [BrS08], [AvJ04], [MeF10]. Detailed discussions on PAE can be found in previous chapters of this thesis.

### 7.4.3 A comparison between BSS and PAE

Both BSS and PAE are extensively applied in sound scene decomposition, and a comparison between these approaches is summarized in Table 7.2. The common objective of BSS and PAE is to extract useful information (mainly the sound sources and their directions) about the original sound scene from the mixtures, and to use this information to facilitate natural sound rendering. Following this objective, there are three common characteristics in BSS and PAE. First, only the mixtures are available and usually no other prior information is given. Second, the extraction of the specific components from the mixtures is based on certain signal models. Third, both techniques require objective and subjective evaluation.

As discussed earlier, the applications of different signal models in BSS and PAE lead to different techniques. In BSS, the mixtures are considered as the sums of multiple sources, and the independence among the sources is one of the most important characteristics. In contrast, the mixing model in PAE is based

Table 7.2 Comparison between BSS and PAE in sound scene decomposition

| | BSS | PAE |
|---|---|---|
| **Objective** | To obtain useful information about the original sound scene from given mixtures, and facilitate natural sound rendering. | |
| **Common characteristics** | • Usually no prior information, only mixtures;<br>• Based on certain signal models;<br>• Require objective as well as subjective evaluation. | |
| **Basic mixing model** | Sums of multiple sources (independent, non-Gaussian, etc.) | Primary components (highly correlated)+ Ambient components (uncorrelated) |
| **Techniques** | ICA [HKO04], sparse solutions [PBD10], time-frequency masking [YiR04], NMF [Vir06], [VBG14], CASA [WaB06], etc. | PCA [MGJ07], LS [Fal06],[HTG14], time-frequency masking [AvJ04],[MGJ07], time/phase-shifting [HTG13], [HGT14], etc. |
| **Typical applications** | Speech, music | Movie, gaming |
| **Related applications** | Speech enhancement, noise reduction, speech recognition, music classification | Sound reproduction,<br>sound localization,<br>coding |
| **Limitations** | • Small number of sources<br>• Sparseness/disjoint<br>• No/simple environment | • Small number of sources<br>• Sparseness/disjoint<br>• Low ambient power<br>• Primary ambient components uncorrelated |

on human perception of directional sources (primary components) and diffuse sound environment (ambient components). The perceptual difference between primary and ambient components is due to the directivity of these components that can be characterized by their correlations. The applications that adopted BSS and PAE also have distinct differences. BSS is commonly used in speech and music applications, where the clarity of the sources is usually more important than the effect of the environment. On the other hand, PAE is more suited for the reproduction of movie and gaming sound content, where the ambient components also contribute significantly to the naturalness and immersiveness of the sound scenes. Subjective experiments revealed that BSS and PAE based headphone rendering can improve the externalization and enlarge the sound stage with minimal coloration [BrS08]. It shall be noted in certain cases, such as extracting sources from their reverberation, BSS shares a similar objective as PAE and hence can be applied in PAE [SRK12].

Despite the recent advances in BSS and PAE, the challenges due to the complexity and uncertainty of the sound scenes still remain to be resolved. One common challenge in both BSS and PAE is the increasing number of audio sources in the sound scenes, while only a limited number of mixtures (i.e., channels) are available. In certain time-frequency representations, the sparse solutions in BSS and PAE would require the sources to be sparse and disjoint [PBD10]. Considering the diversity of audio signals, finding a robust sparse representation for different types of audio signals is extremely difficult. The recorded or post-processed source signals might even be filtered due to physical or equivalently simulated propagation and reflections. Moreover, the audio signals coming from adverse environmental conditions (including reverberation, and strong ambient sound) usually degrade the performance of the decomposition. These difficulties can be addressed by studying the features of the resulting signals and by obtaining more prior information on the sources, the sound environment, the mixing process [VBG14], and combining auditory information with visual information of the scene.

## 7.5 Individualization of HRTF

Binaural technology is the most promising solution for delivering spatial audio in headphones, as it is the closest to natural listening. Unlike conventional microphone recordings, which are meant for loudspeaker playback, the binaural signals are recorded or synthesized at the ears of the listener. In a binaural audio

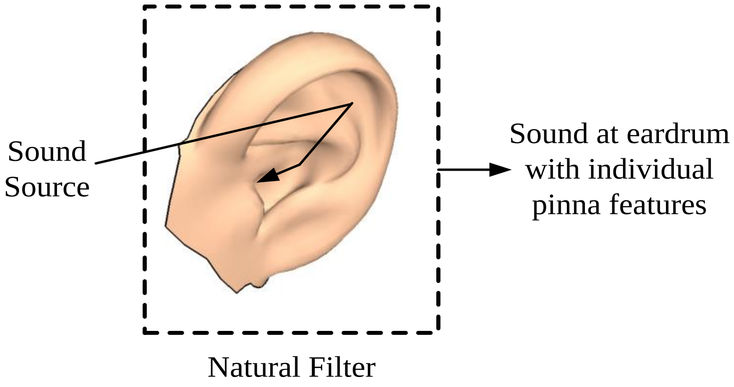


Figure 7.4 Human ears act as a natural filter in physical listening.

system, the spatial encoding (i.e., HRTFs) should encapsulate all the spectral features due to the interaction of the acoustic wave with the listener's morphology (torso, head, and pinna). The pinna, which is also considered as the acoustic fingerprint, embeds the most idiosyncratic spectral features into HRTFs, which are essential for accurate perception of the sound (Fig. 7.4). Thus, the HRTF features of the individuals are extremely unique, as shown in Fig. 7.5. Often the HRTFs used for virtualization are non-individualized HRTFs, typically measured on a dummy head, since they are easily accessible.

However, the use of non-individualized HRTFs leads to several artifacts such as in-head localization, elevation localization confusions, front-back, up-down reversals and inexact location of the auditory image [WAK93]. Thus, individualization of the HRTFs plays a critical role to create an immersive experience closest to the natural listening experience. There are various individualization techniques to obtain the individualized HRTFs from acoustical measurements, anthropometric features of the listener, customizing generic HRTFs with perceptual feedback or frontal projection of sound, as summarized in Table 7.3.

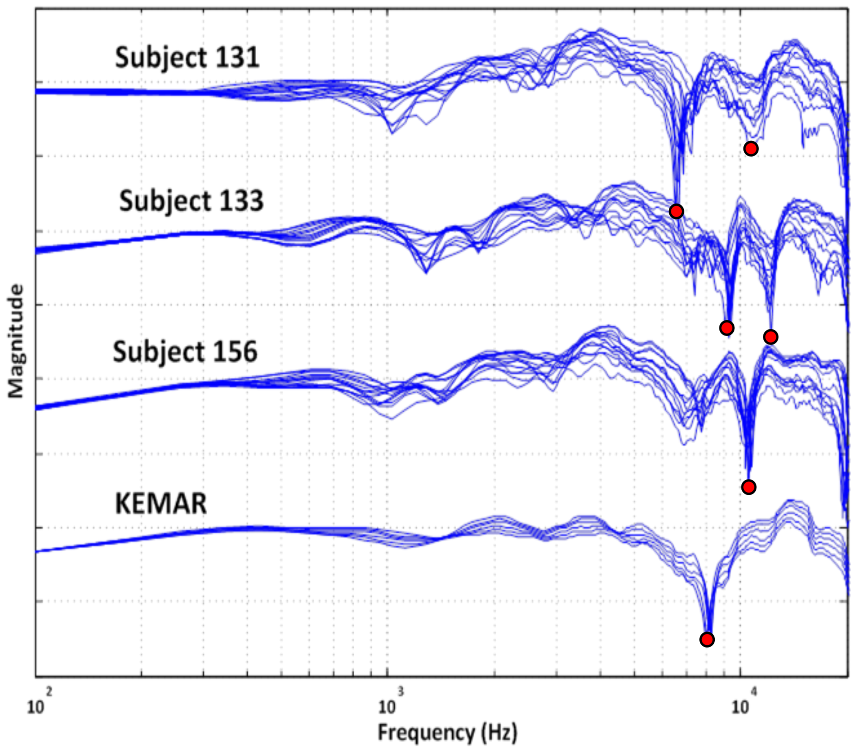


Figure 7.5 The vast variation of the HRTF spectrum at high frequencies of the various subjects from CIPIC database and the MIT KEMAR dummy head database [XLS07]. This is due to the idiosyncratic nature of the pinna.

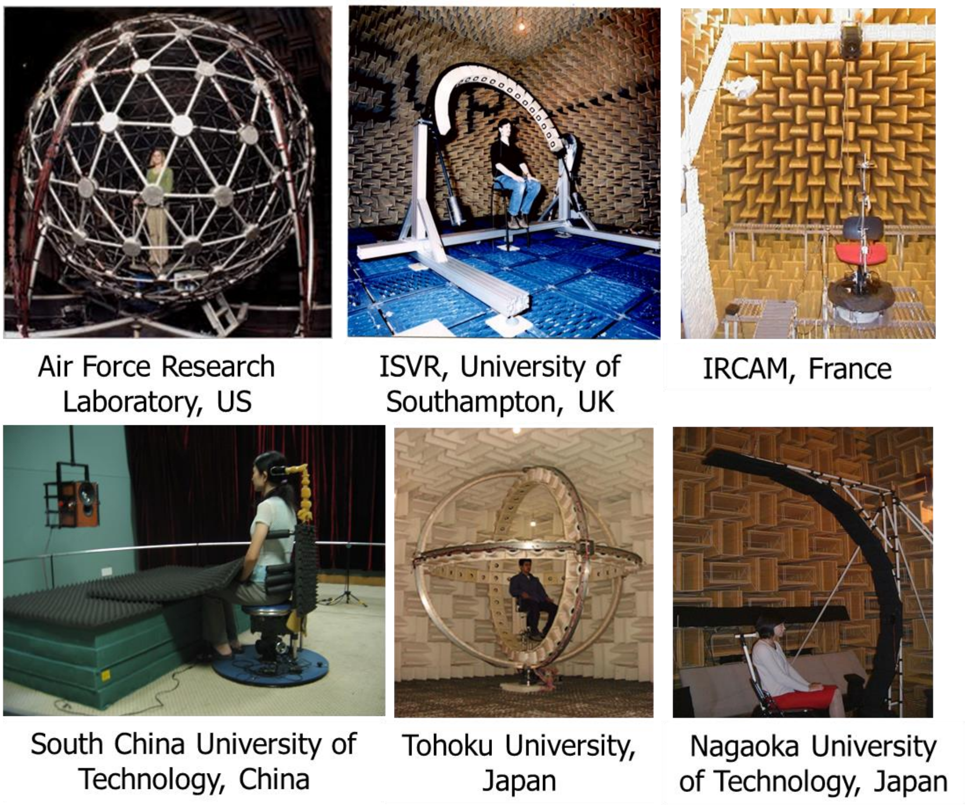
### 7.5.1 Acoustical measurements

The most straightforward individualization technique is to actually measure the individualized HRTFs for every listener at different sound positions [MSH95], [XLS07]. Several examples of HRTF measurement setups are shown in Fig. 7.6. This is the most ideal solution but it is extremely tedious and involves highly precise measurements. These measurements also require the subjects to remain motionless for long periods, which may cause fatigue to the subjects. Zotkin *et al.* developed a fast HRTF measurement system using the technique of reciprocity, where a micro-speaker is placed into the ear and several microphones are placed around the listener [Nic10]. Other researchers

Table 7.3 Comparison of the various HRTF individualization techniques

| How to obtain individual features | Techniques | Pros | Cons | Performance and remarks |
|---|---|---|---|---|
| **Acoustical Measurements** | Individual measurements [MSH95],<br>IRCAM France, CIPIC, Univ. of Maryland, Tohoku Univ, Nagoya Univ., Austrian Academy of Sciences, etc. [XLS07] | Ideal, accurate | Requires high precision;<br>tedious;<br>impractical for every listener | Reference for individualization techniques |
| **Anthropometric data** | Optical Descriptors: 3D mesh, 2D pictures [Nic10] | Based on acoustic principles; studies the effects of independent elements of the morphology | Need a large database;<br>Tedious;<br>Requires high resolution imaging;<br>Expensive equipment;<br>Qualified users | Uses the correlation between individual HRTF and anthropometric data |
| | Analytical or numerical Solutions:<br>PCA + multiple linear regression [XLS07]<br>Finite element method, boundary element method [XLS07], [Nic10], Multiway array analysis [RDS10], Artificial neural network [XLS07] | | | |
| | Structural model of HRTFs [Nic10], HRTF database matching [ZHD03] | | | |
| **Listening/ Training** | Selection from non-individualized HRTF [Nic10], Frequency scaling [Mid99], training [MCD12] | Easy to implement; directly relates to perception | Takes time; requires regular training; causes fatigue | Obtains the best HRTFs perceptually |
| | Tune magnitude spectrum [TaG98], [Nic10], Active sensory tuning [XLS07], PCA weight tuning [FiR12], [FiR15] | | | |
| | Select cepstrum parameters [BBL06] | | | |
| **Playback Mode** | Frontal projection headphone [STG13] | No additional measurement, listening training | New structure; not applicable to normal headphones;<br>Type-2 equalization | Automatic customization, reduced front-back confusions |
| **Non-individualized HRTF** | Generalized HRTF | Easy to implement | Not accurate;<br>Poor localization | Not an individualization technique |

developed a continuous 3D azimuth acquisition system to measure the HRTFs using a multichannel adaptive filtering technique [Enz09]. Interpolation techniques are also employed to synthesize the HRTFs for the directions not measured [Gam13], [Rom12]. However, all these techniques to measure the



Air Force Research Laboratory, US

ISVR, University of Southampton, UK

IRCAM, France

South China University of Technology, China

Tohoku University, Japan

Nagaoka University of Technology, Japan

Figure 7.6 Examples of various setups to measure HRTF directly (pictures obtained online)

individual HRTFs acoustically require a large amount of resources and expensive setups.

### 7.5.2 Anthropometric data

Individualized HRTFs can also be modelled as weighted sums of basis functions, which can be performed either in the frequency or spatial domain. The basis functions are usually common to all individuals and the individualization information is often conveyed by the weights. The HRTFs are essentially expressed as weighted sums of a set of Eigen vectors, which can be derived from PCA or ICA [XLS07], [Nic10]. The individual weights are derived from the anthropometric parameters that are captured by optical

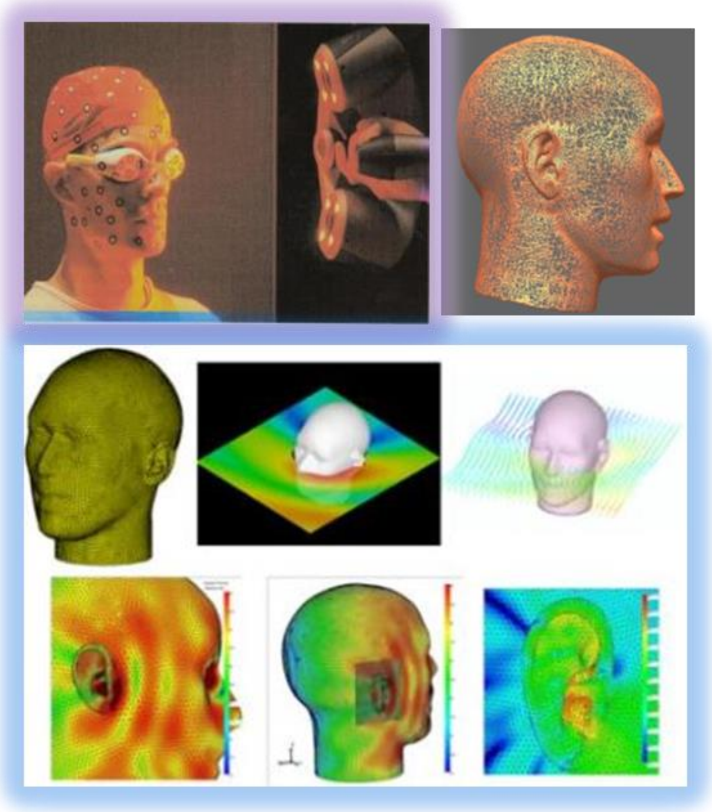


Figure 7.7 Numerical computation of HRTF using 3D meshes (picture extracted from [UoS15])

descriptors, which can be derived from direct measurements, pictures or a 3D mesh of the morphology [Nic10], as shown in Fig. 7.7. The solution to the problem of diffraction of an acoustic wave with the listener’s body results in individual HRTFs. This solution may be obtained by analytical or numerical methods, such as the boundary element method (BEM) or the finite element method (FEM) [Nic10], [XLS07]. Other methods used include multiple linear regressions [XLS07], multiway array analysis [RDS10], and artificial neural networks [XLS07]. The inputs to these methods can be a simple geometrical primitive [DAT02] (e.g., a sphere, cylinder or an ellipsoid), a 3D mesh obtained from MRI or laser scanner or a set of 2D images [Nic10]. An important advantage of these techniques is that the relative effects of a particular morphological element (e.g., torso, head, and pinna) and their variation with size, location and shape can be independently investigated [Nic10]. One of the major challenges today to numerically model the HRTF is the very high resolution of imaging techniques required for accurate prediction of HRTFs at high frequencies. The required resolution of the mesh imaging depends on the shortest wavelength, which is around 17mm at 20 kHz [Nic10]. Moreover, obtaining these optical descriptors demands for the use of extremely expensive laser, MRI scanners, and also requires highly skilled qualified users.

Another type of technique used a simple customization technique, where a HRTF is synthesized based on the matching or training of certain anthropometric features [ZHD03], [HZM08], [LiH13], [GrV07], [MDZ03], [SGA13], [ScK10], [HCT10], [LZD13], [HuG10], [BAT14], [Tas14], as illustrated in Fig. 7.8. The relationship can be trained between the anthropometry database and the corresponding HRTF database [ZHD03], [HZM08], [LiH13], where dimensionality reduction of HRTF database and selection of anthropometric features are critical [HZM08]. To avoid these difficulties, Tashev *el at* [BAT14], [Tas14] proposed an indirect anthropometry based HRTF individualization method. Instead of training the relation between HRTFs and anthropometry, their method obtains a sparse representation for the anthropometry of a new person using the anthropometry of the training subjects. This sparse representation is then used to synthesize the HRTFs of the new person using the HRTFs of the corresponding training subjects. Our study in [HGT15d] revealed that the use of preprocessing and post-processing methods plays an important role in affecting the performance of HRTF individualization.

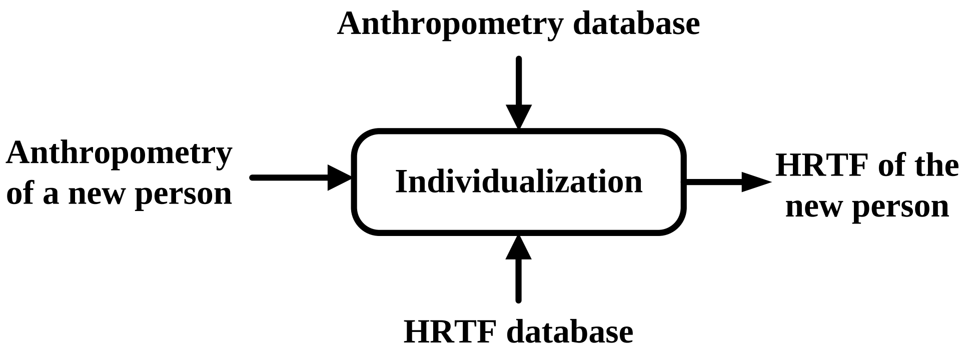


Figure 7.8 Obtaining individualized HRTF using anthropometry

### 7.5.3 Perceptual feedback

Several attempts have been carried out to personalize HRTF from a generic HRTF database using perceptual feedback, as shown in Fig. 7.9. Subjects select the HRTFs through listening tests, where they choose the HRTFs based on the correct perception of frontal sources and reduced front-back reversals [TaG98], [SeF03], [MSC12], [MCD12], [Nic10]. Listeners can also adapt to the non-individualized HRTF by modifying the HRTFs to suit his or her perception. Middlebrooks observed that the peaks and notches of HRTFs are frequency shifted for different individuals and that the extent of the shift is related to the size of pinna [Mid99]. Listeners often tune the spectrum until they achieve a good and natural spatialization [Nic10]. Other techniques involve active sensory tuning [XLS07], and tuning the PCA weights [FiR12], [FiR15] to individualize the HRTFs. These perceptual based methods are much simpler in terms of the required resources, and effort compared to the individualization methods using acoustical measurements or anthropometric data. However, these listening sessions can sometimes be quite long and result in listener fatigue.

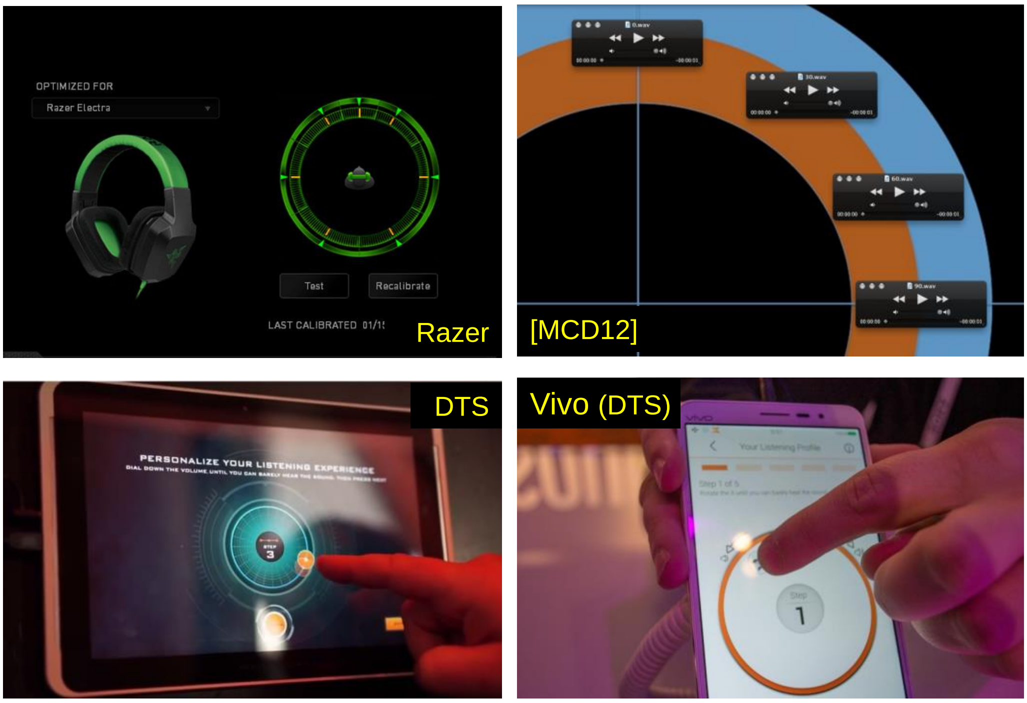


Figure 7.9 Academic and industrial examples of HRTF individualization based on training/tuning.

### 7.5.4 Frontal projection playback

More recently, a study by Sunder *et al.* [STG13] customized the non-individualized HRTFs using a frontal projection headphone, as illustrated in Fig. 7.10. By projecting the sound from the front, the idiosyncratic frontal pinna spectral cues of the listener are captured inherently during the playback [STG13]. The idiosyncratic high frequency pinna cues captured in the frontal projection headphones response match well with the frontal HRTF cues, giving it a better frontal perception (with front-back reversals reduced by almost 50%). The advantage of this technique is that it does not require any measurements, training or the anthropometric data of the listener. However, the frontal projection individualization technique has been limited to only the horizontal

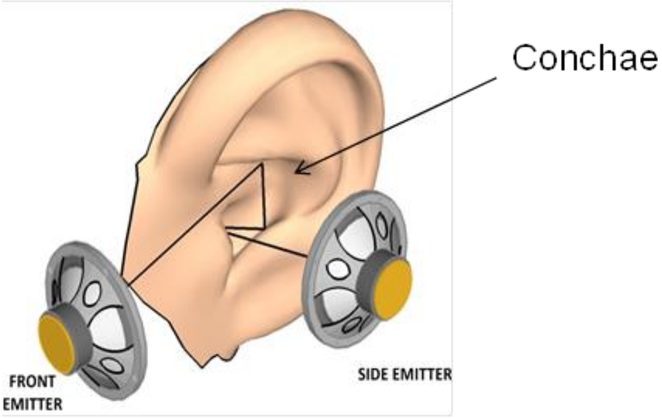


Figure 7.10 A geometric view of front emitter

plane and also requires a special kind of headphone equalization (i.e., Type-2, discussed in Section 7.6).

As discussed in Section 7.3, head tracking is important in the virtualization process. It was found that head tracking, when used with non-individualized HRTFs, can improve the localization [BWA01]. However, head tracking primarily helps in reducing the front-back confusions and has minimal effect in reducing the elevation localization errors, in-head localization [BWA01], and coloration caused by non-individualized HRTFs. Since individualization of HRTFs can alleviate some of these limitations, it is suggested that head tracking be used with individualized rendering.

To sum up, there is a noticeable trend to achieve more and more accurate individualization with lesser data, complexity and effort. However, the effect of individualization of HRTFs can be hindered by the presence of the headphone. Hence, the headphone has to be compensated to ensure that the spectrum at the eardrum has only the individualized HRTF features. Additionally, equalization of the binaural recording itself may be necessary in certain applications (e.g., musical recordings). The challenges and methods of equalization for both binaural and stereo recordings are explained in the Section 7.6.

## 7.6 Equalization

Headphones are not acoustically transparent as they not only color the sound that is played from the headphone but also affect the free-air characteristics at the ear. Typically, the HPTF comprises of the headphones transducer response and the acoustic coupling between the headphones and the listener's ears, as illustrated in Fig. 7.11. To compensate for the headphone response, the HPTF is first measured at the same point where the recording was carried out at the blocked ear canal or at the eardrum [MHJ95]. The binaural recording is then de-convolved with the HPTF to eliminate the effect of the recording microphones and the headphone. This type of direct equalization is known as the "non-decoupled" mode of equalization [LJV98]. This method is often used when the HPTF is measured with the same measurement setup as the recording and particularly works well when the HPTF measurement and recording are carried out on the same dummy head.

It is observed that, headphone equalization is critical in reducing the front-back reversals and elevation localization errors, and improve externalization [Beg00], [XLS07], [Nic10]. However, headphone equalization is challenging since the HPTF depends on individual morphology (headphone-ear coupling). Another difficulty in carrying out accurate headphone equalization is the variability of the HPTFs with repositioning [KuC00]. The positional dependency can only be reduced by taking the average

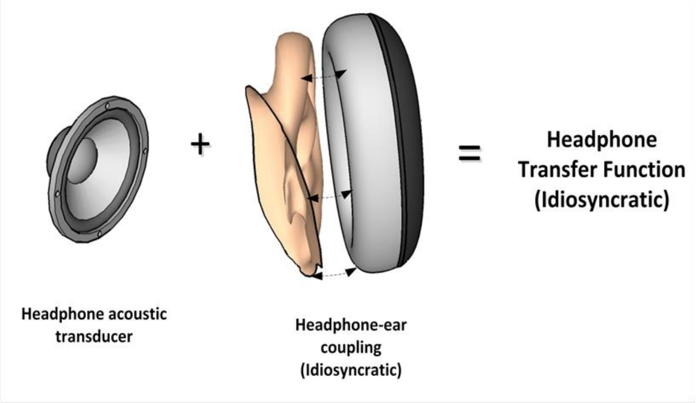

Headphone acoustic transducer
+
Headphone-ear coupling (Idiosyncratic)
=
Headphone Transfer Function (Idiosyncratic)

Figure 7.11 A breakdown of Headphone transfer function

of a number of trials as a representative HPTF [KuC00]. Thus, to create a convincing immersive sound environment, use of individualized HRTFs and individualized equalization is entailed, which may not be viable all the time. To reduce the dependency on individualized equalization, Sunder *et al.* [STG13] designed a Type-2 equalization technique for the playback through frontal projection headphone, which is independent of the headphone-ear coupling. Unlike the conventional equalization technique, Type-2 equalization compensates only for the distortion due to the emitter, thereby preserving the individual pinna cues due to frontal projection.

The other type of equalization is the "decoupled" equalization technique and it is the most commonly used method of equalization for rendering music. In this technique, the binaural recording (BIR or HRTFs) as well as the headphone are equalized using a reference sound field [LJV98]. If the reference sound field (REF) of the recording environment is well known and reproduced reliably, this method of equalization can result in a very natural perception of sound similar to the non-decoupled equalization technique. This method of equalization is mainly carried out to make the binaural recordings compatible with stereophonic (conventional microphone) recordings in terms of timbral quality. Some of the commonly used reference fields are: free-field (FF), diffuse-field (DF), and other more realistic reference fields including modified FF [OWM13], modified DF [MJH95], as well as RR_G and RR1_G proposed by Olive *et al.* [OWM13]. Ideally, the best reference field that preserves the true quality of the recording would be the field where the recording is carried out.

Furthermore, the choice of headphones can also greatly affect the transparency of the binaural rendering even with the correct headphone equalization. The external ear is un-hindered in the natural listening conditions, where the sound pressures at the ear are governed by free-air characteristics. With headphones placed over the ear, the pressure characteristics of the sound arriving at the eardrum are greatly affected compared to the free-air characteristics due to the interaction between the external ear and the headphone enclosure. The closer the coupling characteristic of the headphones with that of the free-air, the more accurate and transparent is the reproduced sound. Such headphones are defined as FEC (free-air equivalent coupling) headphones [MHJ95]. It is important to note that the FEC condition for the headphone is necessary only for binaural recordings made at the blocked ear canal, which is also the most common technique for individualized binaural recording [MHJ95].

## 7.7 Integration

An integration of these signal processing techniques for natural sound rendering reviewed in this chapter is depicted in Fig. 7.12. The original sound

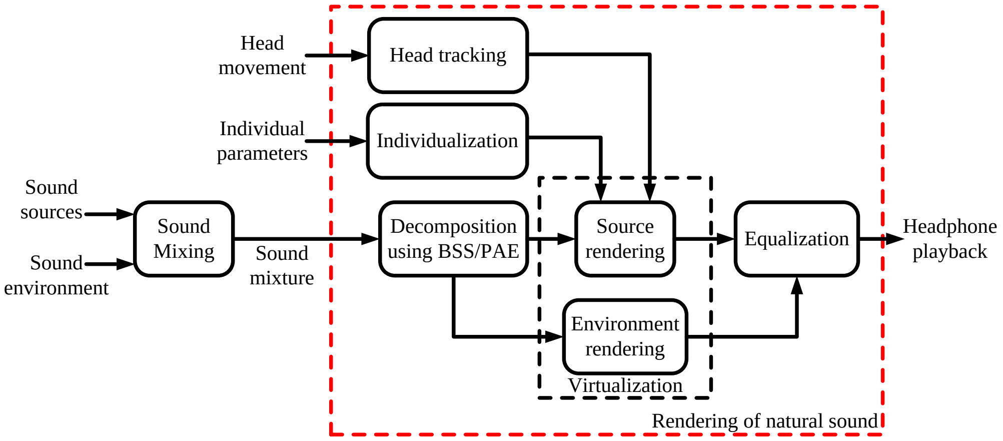


Figure 7.12 Natural sound rendering system for headphones: an integration of all the signal processing techniques reviewed in this chapter.

sources along with their environmental information are represented as a sound mixture after the mixing process. The sound scenes from the mix are then decomposed into primary components (sources) and/or ambient components (environment) using BSS and/or PAE. The extracted primary components, which are basically directional sound sources as perceived by the listener, can be rendered using (individualized) HRTFs [Beg00]. Ambient components are rendered in a manner so as to recreate a natural sound environment. Modelling the acoustics of the natural sound environment by adding the correct amount of early reflections and reverberation also helps in enhancing the perception of the sound environment as well as veridical distance, which is critical for natural listening. Moreover, a suitable individualization technique has to be applied to the directional sources such that the rendered sound scenes played over headphones are maximally tailored for the individual listener. Meanwhile, use of a robust equalization technique can significantly reduce the adverse coloration of the source. Finally, the influence of the head movements on the

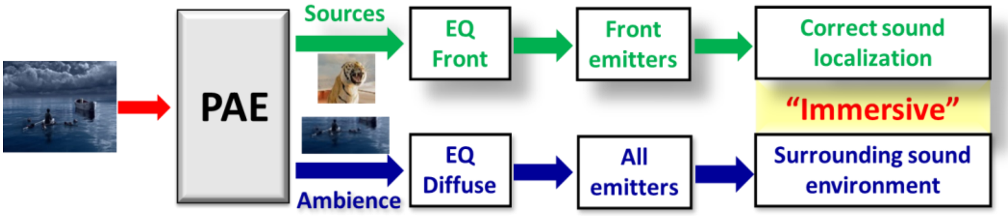


Figure 7.13 Natural sound rendering for 3D audio headphones

rendered sound can be taken into account by incorporating head tracking in virtualization.

In general, natural sound rendering requires both the spatial and timbral quality of the reproduced sound to be realistic. For digital media content that contains plenty of spatial cues (e.g., movies, games), all the five techniques reviewed are important in creating a sense of immersiveness. For other content, where the timbral quality is of utmost importance (e.g., music recordings), a subset of the techniques (e.g., individualization, equalization) are sufficient in natural sound rendering.

## 7.8 Subjective evaluation using 3D audio headphones

Subjective experiments were carried out to validate the reviewed natural sound rendering system by comparing it with the conventional stereo playback system. A total of 18 subjects (15 males and 3 females), who were all between 20-30 years old, participated in this listening experiment. None of the subjects reported any hearing loss. The test was conducted in a semi-anechoic listening room at NTU, Singapore. The two systems of headphone listening tested in this experiment were:

**(i) Conventional stereo system.** The materials are directly played back over headphones without any processing.

**(ii) Natural sound rendering system.** The signal processing techniques introduced in this chapter were applied to the audio content. In this study, we chose PAE as the sound scene decomposition method since our primary interest lies in movie and gaming audio content that contains the individual sound sources and the sound environment [HTG14]. Based on the recommendation in Chapter 3, least-squares is the selected PAE approach and the time-shifting technique discussed in Chapter 5 is employed. Individualization is carried out by frontal projection headphone since it inherently embeds the personal pinna cues during playback and does not require any individual acoustical experiments, anthropometric data or training [STG13]. To fully exploit the frontal projection in the natural sound rendering, we have developed a new four-emitter headphone [GaT14] that houses a frontal emitter and a conventional side emitter in each ear cup of the headphone [STG13]. In the virtualization process, the frontal emitters are used to render the directional sources, while all the emitters (both frontal and side) are used to render the sound environment. Type-2 EQ is applied to the frontal emitters for source rendering [STG13], and diffuse-field EQ is used to render environment signals over all the emitters. Head tracking has not been incorporated in this system. Fig. 7.13 indicates the specific natural sound rendering techniques for 3D audio headphones.

The stimuli used in this experiment were binaural (motorcycle in a storm and bee at a waterfall), movie (Brave, Prometheus), and gaming tracks (Battlefield 3), which contain plenty of spatial cues. Each track was played back using the two headphone playback systems tested here. The tracks corresponding to the two systems were named “A” and “B” and played back in a random order. The listening tests were conducted in a double-blind manner, where both the experimenter and the subjects were unaware of the order of the stimuli. In this experiment, four audio quality measures were considered to evaluate the performance of the two systems. Their descriptions are given below:

1. *Sense of direction*: how clear or distinct are the perceived directions of the sound objects?
2. *Externalization*: how clear is the stimulus perceived outside the head?
3. *Ambience*: how clear and natural is the ambience of the sound environment perceived?
4. *Timbral quality*: how realistic is the timbral quality of the sound?

Subjects were asked to give the scores for the four measures for each of the two tracks “A” and “B”. The scores were based on a 0-100 scale where subjects rated 0-20 (Bad), 21-40 (Poor), 41-60 (Fair), 61-80 (Good), and 81-100 (Excellent). Finally, the subjects were also required to indicate their overall preference for the two tracks by selecting one of the following three choices: “Prefer A”, “Not sure”, or “Prefer B”. To carry out this experiment, a GUI was created which randomized the order of the stimuli and automatically stored the responses of the subjects in a file.

The responses of the subjects were analyzed for both sound rendering systems. Fig. 7.14 shows the overall comparison between the two systems in terms of the mean opinion score (MOS), scatter plot and the overall preference of the subjects. In Fig. 7.14(a), MOS of the four measures for the two systems

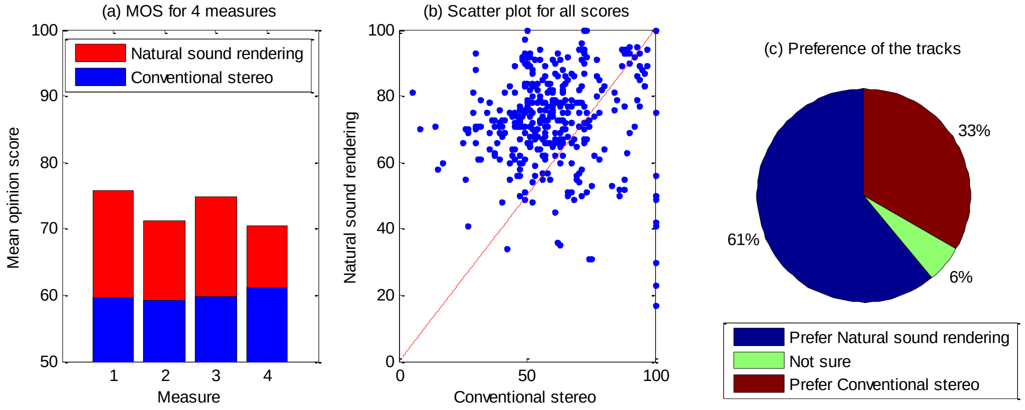


Figure 7.14 Results of the subjective experiments: (a) MOS, (b) scatter plot, and (c) overall preference.

were computed across all the 18 subjects and five stimuli. While the MOS for the conventional stereo system for all the measures were around 60, the natural sound rendering system performed much better with MOS of over 70. An analysis of variance (ANOVA) was conducted to generalize these results to the whole population of listeners. The *p*-values were found to be very small (<< 0.01) for all the measures, indicating that the improved performance of the natural sound rendering system over the conventional stereo system is statistically significant. The scatter plot in Fig. 7.14(b) implies that most of the subjects gave a higher score for the natural sound rendering system for all the four measures. The overall preference of the subjects across all the five tracks is shown in Fig. 7.14(c). The pie chart suggests that 61% of the subjects preferred the natural sound rendering, whereas only 33% preferred the conventional stereo rendering.

To sum up the subjective test results, we found that the natural sound rendering system using the various signal processing techniques explained in this chapter enhances the listening experience compared to a conventional stereo system. Additionally, the presence of head tracking in the system will only improve the natural sound rendering as observed in several studies [BWA01].

## 7.9 Conclusions and future directions

With the advent of low cost, low power, small form factor, and high speed multi-core embedded processor, we can now implement the above signal processing techniques in real-time and embed processors into the headphone design. However, various implementation issues regarding the computation cost of sound scene decomposition, HRTF/BRIR filtering in virtualization, and equalization as well as the latency in head tracking should be carefully considered. One example of such a natural sound rendering system is the four-emitter 3D audio headphone [GaT14] developed at the DSP Lab in NTU. This system has been psychophysically validated and found to perform much better than the conventional stereo headphone playback system.

Besides the five types of techniques discussed in this chapter, there have been other efforts to enhance the natural experience of headphone listening. To enable the natural pass through of the sound from outside world without coloration, headphones can be designed with suitable acoustically transparent materials. When this is not effective, microphones integrated into headphones and associated signal processing techniques, such as equalization [HJT04], and active noise control (ANC) [ScA05], are employed. The headphones with built-in microphones open a new dimension to augment the listening experience with the physical world [VFR15].

The future of headphones for assistive listening applications would be the one where listeners cannot differentiate between the virtual acoustic space created from headphone playback and the real acoustic space. This would require the combined effort from the whole audio community from the headphone manufacturers, sound engineers to audio scientists. More information about the content production has to be distributed from the content developers to the end user to enhance the extraction process. Moreover, obtaining and exploiting every individual’s anthropometrical features or hearing profiles is crucial for a natural listening experience. Finally, with more sensors, such as GPS, gyroscopes, and microphones that can be integrated into headphones, future headphones are becoming more content-aware, location-aware, listener-aware, and hence more intelligent and assistive.

# Chapter 8

# Conclusions and Future Works

In this chapter, we will summarize this thesis with conclusions drawn from our works as well as future works to be carried out as an extension of the thesis.

## 8.1 Conclusions

Spatial audio reproduction is essential in creating immersive and authentic listening experience, as per the increasing need from the consumer market. Primary ambient extraction can be applied in spatial audio reproduction to alleviate the rigorous requirements of the channel-based audio format on the audio reproduction system configuration. Thereby, PAE facilitates flexible, efficient, and immersive spatial audio reproduction. With the PAE approaches proposed for signals in the ideal case, little work has been carried out to study PAE for more practical real-world signals encountered in digital media content. Thus, spatial audio reproduction based on PAE was investigated in this thesis on the following five aspects.

First, a comprehensive study on existing PAE approaches was carried out. Our observations on existing PAE approaches like PCA, least-squares led us to a unified linear estimation framework, where the extracted (primary or ambient) components can be estimated as a weighted sum of the input signals. Furthermore, in order to quantify the objective performance of PAE, we introduced two groups of performance measures, namely, the measures for extraction accuracy and measures for spatial accuracy. For extraction accuracy, we identified three types of errors that contribute to the extraction error: the distortion, the interference, and the leakage. Dividing the extraction error into these three parts helps us understand the performance of PAE approaches. With the objectives of minimum leakage, minimum distortion, and adjustable performance, three variants of the least-squares method were proposed. The key relationships and differences among these linear estimation based PAE approaches were established in this thesis. Comparatively better performance was found in primary component extraction than in ambient component extraction, where primary power ratio also plays an important role. As a result of this comparative study, guidelines and recommendations on selecting the more suitable PAE approaches for various spatial audio reproduction applications were suggested.

Secondly, a novel ambient spectrum estimation (ASE) framework was proposed to improve the performance of PAE, especially when the ambient power is strong. Based on the relation of equal magnitude of ambient components in two channels, the ASE framework can be analyzed from two perspectives, i.e., ambient phase estimation (APE), and ambient magnitude estimation (AME). Equivalence between APE and AME was verified. The sparsity constraint of the primary components was employed in this thesis to solve the ASE problem, leading to two PAE approaches, APES, and AMES. To improve the computational efficiency, an approximate solution to the ASE problem with sparsity constraint, known as APEX, was further proposed. With the aim to apply the evaluation framework of extraction accuracy (as introduced earlier) in PAE approaches without analytical solutions (as is the case with these ASE approaches), an optimization method was proposed. It was evident from our objective and subjective experiments that the ASE approaches can improve the performance of PAE with 3-6 dB less extraction error (all cases, on average) and closer spatial cues. Furthermore, the experiments with variant ambient magnitude difference indicated the robustness of the ASE approaches.

Thirdly, when dealing non-ideal signals (signals that do not fit the signal model), we observed a significant performance degradation using conventional PAE approaches. One of the most often occurring case is the primary-complex case where the primary components are partially correlated at zero lag. The performance degradation generally increases as primary correlation decreases. Therefore, a time-shifting technique was proposed to maximize the primary correlation prior to PAE. Overlapped output mapping is introduced to alleviate the frame boundary switching artifacts due to varied time-shifting amounts. Simulations using synthesized signals and real recordings showed that the time-shifting technique can greatly enhance the performance PAE with around 50% lower extraction error and much more accurate spatial cues. Furthermore, the time-shifting technique can be seamlessly incorporated into any existing PAE approaches.

Fourthly, it is possible to encounter even more complex signals when dealing with actual sound scenes from digital media, where multiple concurrent dominant sources pose a challenge for PAE. Our study revealed that multi-shifting technique with ICC based weighting and subband technique with adaptive frequency bin partitioning could enhance the PAE performance with multiple sources.

Lastly, we discussed how PAE can be applied in spatial audio reproduction over headphones. Differences between headphone listening and natural listening were examined, which leads to a natural sound rendering paradigm for headphones. Five types of signal processing techniques, including PAE based sound scene decomposition and HRTF individualization, were discussed and integrated in the natural sound rendering system. Finally, an example of 3D audio headphones that implements the natural sound rendering was evaluated using subjective listening tests, which achieved a significant performance improvement over conventional headphone playback.

## 8.2 Future works

Through the investigations of PAE approaches reported in this thesis, there are several interesting future works that can be further explored, which are suggested as follows:

Firstly, the performance of various PAE approaches is only studied in the ideal case. In the non-ideal cases, only the performance of PCA is analyzed. Therefore, it is also interesting and beneficial to understand how the performance varies for the other PAE approaches. Some interesting results from this study could shed lights on how to design more specific techniques to improve the performance of PAE in non-ideal cases.

Secondly, it is commonly known that for spatial audio evaluation, timbre quality and spatial quality are two important aspects. Previous studies from Rumsey *et al.* showed that it is possible to combine the two aspects [RZK05]. Yet, it is unknown whether the model developed in [RZK05] is applicable to

PAE. Hence, one future work would involve extensive subjective listening tests to understand the relative importance of the timbre and spatial quality in PAE applications. Furthermore, considering that conducting subjective tests to evaluate all the PAE approaches is very tedious and impractical, objective evaluation is more preferred. Therefore, the relations between the subjective quality and objective quality would lead to a more meaningful and reliable objective evaluation. Besides, evaluation of PAE approaches in a more specific spatial audio reproduction application could help us understand the final performance of these PAE approaches.

Thirdly, further studies can be extended based on our ambient spectrum estimation (ASE) framework. In the current study, only the sparsity constraint is employed. Employing other constraints, such as the diffuseness of the ambient components and the independence between primary and ambient components, could improve the performance of PAE (or the performance in certain cases). Probabilistic approaches could be developed to model the ambient magnitude variations better.

Fourthly, more work still has to be carried out for complex signals in PAE. For those signals with multiple dominant source, we could further combine the multi-shift technique with the subband technique, which might lead to an optimal filtering method. Moreover, blind source separation techniques could be incorporated into PAE to separate the multiple sources. On the other hand, more comprehensive study shall be carried out on PAE for multichannel signals.

Lastly, PAE is a blind process, which implies that its performance relies heavily on how effective the signal model is. Due to the complexity of the actual sound scenes, not one signal model could satisfy any audio content. Therefore, machine learning techniques could be introduced to solve the PAE and spatial audio reproduction problem, thanks to the vast amount of digital media data. Furthermore, real-time implementation of the PAE approaches for spatial audio reproduction applications shall also be seriously considered.

# Author’s Publications

## Journal papers

[J1] **J. He**, E. L. Tan, and W. S. Gan, “Linear estimation based primary-ambient extraction for stereo audio signals,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 505-517, Feb .2014.

[J2] K. Sunder, **J. He**, E. L. Tan, and W. S. Gan, “Natural sound rendering for headphones: Integration of signal processing techniques,” *IEEE Sig. Process. Mag.*, vol. 32, no. 2, Mar 2015, pp. 100-113.

[J3] **J. He**, W. S. Gan, and E. L. Tan, “Primary-ambient extraction using ambient phase estimation with a sparsity constraint,” *IEEE Signal Process. Letters*, vol. 22, no. 8, pp. 1127-1131, Aug. 2015.

[J4] **J. He**, E. L. Tan, and W. S. Gan, “Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1430-1443, Sept. 2015.

[J5] **J. He**, W. S. Gan, and E. L. Tan, “Time-shifting based primary-ambient extraction for spatial audio reproduction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1576-1588, Oct. 2015.

## Conference papers

[C1] **J. He**, E. L. Tan, and W. S. Gan, “Time-shifted principal component analysis based cue extraction for stereo audio signals,” in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 266-270.

[C2] **J. He**, W. S. Gan, and E. L. Tan, “A study on the frequency-domain primary-ambient extraction for stereo audio signals,” in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2892-2896. (*Awarded SPS travel grant*)

[C3] **J. He**, W. S. Gan and Y. K. Chong, “Study on the use of error term in parallel-form narrowband feedback active noise control systems,” in *Proc. 2014 Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (invited)*, Cambodia, Dec. 2014.

[C4] **J. He**, W. S. Gan, and E. L. Tan, “On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometry features,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 639-643. (*Awarded SPS travel grant*)

[C5] **J. He**, and W. S. Gan, “Multi-shift principal component analysis based primary component extraction for spatial audio reproduction,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 350-354. (*Awarded SPS travel grant*)

[C6] S. Fasciani, **J. He**, B. Lam, T. Murao, and W. S. Gan, “Comparative study of cone-shaped versus flat-panel speakers for active noise control of multi-tonal signals in open windows,” in *Proc. Internoise 2015 (invited)*, San Francisco, Aug. 2015.

[C7] **J. He**, and W. S. Gan, “Applying primary ambient extraction for immersive spatial audio reproduction,” *2015 Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (invited)*, Hong Kong, Dec. 2015.

[C8] **J. He**, R. Ranjan, and W. S. Gan, “Fast continuous HRTF acquisition with unconstrained movements of human subjects,” in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp.

**[Tutorial]** W. S. Gan, and **J. He**, “Assisted listening for headphones and hearing aids: signal processing techniques,” Tutorial at *APSIPA ASC 2015*, Hong Kong, Dec. 2015.

**[Show & Tell]** D. H. Nguyen, **J. He**, K. K. Phyo, and W. S. Gan, “Real-time audio signal processing platform for natural 3D sound rendering,” Show & Tell at *ICASSP 2016*, Shanghai, China, Mar. 2016.

# References

[ADD02] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov and Z. Tang, "Approximating the head-related transfer function using simple geometric models of the head and torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053-2064, Nov. 2002.

[ADM01] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," in *Proc. IEEE WASSAP*, New Paltz, NY, Oct. 2001.

[ADT01] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IWASPAA*, New Paltz, NY, USA, Oct. 2001.

[Air15] Airforce Technology, 1 July 2015. "Terma to integrate 3D-Audio/ANR headset in BAE helmets," Available online: http://www.airforce-technology.com/news/newsterma-to-integrate-3d-audioanr-headset-in-bae-helmets-4554218

[AlB79] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943-950, 1979.

[AlD11] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Processing Mag*., vol. 28, no. 1, pp. 33–42, Jan. 2011.

[AnC09] Y. Ando, and P. Cariani, *Auditory and Visual Sensation*. New York: Springer, 2009.

[ASI08] Y. C. Arai, S. Sakakibara, A. Ito, K. Ohshima, T. Sakakibara, T. Nishi, et al., "Intra-operative natural sound decreases salivary amylase activity of patients undergoing inguinal hernia repair under epidural anesthesia," *Acta Anaesthesiologica Scandinavica*, vol. 52, no. 7, pp. 987-990, May 2008.

[ATT12] S. Aoki, M. Toba, and N. Tsujita, "Sound localization of stereo reproduction with parametric loudspeakers," *Applied Acoustics*, vol. 73, no.12, 1289-1295, Dec. 2012.

[Aur15] AURO-3D concept. 1 May 2015. Available online: http://www.auro-3d.com/system/concept/

[AvJ02] C. Avendano, and J. M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," In *Proc. ICASSP*, pp. 13-17, May 2002.

[AvJ04] C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740-749, Jul./Aug. 2004.

[BaF03] F. Baumgarte, and C. Faller, "Binaural cue coding-Part I: psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp.509-519, Nov. 2003.

[BaS07] M. R. Bai and G. Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *IEEE Trans. Consumer Electron.*, vol. 53, no. 3, pp. 1011-1019, Aug. 2007.

[BAT14] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. Tashev, and J. C. Plata, "HRTF magnitude synthesis via sparse representation of anthropometric features," in Proc. *IEEE ICASSP*, Florence, Italy, pp. 4501-4505, May 2014.

[BCH11] J. Benesty, J. Chen, and Y. Huang, "Binaural noise reduction in the time domain with a stereo setup," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no.8, pp. 2260-2272, Nov. 2011.

[Beg00] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Cambridge, MA: AP Professional, 2000.

[Ber88] A. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, Dec. 1988.

[BeZ07] S. Bech, and N. Zacharov. *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons, 2007.

[BHK07] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen and S. van de Par, "Background, concept, and architecture for the recent MPEG Surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 331-351, May, 2007.

[Bla97] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*. Cambridge, MA, USA: MIT Press, 1997.

[Blu31] A. D. Blumlein, “Improvements in and relating to sound-transmission, sound-recording and sound reproducing systems.” British Patent 394 325, 1931.

[BJP12] Y. H. Baek, S. W. Jeon, Y. C. Park, and S. Lee, “Efficient primary-ambient decomposition algorithm for audio upmix,” in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.

[BNK10] J. Breebaart, F. Nater, and A. Kohlrausch, “Spectral and spatial parameter resolution requirements for parametric, filter-bank-based HRTF processing,” *J. Audio Eng. Soc.*, vol. 58, no. 3, pp. 126-140, Mar. 2010.

[BrD98] C. P. Brown and R. O. Duda, “A structural model for binaural sound synthesis,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 5, pp. 476-488, Sept. 1998.

[Bre90] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990.

[BrF07] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and other Applications.* Hoboken, NJ: Wiley, 2007.

[BrS08] J. Breebaart and E. Schuijers, “Phantom materialization: a novel method to enhance stereo audio reproduction on headphones,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol.16, no. 8, pp. 1503-1511, Nov. 2008.

[BVM06] M. Briand, D. Virette and N. Martin, “Parametric representation of multichannel audio based on principal component analysis,” in *Proc. 120th Audio Eng. Soc. Conv.*, Paris, 2006.

[BVV93] A. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[BWA01] 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, Oct. 2001.

[BWG10] D. R. Begault, E. M. Wenzel, M. Godfroy, J. D. Miller, and M. R. Anderson, “Applying spatial audio to human interfaces: 25 years of NASA experience,” in *Proc. 40th AES Intl. Conf. Spatial Audio,* Tokyo, Oct. 2010.

[Cap69] J. Capon, “High resolution frequency wave number spectrum analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[CCK14] H. Chung, S. B. Chon, and S. Kim, “Flexible audio rendering for arbitrary input and output layouts,” in *Proc. 137th AES Conv.*, Los Angeles, CA, Oct. 2014.

[DAT02] R. O. Duda, V. R. Algazi, and D. M. Thompson, “The use of head-and-torso models for improved spatial sound synthesis,” in *Proc. 113th AES Conv.*, Los Angeles, Oct. 2002.

[DLH03] G. B. Diette, N. Lechtzin, E. Haponik, A. Devrotes, and H. R. Rubin, “Distraction therapy with nature sights and sounds reduces pain during flexible bronchoscopy: a complementary approach to routine analgesia,” *Chest Journal*, vol. 123, no. 3, pp. 941-948, Mar. 2003.

[DHT12] S. Dong, R. Hu, W. Tu, X. Zheng, J. Jiang, and S. Wang, “Enhanced principal component using polar coordinate PCA for stereo audio coding,” in *Proc. ICME*, Melbourne, Australia, 2012, pp. 628-633.

[Dol13] Dolby Atmos-Next Generation Audio for Cinema (White Paper). 2013. Available online:
http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/Dolby-Atmos-Next-Generation-Audio-for-Cinema.pdf

[Enz09] G. Enzner, “3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering,” *Proc. IWASPAA,* pp. 325-328, Oct. 2009.

[EVH11] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “subjective and objective quality assessment of audio source separation,” *IEEE Trans. Audio Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046-2057, Sept. 2011.

[FaB03] C. Faller, and F. Baumgarte, “Binaural cue coding-Part II: Schemes and applications,” *IEEE Trans. Speech Audio Process.,* vol. 11, no. 6, pp. 520-531, Nov. 2003.

[FaB11] C. Faller and J. Breebaart, “Binaural reproduction of stereo signals using upmixing and diffuse rendering,” in *Proc. 131th Audio Eng. Soc. Conv.,* New York, 2011.

[Fal04] C. Faller, “Coding of spatial audio compatible with different playback formats,” in *Proc. 117th AES Conv.*, San Francisco, CA, Oct. 2004.

[Fal06] C. Faller, “Multiple-loudspeaker playback of stereo signals,” *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051-1064, Nov. 2006.

[Fal06b] C. Faller, “Parametric multichannel audio coding: synthesis of coherence cues,” *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 14, no. 1, pp. 299-310, Jan. 2006.

[Fal07] C. Faller, “Matrix surround revisited,” in *Proc. 30$^{th}$ AES Int. Conf.,* Saariselka, Finland, Mar. 2007.

[FaM04] C. Faller, and J. Merimaa, “Source localization in complex listening situations: selection of binaural cues based on interaural coherence,” *J. Acoust. Soc. Am.,* vol. 116, no. 5, pp. 3075-3089, Nov. 2004.

[FiR12] K. J. Fink and L. Ray, “Tuning principal component weights to individualize HRTFs,” in *Proc. ICASSP,* Kyoto, pp. 389-392, Mar. 2012.

[FiR15] K. J. Fink, and L. Ray, “Individualization of head related transfer functions using principal component analysis,” *Applied Acoustics*, vol. 87, pp. 162-173, 2015.

[Fle40] H. Fletcher, “Auditory patterns,” *Rev. Mod. Psys.*, vol. 12, no. 1, pp. 47-65, 1940.

[Gam13] H. Gamper, “Head-related transfer function interpolation in azimuth, elevation, and distance,” *J. Acoust. Soc. Am.*, vol. 134, pp. EL547–554, Dec. 2013.

[GaM95] W. G. Gardner, and K. D. Martin, “HRTF Measurements of a KEMAR,” *J. Acoust. Soc. Am., vol.*, vol. 97, pp. 3907-3908, 1995.

[Gar97] W. Gardner, “3-D audio using loudspeakers,” PhD thesis, School of Architecture and planning, MIT, USA, 1997.

[Gar00] J. Garas, *Adaptive 3D sound systems*. Springer, 2000.

[GaS79] A. Gabrielsson and H. Sjogren, “Perceived sound quality of sound reproducing systems,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 1019–1033, 1979.

[GaT14] W. S. Gan and E. L. Tan, “Listening device and accompanying signal processing method,” US Patent 2014/0153765 A1, 2014.

[Ger73] M. A. Gerzon, “Perophony: with-height sound reproduction,” *J. Audio Eng. Soc.,* vol. 21, no. 1, pp. 3-10, 1973.

[Ger92] M. A. Gerzon, “Optimal reproduction matricies for multispeaker stereo,” *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 571–589, Jul./Aug. 1992.

[God08] M. Goodwin, “Geometric signal decompositions for spatial audio enhancement,” in *Proc. ICASSP,* Las Vegas, 2008, pp. 409-412.

[GoJ06a] M. Goodwin and J.-M. Jot. “A frequency-domain framework for spatial audio coding based on universal spatial cues,” in *Proc. 120th AES Conv*., May 2006.

[GoJ06b] M. Goodwin and J.-M. Jot. “Analysis and synthesis for universal spatial audio coding,” in *Proc. 121st AES Conv.,* Oct. 2006.

[GoJ07a] M. Goodwin and J. M. Jot, “Binaural 3-D audio rendering based on spatial audio scene coding,” in *Proc. 123rd AES Conv.,* New York, 2007.

[GoJ07b] M. Goodwin and J. M. Jot, “Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement,” in *Proc. ICASSP,* Hawaii, 2007, pp. 9-12.

[GoJ07c] M. Goodwin and J. M. Jot. “Multichannel surround format conversion and generalized up-mix,” in *Proc. AES 30th Intl. Conf.,* March 2007.

[GoJ08] M. Goodwin and J. M. Jot, “Spatial audio scene coding,” in *Proc. 125th AES Conv.*, San Francisco, 2008.

[GrV07] G. Grindlay and M. A. O. Vasilescu, “A multilinear approach to HRTF personalization,” in Proc. *IEEE ICASSP*, Honolulu, Hawaii, USA, Apr. 2007.

[GTK11] W. S. Gan, E. L. Tan, and S. M. Kuo, “Audio projection: directional sound and its application in immersive communication,” *IEEE Sig. Process. Mag.*, vol. 28, no. 1, pp. 43-57, Jan. 2011.

[Hab14] E. Habets. (2014, Aug. 1). Emanuel Habets’s website | RIR Generator [Online], Available: http://home.tiscali.nl/ehabets/rir_generator.html

[HaR09] S. Haykin, and K. J. Ray Liu, *Handbook on array processing and sensor networks*, Wiley-IEEE Press, 2009.

[Har11] A. Härmä, “Classification of time-frequency regions in stereo audio,” *J. Audio Eng. Soc*., vol. 59, no. 10, pp. 707-720, Oct. 2011.

[HCT10] Z. Haraszy, D.-G. Cristea, V. Tiponut, and T. Slavici, “Improved head related transfer function generation and testing for acoustic virtual reality development,” in *Proc. WSEAS CSCC ICS*, Corfu Island, Greece, Jul. 2010.

[HeG15] J. He, and W. S. Gan, “Multi-shift principal component analysis based primary component extraction for spatial audio reproduction,” in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 350-354.

[HeG15b] J. He, and W. S. Gan, "Applying primary ambient extraction for immersive spatial audio reproduction," in *Proc. APSIPA ASC*, Hong Kong, Dec. 2015.

[HGC09] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Letters*, vol. 16, no. 9, pp. 770-773, Sep. 2009.

[HGT14] J. He, W. S. Gan, and E. L. Tan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2892-2896.

[HGT15a] J. He, W. S. Gan, and E. L. Tan, "Primary-ambient extraction using ambient phase estimation with a sparsity constraint," *IEEE Signal Process. Letters*, vol. 22, no. 8, pp. 1127-1131, Aug. 2015.

[HGT15b] J. He, W. S. Gan, and E. L. Tan, "Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1430-1443, Sept. 2015.

[HGT15c] J. He, W. S. Gan, and E. L. Tan, "Time-shifting based primary-ambient extraction for spatial audio reproduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1576-1588, Oct. 2015.

[HGT15d] J. He, W. S. Gan, and E. L. Tan, "On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometry features," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 639-643.

[HHK14] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio – the new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, Dec. 2014.

[HHK15] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio – the new standard for coding of immersive spatial audio," *J. Sel. Topics Sig. Process.*, to appear, 2015.

[HMS11] K. Hamasaki, K. Matsui, I. Sawaya, and H. Okubo, "The 22.2 multichannel sounds and its reproduction at home and personal environment," in *Proc. AES 43rd Intl. Conf. Audio for Wirelessly Networked Personal Devices*, Pohang, Korea, Sept. 2011.

[HiD09] J. Hilpert, and S. Disch, "The MPEG surround audio coding standard," *IEEE Sig. Process. Mag.*, vol. 26, no. 1, pp.148-152, Jan. 2009.

[HJT04] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, et al., "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, no. 6, pp. 618-639, Jun. 2004.

[HKB08] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, K.S. Chong, "MPEG Surround-the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, Nov. 2008.

[HKO04] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ: Wiley, 2004.

[Hol08] T. Holman, *Surround sound up and running 2nd ed*., MA: Focal Press, 2008.

[HPB08] A. Härmä, S. V. D. Par, and W. D. Bruijin, "On the use of directional loudspeakers to create a sound source close to the listener," in *Proc. 124th AES Conv.*, Amsterdam, The Netherlands, May 2008.

[HPK12] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, H., and H. Oh, "MPEG Spatial Audio Object Coding – the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 655-673, Sept. 2012.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[HTG13] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 266-270.

[HTG14] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process*., vol. 22, no. 2, pp. 505-517, Feb. 2014.

[HuG10] W. W. Hugeng and D. Gunawan, "Improved method for individualization of head-related transfer functions on horizontal plane using reduced number of anthropometric measurements," *J. Telecom.*, vol. 2, no. 2, pp. 31–41, May 2010.

[HZM08] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, no. 2, pp. 163–172, Feb. 2008.

[JMG07] J. M. Jot, J. Merimaa, M. Goodwin, A. Krishnaswamy, and J. Laroche, "Spatial audio scene coding in a universal two-channel 3-D stereo format," in *Proc. 123rd AES Conv.,* New York, NY, Oct. 2007.

[HPB05] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and F. Myburg, "The Reference Model Architecture for MPEG Spatial Audio Coding," in *Proc. 118th AES Conv.*, Barcelona, Spain, May, 2005.

[HWZ14] R. Hu, X. Wang, M, Zhao, D. Li, S. Wang, L. Gao, C. Yang, and Y. Yang, "Review on three-dimension audio technology," *J. Data Acquisition and Process.*, vol. 29, no. 5, pp. 661-676, Sep. 2014.

[Ios15] IOSONO, http://www.iosono-sound.com/home/, 2015.

[IrA02] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 914-926, Nov. 2002.

[ITU93] Rec. ITU-R BS.775, "Multi-Channel Stereophonic Sound System with or without Accompanying Picture," ITU, 1993, Available: http://www.itu.org.

[ITU03] Rec. ITU-R BS.1284-1, "RECOMMENDATION ITU-R BS.1284-1* General methods for the subjective assessment of sound quality," 2003.

[ITU03b] ITU, "ITU-R Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," 2003.

[ITU12] Recommendation ITU-R BS.775-3, "Multichannel stereophonic sound system with and without accompanying picture," Geneva, Aug. 2012.

[ITU12b] ITU, "Report ITU-R BS.2159-4: Multichannel sound technology in home and broadcasting applications," 2012.

[ITU14] ITU, "ITU-R Recommendation BS.1534-2: Method for the subjective assessment of intermediate quality levels of coding systems," 2014.

[Jef48] A. Jeffress, "A place theory of sound localization," *J. Comput. Physiol. Psychol.,* vol. 41, no. 1, pp. 35-39, Feb. 1948.

[JHS10] S. W. Jeon, D. Hyun, J. Seo, Y. C. Park, and D. H. Youn, "Enhancement of principal to ambient energy ratio for PCA-based parametric audio coding," in *Proc. ICASSP*, Dallas, 2010, pp. 385-388.

[JJF10] J. D. Johnston, J. M. Jot, Z. Fejzo, and S. R. Hastings, "Beyond Coding: Reproduction of Direct and Diffuse Sounds in Multiple Environments," in *Proc. 129$^{th}$ AES Conv.*, San Francisco, CA, Nov. 2010.

[JLP99] J. M. Jot, V. Larcher, and J. M. Pernaux, "A Comparative Study of 3-D Audio Encoding and Rendering Techniques," in *Proc. 16th AES Int. Conf.,* Rovaniemi, Finland, 1999.

[JoF11] J. M. Jot, and Z. Fejzo, "Beyond surround sound - creation, coding and reproduction of 3-D audio soundtracks," in *Proc. 131st AES Conv.*, New York, NY, Oct. 2011.

[Jol02] I. Jolliffe, *Principal component analysis, 2nd ed.*, New York: Springer-Verlag, 2002.

[JPL10] S. W. Jeon, Y. C. Park, S. Lee, and D. Youn, "Robust representation of spatial sound in stereo-to-multichannel upmix," in *Proc. 128th AES Conv.*, London, UK, 2010.

[JSY98] P. X. Joris, P. H. Smith, and T. Yin, "Coincidence detection in the auditory system: 50 years after Jeffress," *Neuron*, vol. 21, no. 6, pp.1235-1238, Dec. 1998.

[KaN14] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. 3530-3541, Jun. 2014.

[KDN09] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 17, no. 4, pp. 534–545, May. 2009.

[Ken95a] G. Kendall, "A 3-D sound primer: directional hearing and stereo reproduction," *Computer Music Journal*, vol. 19, no. 4, pp. 23-46, Winter 1995.

[Ken95b] G. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, Winter 1995.

[SHT15] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones: integration of signal processing techniques," *IEEE Sig. Process. Mag.,* vol. 32. no. 2, pp. 100-113, Mar. 2015.

[KKL07] S. J. Kim, K. Koh, M. Lusig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale $l_1$-regularized least squares," *J. Sel. Topics Sig. Process.*, vol. 1, no. 4, pp. 606-617, Dec. 2007.

[KKM15] P. Kleczkowski, A. Krol, and P. Malecki, "Multichannel sound reproduction quality improves with angular separation of direct and reflected sounds," *J. Audio Eng. Soc.*, Vol. 63, No. 6, pp. 427-442, Jun. 2015.

[KTT15] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing," *IEEE Sig. Process. Mag.*, vol. 32, no. 2, Mar 2015, pp. 31- 42.

[KuC00] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1071-1074, Feb. 2000.

[LaA87] P. J. V. Laarhoven, and E. H. Aarts, *Simulated annealing*, Netherlands: Springer, 1987.

[LBH09] T. Lossius, P. Baltazar, and T. de la Hogue. "DBAP–distance based amplitude panning," In *Proc. Intl. Computer Music Conf.*, Montreal, Canada, 2009

[LBP14] T. Lee, Y. Baek, Y. C. Park, and D. H. Youn, "Stereo upmix-based binaural auralization for mobile devices," *IEEE Trans. Consum. Electron.*, vol. 60, no. 3, pp.411-419, Aug. 2014.

[LiH13] L. Li and Q. Huang, "HRTF personalization modeling based on RBF neural network," in Proc. *IEEE ICASSP*, Vancouver, British Columbia, Canada, May 2013.

[LJV98] V. Larcher, J. M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Proc. 105th AES Conv.*, SanFrancisco, Sep. 1998.

[LMG05] J. M. Loomis, J. R. Marston, R. G. Golledge, and R. L. Klatzky, "Personal guidance system for people with visual impairment: a comparison of spatial displays for route guidance," *J. Vis. Impair blind*, vol. 99, no. 4, pp. 219-232, Jan. 2005.

[LNZ14] J. Liebetrau, F. Nagel, N. Zacharov, K. Watanabe, C. Colomes, P. Crum, T. Sporer, and A. Mason, "Revision of Rec. ITU-R BS. 1534," in *Proc. 137th AES conv.*, LA, Oct. 2014.

[LWW09] Y. Li, J. Woodruff, and D. L. Wang. "Monaural musical sound separation using pitch and common amplitude modulation," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 7, pp. 1361–1371, Jul. 2009.

[LZD13] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process data fusion for heterogeneous HRTF datasets," in *Proc. IWASPAA*, New Paltz, New York, USA, Oct. 2013.

[Mat13] The MathWorks, Inc. "Find the local maxima." Internet: http://www.mathworks.com/help/signal/ref/findpeaks.html, [Apr. 10, 2013].

[MCD12] C. Mendonca, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the Improvement of Localization Accuracy with Non-individualized HRTF-Based Sounds," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 821-830, Oct. 2012.

[MDZ03] A. Mohan, R. Duraiswami, D. N. Zotkin, D. DeMenthon, and L. S. Davis, "Using computer vision to generate customized spatial audio," in *Proc. IEEE ICME*, Baltimore, Maryland, USA, Jul. 2003.

[MeF09] F. Menzer, and C. Faller, "Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching," in *Proc. 126th AES Conv.*, Munich, Germany, May 2009.

[MeF10] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th AES Conv.*, London, UK, 2010.

[MGJ07] J. Merimaa, M. Goodwin, J. M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd AES Conv.*, New York, Oct. 2007.

[MHJ95] H. Møller, D. Hammershoi, C. B. Jensen, and M. F. Sorensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203-217, Apr. 1995.

[Mic14] The next web, 1 July 2015. "Microsoft pilots 3D audio technology to help blind people navigate," Available online: http://thenextweb.com/microsoft/2014/11/06/microsoft-pilots-3d-audio-technology-help-blind-people-navigate/

[Mid99] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480-1492, Sept. 1999.

[Mil72] W. Mills, Auditory localization. In J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory*. New York: Academic Press. 1972.

[Mit06] S. K. Mitra, *Digital signal processing: a computer-based approach, 3rd ed*. New York: McGraw-Hill, 2006.

[MJH95] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, “Design criteria for headphones,” *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218-232, Apr. 1995.

[MMS11] F. Melchior, U. Michaelis, and R. Steffens, “Spatial mastering-a new concept for spatial sound design in object-based audio scenes,” in *Proc. Intl Computer Music Conf.*, University of Huddersfield, UK, Jul. 2011.

[Mol92] H. Møller, “Fundamentals of binaural technology,” *Applied Acoustics*, vol. 36, no. 3, pp.171-218, Mar. 1992.

[Moo98] B. C. J. Moore, *Cochlear hearing loss*. London: Whurr Publishers Ltd. 1998.

[MSC12] C. Mendonca, J. A. Santos, G. Campos, P. Dias, and J. Vieira, “On the adaptation to non-individualised HRTF auralisations: a longitudinal study,” in *Proc. 45th AES Intl Conf.*, Helsinki, Finland, Mar. 2012.

[MSH95] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, “Head-Related Transfer Functions of Human Subjects,” *J. Aud. Eng. Soc.*, vol. 43, no. 5, pp. 300-321, May 1995.

[MWC99] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, “Controlling phantom image focus in a multichannel reproduction system,” in *Proc. 107th AES Conv.*, New York, NY, Sept. 1999.

[Nic10] R. Nicol, *Binaural Technology*. AES, 2010.

[NLB06] R. Nicol, V. Lemaire, A. Bondu, and S. Busson, “Looking for a relevant similarity criterion for HRTF clustering: a comparative study,” in *Proc. 120th AES Conv.*, Paris, France, May 2006.

[OWM13] S. Olive, T. Welti, and E. McMullin, “Listener preferences for different headphone target response curves,” in *Proc. 134th AES Conv.*, Rome, Italy, May 2013.

[PBD10] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, “Sparse representation in audio and music: from coding to source separation,” *Proc. IEEE*, vol. 98, no. 6, pp. 995-1016, Jun. 2010.

[PKV99] V. Pulkki, M. Karjalainen, and V. Välimäki, “Localization, coloration, and enhancement of amplitude-panned virtual sources,” in *Proc. 16th AES Intl. Conf. Spatial Sound Reproduct.*, Rovaniemi, Finland, Apr. 1999.

[PoB04] G. Potard, I. Burnett, “Decorrelation techniques for the rendering of apparent sound source width in 3D audio displays,” in *Proc. DAFx’04*, Naples, Italy, Oct. 2004.

[Pol05] M. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, Nov. 2005.

[PuK08] V. Pulkki, and M. Karjalainen, “Multichannel audio rendering using amplitude panning [DSP Applications],” *IEEE Sig. Process. Mag.*, vol. 25, no. 3, pp.118-122, May 2008.

[Pul07] V. Pulkki, “Spatial Sound Reproduction with Directional Audio Coding,” *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503-516, Jun. 2007.

[Pot06] G. Potard, *3D-audio object oriented coding*, PhD thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, 2006, Available: http://ro.uow.edu.au/theses/539.

[Pul97] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[Ray07] L. Rayleigh, “On Our Perception of Sound Direction,” *Philosophical Magazine*, *6th Series.*, vol. 13, pp. 214-323, 1907.

[RDS10] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," in *Proc. 18th EUSIPCO*, Aalborg, Denmark, pp. 229-23, Aug. 2010.

[Rom12] G. D. Romigh, “Individualized hread-related transfer functions: efficient modeling and estimation from small sets of spatial samples,” PhD dissertation, School of Electrical and Computer Engineering, Carnegie Mellon University, 2012.

[RoW08] N. Roman and D. L. Wang, “Binaural tracking of multiple moving sources,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 728-739, May 2008.

[Rum99] F. Rumsey, “Controlled subjective assessments of two-to-five channel surround sound processing algorithms,” *J. Audio Eng. Soc.*, vol. 47, no. 7/8, pp. 563–582, Jul./Aug. 1999.

[Rum01] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001.

[Rum02] F. Rumsey, “Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm,” *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666, Sept. 2002.

[Rum10] F. Rumsey, “Time-frequency processing for spatial audio,” *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 655–659, Jul./Aug. 2010.

[Rum11] F. Rumsey, “Spatial audio: eighty years after Blumlein,” *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 57–62, Jan./Feb. 2011.

[Rum13] F. Rumsey, “Spatial audio processing: upmix, downmix, shake it all about,” *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 474–478, Jun. 2013.

[RVE10] M. Raspaud, H. Viste, and G. Evangelista, “Binaural source localization by joint estimation of ILD and ITD,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[RZK05] F. Rumsey, S. Zielicski, R. Kassier, and S. Bech, “On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality,” *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976, Feb. 2005.

[SAM06] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Blind extraction of dominant target sources using ICA and time–frequency masking,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2165–2173, Nov. 2006.

[SBP04] E. Schuijers, J. Breebaart, H. Purnhagen, J. Engdegard, “Low complexity parametric stereo coding,” in *Proc. 116th AES Conv*., Berlin, Germany, May 2004.

[ScA05] D. W. Schobben and R. M. Aarts, “Personalized multi-channel headphone sound reproduction based on active noise cancellation,” *Acta acustica united with acustica*, vol. 91, no. 3, pp. 440-450, May/Jun. 2005.

[Sch58] M. Schroeder, “An Artificial Stereophonic Effect Obtained from a Single Audio Signal,” *J. Audio Eng. Soc.*, vol. 6, no. 2, pp. 74–79, Feb. 1958.

[ScK10] D. Schonstein and B. F. G. Katz, “HRTF selection for binaural synthesis from a database using morphological parameters,” in *Proc. ICA*, Sydney, Australia, Aug. 2010.

[SeF03] B. U. Seeber and H. Fastl, “Subjective selection of non-individual head-related transfer functions,” in *Proc. ICAD*, pp. 259-262, Boston, Jul. 2003.

[SGA13] S. Spagnol, M. Geronazzo, and F. Avanzini, “On the relation between pinna reflection patterns and head-related transfer function features,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 508–519, Mar. 2013.

[SPL10] T. Särkämö, E. Pihko, S. Laitinen, A. Forsblom, S. Soinila, M. Mikkonen, et al., “Music and speech listening enhance the recovery of early sensory processing after stroke,” *J. Cog. Neuroscience*, vol. 22, no. 12, pp. 2716-2727, Dec. 2010.

[SoE15] Sonic Emotion, http://www2.sonicemotion.com/professional/, 2015

[SRK12] A. Schwarz, K. Reindl, W. Kellermann, “A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering,” *Proc. ICASSP*, pp.113-116, 2012.

[STG13] K. Sunder, E. L. Tan, and W. S. Gan, “Individualization of binaural synthesis using frontal projection headphones,” *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 989-1000, Dec. 2013.

[StM15] N. Stefanakis, and A. Mouchtaris, “Foreground suppression for capturing and reproduction of crowded acoustic environments,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 51-55.

[SWB06] R. M. Stern, D. Wang, and G. J. Brown, *Computational auditory scene analysis*. Piscataway, NJ: Wiley/IEEE Press, 2006.

[SWR13] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, “Spatial sound with loudspeakers and its perception: a review of the current state,” *Proc. IEEE*, vol. 101, no. 9, pp.1920-1938, Sept. 2013.

[TaG98] C. J. Tan and W. S. Gan, “User-defined spectral manipulation of HRTF for improved localisation in 3 D sound systems,” *Electronics letters*, vol. 34, no. 25, pp. 2387-2389, Dec. 1998.

[TaG12] E. L. Tan, and W. S. Gan, “Reproduction of immersive sound using directional and conventional loudspeakers,” *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3215-3215, Apr. 2012.

[Tas14] I. Tashev, “HRTF phase synthesis via sparse representation of anthropometric features,” in Proc. *Information Theory and Applications Workshop (ITA)*, San Diego, CA, pp. 1-5, Feb. 2014.

[TGC12] E. L. Tan, W. S. Gan, and C. H. Chen, “Spatial sound reproduction using conventional and parametric loudspeakers,” in *Proc. APSIPA ASC*, Hollywood, CA, 2012.

[ThH12] O. Thiergart and E.A.P. Habets, “Sound field model violations in parametric spatial sound processing,” *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2012.

[ThP77] G. Theile and G. Plenge, “Localization of lateral phantom sources,” *J. Aud. Eng. Soc.*, vol. 25, no. 4, pp. 196–200, Apr. 1977.

[TSW12] J. Thompson, B. Smith, A. Warner, and J. M. Jot, “Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations,” in *Proc. 133rd AES Conv.*, San Francisco, 2012.

[UhH15] C. Uhle, and E. A. P. Habets, “Direct-ambient decomposition using parametric wiener filtering with spatial cue control,” in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 36-40.

[UhP08] C. Uhle, and C. Paul, “A Supervised Learning Approach to Ambience Extraction,” in *Proc. DAFx*, Espoo, Finland, 2008.

[UoS15] University of Southampton, UK. Head Related Transfer Functions (HRTFs), Jun. 10, 2015, Available: http://www.southampton.ac.uk/engineering/research/groups/fluid_dynamics/electroacoustics/nmh_hrtf.page

[UsB07] J. Usher and J. Benesty, “Enhancement of spatial sound quality: a new reverberation-extraction audio upmixer,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2141-2150, Sept. 2007.

[UWH07] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, “Ambience separation from mono recordings using non-negative matrix factorization,” in *Proc. 30th AES Int. Conf.*, Saariselka, Finland, 2007.

[VaB88] B. D. Van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[VBG14] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, “From blind to guided audio source separation,” *IEEE Sig. Process. Mag.*, vol. 31, no. 3, pp. 107-115, May 2014.

[Vas05] P. N. Vassilakis, Introduction to Psychoacoustics, Columbia College lecture notes, http://acousticslab.org/psychoacoustics/PMFiles/Module07a.htm, 2015

[VFR15] V. Välimäki, A. Franck, J. Rämö, H. Gamper, and L. Savioja, “Assisted listening using a headset,” *IEEE Sig. Process. Mag.*, vol. 32, no. 2, pp. 92-99, Mar. 2015.

[VGF06] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[Vin13] E. Vincent, “MUSHRAM: a MATLAB interface for MUSHRA listening tests, 2005.” Available: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/. [Apr. 05, 2013].

[Vir06] T. Virtanen, “Sound source separation in monaural music signals,” PhD Thesis, Tampere University of Technology, 2006.

[VPS12] Vesa Välimäki, Julian D. Parker, Lauri Savioja, Julius O. Smith, and Jonathan S. Abel, “Fifty years of artificial reverberation,” *IEEE Trans. Audio, Speech, Lang. Process.*,vol. 20, no. 5, pp. 1421-1448, Jul. 2012.

[WaB06] D. L. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley/IEEE Press, Hoboken, NJ, 2006.

[WaF11] A. Walther, and C. Faller, “Direct-ambient decomposition and upmix of surround signals,” in *Proc. IWASPAA*, New Paltz, NY, Oct. 2011, pp. 277-280.

[WAK93] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization Using Non-individualized Head-Related Transfer Functions,” *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111-123, Jan. 1993.

[WHO06] World Health Organization. “Primary ear and hearing care training resource,” Switzerland, 2006.

[Wik13] WIKIPEDIA. (2013, August 23). Stereophonic sound [Online]. Available: http://en.wikipedia.org/wiki/Stereophonic_sound.

[WoW12] J. Woodruff and D. L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503-1512, Jul. 2012.

[WSL06] S. Wang, D. Sen, and W. Lu, “Subband analysis of time delay estimation in STFT domain,” in *Proc. 11th Aust. Intl. Conf. on Speech Science & Technology*, New Zealand, 2006.

[Xie13] B. S. Xie, *Head-related transfer function and virtual auditory display, 2$^{nd}$ edition.* J. Ross Publishing, US, 2003.

[XLS07] S. Xu, Z. Li, and G. Salvendy, “Individualization of head-related transfer function for three-dimensional virtual auditory display: a review,” in R. Shumaker (Ed.): Virtual Reality, HCII 2007, LNCS 4563, pp. 397–407, 2007.

[YiR04] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Processing,* vol. 52, no.7, pp. 1830-1847, Jul. 2004.

[Yos93] W. A. Yost, “Perceptual models for auditory localization,” in *Proc. 12th Audio Eng. Soc. Int. Conf.*, Copenhagen, Denmark, 1993.

[Zar02a] P. Zahorik, “Assessing auditory distance perception using virtual acoustics,” *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1832-1846, Apr. 2002.

[Zar02b] P. Zahorik, “Direct-to-reverberant energy ratio sensitivity,” *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2110-2117, May 2002.

[ZHD03] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, “HRTF personalization using anthropometric measurements,” in *Proc. IWASPAA*, New York, pp. 157-160, Oct. 2003.

[ZiR03] S. K. Zielinski, and F. Rumsey, “Effects of down-mix algorithms on quality of surround sound,” *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 780–798, Sept. 2003.

[Zwi61] E. Zwicker, “Subdivision of the Audible Frequency Range into Critical Bands,” *J. Acoust. Soc. Am.*, vol. 33, no. 2, pp. 248, 1961.

# Appendix A: Derivation of Simplied Solution for PCA based PAE

In the following, we show the derivations for the extracted primary component in channel 0. From (3.8) and (3.20), we can find

$$\frac{\lambda_{\mathrm{P}} - r_{00}}{r_{01}} = k. \tag{A.1}$$

From (3.5)-(3.7), we have

$$r_{11} = \frac{k^2 - 1}{k} r_{01} + r_{00}. \tag{A.2}$$

Based on (A.1), we can rewrite (3.21) as

$$\mathbf{u}_{\mathrm{P}} = r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right). \tag{A.3}$$

Substitute (A.3) into (3.22),

$$\begin{aligned}\hat{\mathbf{p}}_0 &= \frac{\mathbf{u}_{\mathrm{P}}{}^H \mathbf{x}_0}{\mathbf{u}_{\mathrm{P}}{}^H \mathbf{u}_{\mathrm{P}}} \mathbf{u}_{\mathrm{P}} \\ &= \frac{r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right)^H \mathbf{x}_0}{r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right)^H r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right)} r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right) \\ &= \frac{r_{01}\left(r_{00} + kr_{01}\right)}{r_{01}{}^2\left(r_{00} + k^2 r_{11} + 2kr_{01}\right)} r_{01}\left(\mathbf{x}_0 + k\mathbf{x}_1\right) \\ &= \frac{r_{00} + kr_{01}}{r_{00} + k^2 r_{11} + 2kr_{01}}\left(\mathbf{x}_0 + k\mathbf{x}_1\right).\end{aligned} \tag{A.4}$$

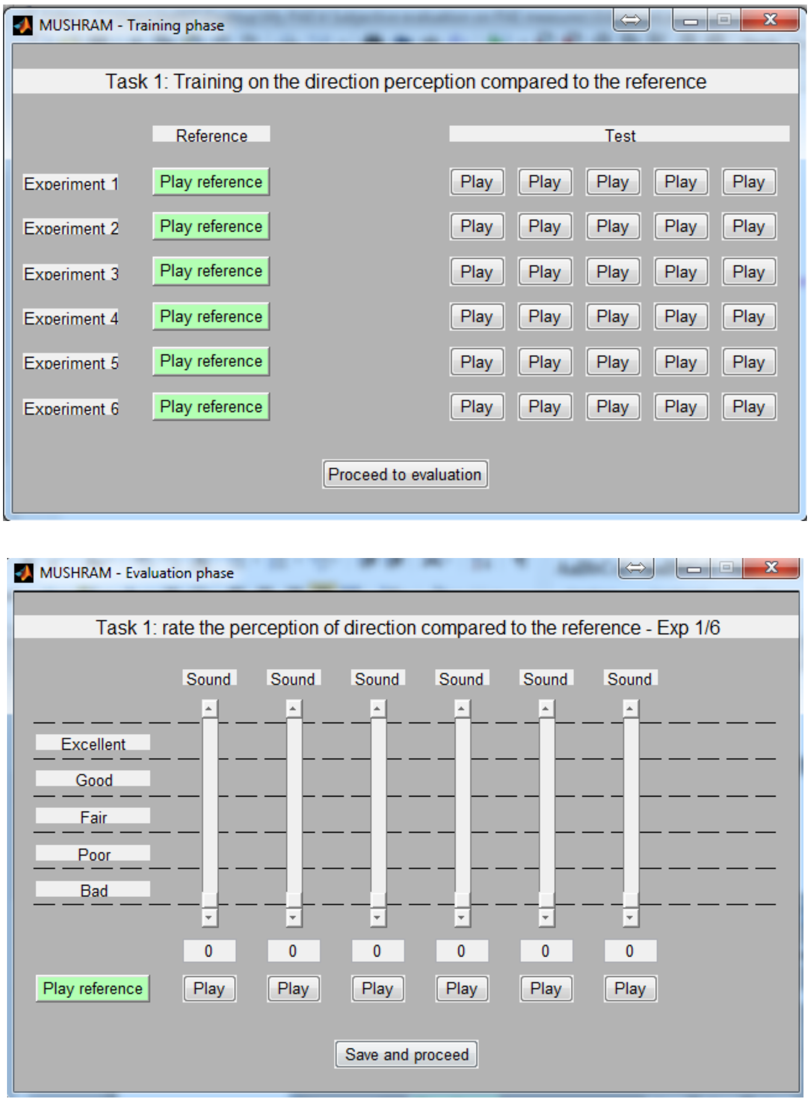
Substitute (A.2) into (A.4),

$$
\begin{aligned}
\hat{\mathbf{p}}_0 &= \frac{r_{00}+kr_{01}}{r_{00}+k^2\left(\frac{k^2-1}{k}r_{01}+r_{00}\right)+2kr_{01}}\left(\mathbf{x}_0+k\mathbf{x}_1\right)\\
&= \frac{r_{00}+kr_{01}}{\left(1+k^2\right)r_{00}+k\left(k^2+1\right)r_{01}}\left(\mathbf{x}_0+k\mathbf{x}_1\right)\\
&= \frac{1}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right).
\end{aligned}
\tag{A.5}
$$

Thus, we obtain the simplified expression of the extracted primary component in channel 0, as shown in (3.23). The primary component in channel 1 and ambient components can also be derived in the same way.

# Appendix B: Subjective Listening Tests for PAE

In this appendix, we show the MATLAB GUI screenshots (Fig B.1) and written guidelines that were presented to the participants of the subjective listening tests conducted to evaluate the perceptual performance of different PAE approaches. Some results of the listening tests are reported in Chapter 4.



MUSHRAM - Training phase

Task 1: Training on the direction perception compared to the reference

MUSHRAM - Evaluation phase

Task 1: rate the perception of direction compared to the reference - Exp 1/6

Figure B.1Two screenshots of the MATLAB GUI

## Guidelines for listening tests (timbre quality)

This listening test aims to rate the quality of a set of signals produced by primary-ambient extraction systems. Primary-ambient extraction aims to extract the primary components and ambient components from a mixture of them. The resulting signals may include several types of degradations compared to the clean target signals, which mainly includes distortions of the target signal, and residual leakage from other undesired signals.

During the test, you will be asked to address *3 successive tasks* (6 sub-tasks):

1. Rate the **global quality** compared to the reference for each test signal (tasks 1-2);
2. Rate the quality in terms of **preservation of the target signal** in each test signal (tasks 3-4);
3. Rate the quality in terms of **suppression of other undesired signals** in each test signal (tasks 5-6).

For each task, the test will involve a *training phase* and an *evaluation phase*.

During the training phase, you will have to listen to all the sounds to

- train yourself to address the required task and learn the range of observed quality according to that task;
- set the volume of your headphones so that it's comfortable but you can clearly hear differences between sounds (the volume can't be changed later on).

The evaluation phase involves 6-8 trials. In each trial, you will have to rate the quality of 6 test sounds compared to a reference sound (clean target) on a scale from 0 to100, where larger ratings indicate better quality. You can listen to the sounds as many times as you want. You should make sure that

- **the ratings between pairs of sounds are consistent**, *i.e.* if one sound has better quality than another, it should be rated higher,
- **the ratings between different experiments are consistent**, *i.e.* if two sounds from different experiments have the same quality, they should be rated equally,
- **the whole rating scale is used**, *i.e.* sounds with perfect quality (as compared to the reference signal) should be rated 100 and the worst test sound over all experiments (but not necessarily the worst test sound in each experiment) should be rated 0.

The expected total duration of the test is 30 minutes. You can make short breaks between each two trials. Feel free to contact the experimenter whenever you have any doubts or need any assistance. Thank you for your participation.

**Guidelines for listening tests (spatial quality)**

This listening test aims to rate the quality of a set of signals produced by primary-ambient extraction systems. Primary-ambient extraction aims to extract the primary components and ambient components from a mixture of them. The resulting signals may include inaccurate spatial perception as compared to the clean target signals.

During the test, you will be asked to address *3 successive tasks according to these two criteria*:

4. Rate the quality in terms of the **perception of direction (able to localize the sound from the same direction)** as compared to the reference signal (task 1);
5. Rate the quality in terms of **perception of diffuseness (feeling of the sound coming from any directions)** as compared to the reference signal (tasks 2-3);

For each task, the test will involve a *training phase* and an *evaluation phase*.

During the training phase, you will have to listen to all the sounds to

- train yourself to address the required task and learn the range of observed quality according to that task;
- set the volume of your headphones so that it's comfortable but you can clearly hear differences between sounds (the volume can't be changed later on).

The evaluation phase involves 6 experiments. In each experiment, you will have to rate the quality of 6-8 test sounds compared to a reference sound (clean target) on a scale from 0 to100, where larger ratings indicate better quality. You can listen to the sounds as many times as you want. You should make sure that

- **the ratings between pairs of sounds are consistent**, *i.e.* if one sound has better quality than another, it should be rated higher,
- **the ratings between different experiments are consistent**, *i.e.* if two sounds from different experiments have the same quality, they should be rated equally,
- **the whole rating scale is used**, *i.e.* sounds with perfect quality (as compared to the reference signal) should be rated 100 and the worst test sound over all experiments (but not necessarily the worst test sound in each experiment) should be rated 0.

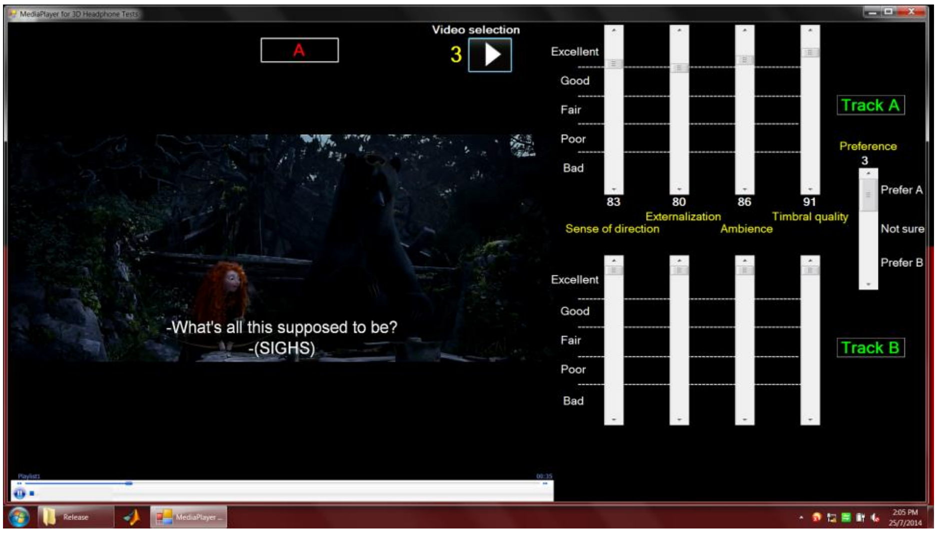
The expected total duration of the test is 15 minutes. You can make short breaks between each two experiments. Feel free to contact the experimenter whenever you have any doubts or need any assistance. Thank you for your participation.

# Appendix C: Subjective Listening Tests for Natural Sound Rendering Headphone System

In this appendix, we show the GUI screenshot (Fig. C.1) and written guidelines presented to the participants of the subjective listening tests conducted to evaluate the two headphone rendering systems as discussed in Chapter. Table B.1 shows the specifications of the five audio/video tracks used in this experiment. The results of the listening tests are reported in Chapter 7.

Table C.1 Specifications of five audio/video tracks

| # | Type | Track | Duration |
|---|---|---|---|
| 1 | Binaural audio track | Motorcycle in the storm | 1:07 |
| 2 | | Bee at the waterfall | 0:20 |
| 3 | Movie video track | Brave | 2:59 |
| 4 | | Prometheus | 2:24 |
| 5 | Gaming video track | Batterfield 3 | 1:49 |



Video selection 3

Sense of direction 83

Externalization 80

Ambience 86

Timbral quality 91

Preference 3

-What's all this supposed to be?
-(SIGHS)

Figure C.1 A screenshot of the GUI designed for the headphone listening test

### Instructions

Thank you for participating in this experiment.

There are 5 sessions. In each session, there is a video with two sound tracks (A, B). Listen to A and B and give your scores based on the each of the following 4 criteria:

1. **Sense of direction**: how clear can you perceive the directions of the sound objects?

2. **Externalization**: how clear can you perceive the sound coming from outside your head?

3. **Ambience**: how clear can you perceive the ambience of the sound environment?

4. **Timbral quality**: how realistic is the sound?

Finally, please give your **preference** between A and B.

You can switch between A and B anytime, and give the score before the track ends. There is a number indication once you give the score.

In case of the video and sound not sync, click the **stop** button of the player and then click the **play** button.

Feel free to contact the experimenter if you have any doubts. Thanks.