3D Sound Effect Analysis, Synthesis and Application Design
- A Primary-Ambient Extraction Approach

Abstract

In spatial audio analysis-synthesis, one of the key issues is to decompose a signal into primary and ambient components based on their spatial features. Stereo audio signals are often modeled as a linear mixture of primary and ambient components. Existing approaches like principal component analysis (PCA) and least squares (LS) have been widely employed to extract primary and ambient components from stereo signals. However, the performance and comparisons of these approaches in primary-ambient extraction (PAE) have not been well studied. In this report, we show that existing approaches can be generalized into a linear estimation framework. Under this framework, we propose a systematic series of performance measures identifying the components comprising the extraction error. Based on the linear estimation framework and the performance measures, we present a comparative study of the linear estimation based PAE approaches including existing PCA, LS, and two proposed variant LS approaches for more practical objectives in their performance. Experimental results are provided to justify the relationships and differences of these approaches.

However, the performance of PCA based primary ambient extraction (PAE) is highly dependent on the assumptions of the input signal model, where the primary components in the stereo signal are assumed to be completely correlated at zero lag. One of the most frequently encountered cases where the primary component is partially correlated, namely the primary-complex case, is not well-studied. To alleviate the performance degradation in this case, the time-shifted PCA based PAE is proposed in this work. This approach involves time-shifting the input signal according to the estimated inter-channel time difference (ICTD) of the input signal prior to the linear estimation in PAE. Based on the results from our simulation and informal listening tests, the shifted PAE approach is found to be superior to the conventional PCA based PAE methods.

From our study, we find that the existing approaches are still unable to handle more complicated cases of the input signals. For example, the stereo signal model requires the primary components to come from one direction and the primary and ambient components are only characterized by their inter-channel correlations. These remaining problems stimulate our further study on formulating more realistic signal models as well as the classification of the input signal into a specific model. With the proper extraction of primary and ambient components as well as the appropriate post-processing techniques, a more immersive 3D audio experience can be achieved.

This page is left intentionally blank

Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 3D audio

With the increasing prevalence of 3D video technology, consumers are demanding a more immersive listening experience to better match the 3D visual effects. This results in a growing need for a better spatial audio reproduction, which can create the feeling of *being there* at the scene where the sound events take place. 3D audio or spatial audio refers to the perception of soundscape in a three-dimension space [1], as what we can perceive in our daily life. A number of audio cues are found to be useful to accurately localize sound. Two most important cues are interaural time differences (ITDs) and interaural level differences (ILDs), which account for the time differences and level differences as sound propagates to our left and right ears [2]. However, identical ITD and ILD will incur the cone of confusion [3], which is the major cause of front-back ambiguity as well as elevation ambiguity. To address these issues, spectral cues based on the response of our pinnae are used, as the characterization of the human ears is given by the head-related transfer function (HRTF) [4]. HRTF has been widely applied in 3D audio systems, and can be directly implemented by digital filtering [1]. Other cues including head movement, visual cues, and sound familiarity also play influential roles in sound localization. 3D audio systems can be implemented using loudspeakers or headphones, but headphones are generally preferred as there is no crosstalk. Using proper crosstalk cancellation, a richer and more natural soundscape can be created in loudspeaker based 3D audio systems. However, such loudspeaker systems are unable to produce a very accurate auditory images especially when number of loudspeakers is limited [5], [6]. On the other hand, headphone based 3D audio system is more suitable for personal listening. Binaural techniques with non-individualized HRTF are normally used in producing 3D audio for headphones [7], but the main drawbacks of such approaches are the lack of externalization and relatively high front-back confusions [8]. Recent study in [9] shows that better localization with reduced front-back confusion can be achieved by introducing the frontal emitters in headphones.

1.2 Primary-ambient extraction for spatial audio

Audio signals generally consist of directional and diffuse components, which are referred as primary and ambient components, respectively. These two elements are critical in creating an immersive 3D audio experience. While ambient component recreates the sound environment of the auditory scene, the primary component, on the other hand, allows the listener to localize the sound of interest. An example of such a scenario from the movie *Life of PI* is given in Fig. 1.1. Immersive 3D audio experience aims to create the experience as if the viewer is placed at the same scene as the actor, and the listener expects to experience the ambience of the sea and a tiger in close proximity.

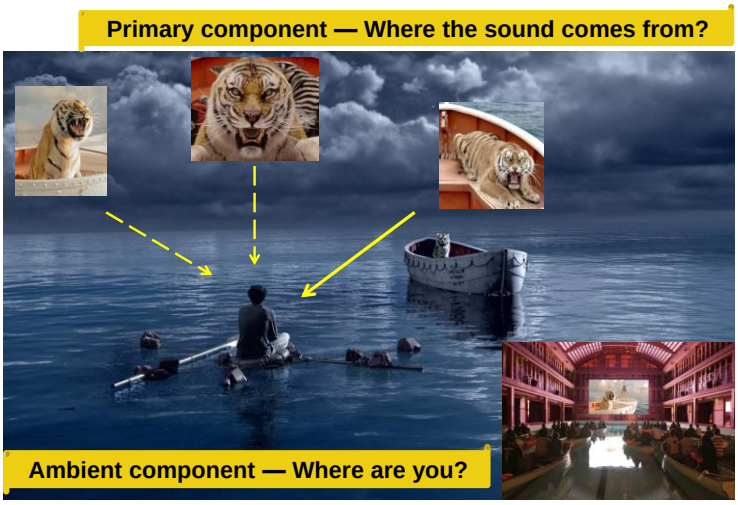

Figure 1.1 Primary and ambient components in spatial audio.

To achieve accurate spatial audio rendering, different processing schemes should be applied on the primary and ambient components of the audio signals [10], [11]. However, the primary and ambient components are not separately stored in conventional audio formats, such as stereo and surround sound. Such audio formats necessitate primary-ambient extraction (PAE).

Figure 1.2 shows a general block diagram of PAE in spatial audio processing. Pre-processing techniques such as digital Fourier transform (DFT) are applied before PAE. Next, the extracted primary and ambient components are rendered separately, based on the associated spatial attributes. For example, primary components can be processed by HRTF filtering of appropriate azimuth and elevation, while ambient components can be rendered using binaural room impulse response [11]. Finally, the rendered primary and ambient components can be re-mixed or sent to playback systems with different configurations. Recent years have seen applications of PAE in spatial audio processing [5], [11]-[13] spatial audio coding [14], [15], audio mixing [16]-[18], and immersive 3D sound system [6], [19]-[20]. An overall diagram view of the application and related research areas of PAE is shown in Fig. 1.3.

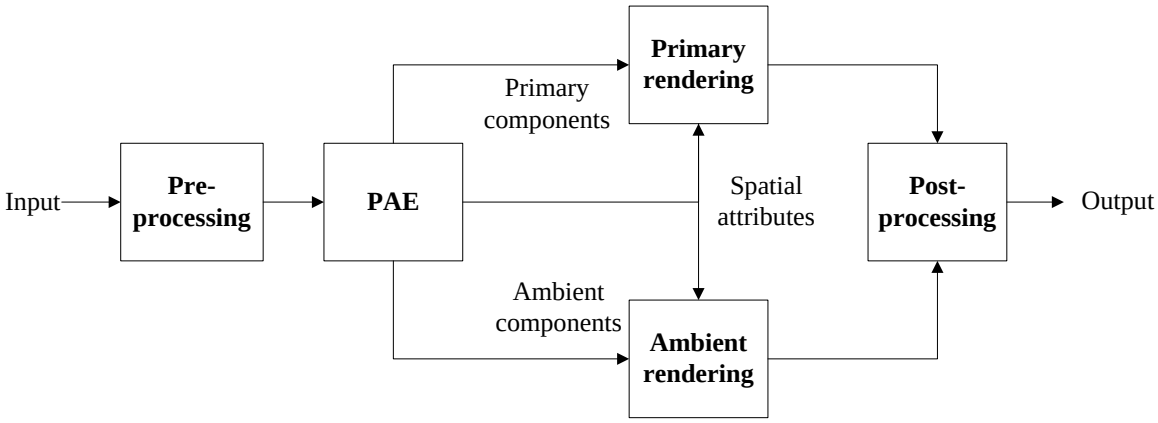

Figure 1.2 Block diagram of PAE in spatial audio processing.

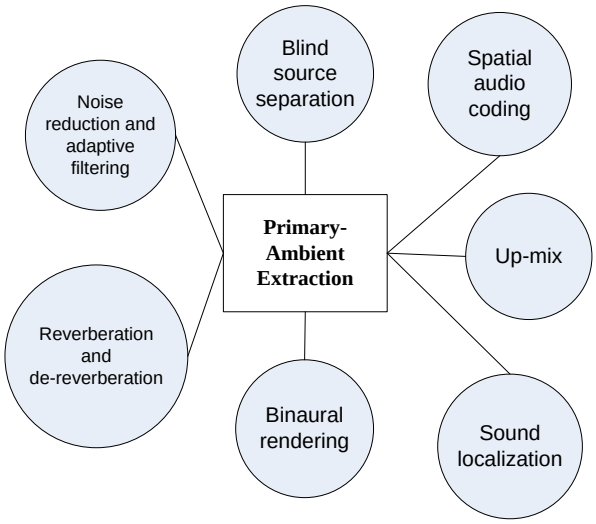

Figure 1.3 Overview of the application and related research areas of primary-ambient extraction.

In spatial audio coding, there are two widely studied frameworks, namely, spatial audio scene coding (SASC) [15], [21] and directional audio coding (DirAC) [14]. In SASC, primary-ambient decomposition is performed on multichannel audio input. The localization analysis for the decomposed primary and ambient components is carried out separately using Gerzon localization vector [22] and the spatial cues are applied in the final synthesis. On the other hand, DirAC aims to reproduce spatial sound using any sound system configurations. One of the most essential stages of DirAC is to divide the input audio signal into diffuse (ambient) and non-diffuse (primary) sound. The primary sound is then reproduced using vector base amplitude panning (VBAP)

[23], while the ambient sound is usually decorrelated to create a better surround sound effect.

Various up-mix techniques based on PAE have been discussed in [17], [18], [24]. The PAE based up-mix approach is extremely suitable for immersive 3-D (i3D) sound system proposed by Gan *et al*. [19], [20]. The i3D sound system comprises of a unique combination of conventional and parametric loudspeakers, so as to exploit the high directivity of the parametric loudspeaker to render sharp images of the primary component, and to reproduce the spaciousness of the ambience using the conventional loudspeaker. By accurately reproducing the primary and ambient components of the spatial sound, the i3D sound system creates an immersive and realistic soundscape for gaming and home entertainment applications [6].

Primary-ambient extraction can also be related to other research problems in audio signal processing. Similar to blind source separation (BSS), PAE also deals with the decomposition of significantly different components from mixed signals. The key difference between BSS and PAE is the characteristics of the separated components: BSS separates the independent components [25], while PAE extracts components with different perceptual features. One of the most important features of PAE is the extraction of accurate spatial auditory image from the input signal, which is in line with the objectives in sound localization problems. By considering the primary component as the direct sound and the ambient component as the reverberated sound, PAE can be applied to extract reverberation [26]. PAE can also be applied to noise reduction and adaptive filtering operations if the extraction of the primary component from a noise-like ambience is considered [27].

To date, many approaches have been proposed for PAE, especially for stereo signals. For these approaches, the stereo signal is generally modeled as a directional sound source mixed with uncorrelated ambience. Figure 1.4 shows the extraction of the primary and ambient components from a stereo signal using PAE. In [24], a single-channel time-frequency mask was used to extract the ambience from a stereo signal with the assumptions of equal level of ambience and equal ratio of ambience to primary components in the two channels of the stereo signal. Principal component analysis (PCA) remains one of the most widely studied approaches applied in PAE [6], [10], [28]-[35]. Taking into consideration of the independence between primary and ambient components, the stereo signal is decomposed into two orthogonal bases using the Karhunen-Loève transform [36]. Based on the assumption that the primary component is relatively stronger than ambience, the projected signal on the basis vector with larger variance is assumed to be the primary component, and the projected signal on the other basis vector is assumed as the ambience.

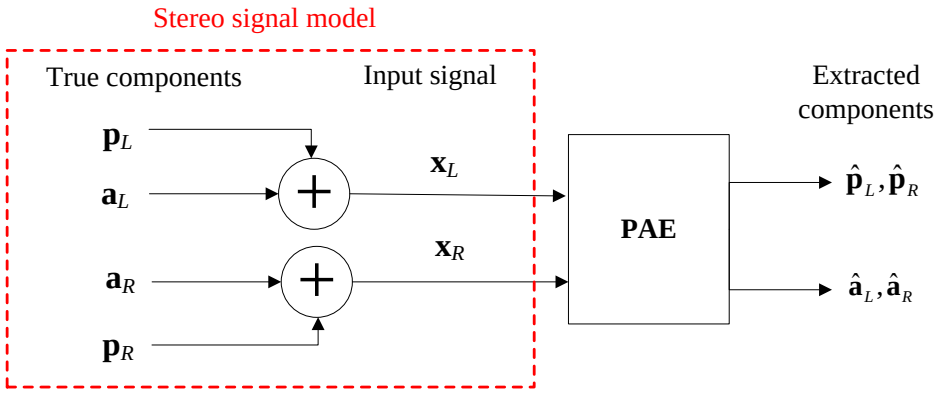

Figure 1.4 Extraction of primary and ambient components using PAE, where $\mathbf{x}_L, \mathbf{x}_R$ are the input stereo signals in the left and right channel; $\mathbf{p}_L, \mathbf{p}_R$ and $\mathbf{a}_L, \mathbf{a}_R$ are the true primary and ambient components; $\hat{\mathbf{p}}_L, \hat{\mathbf{p}}_R$ and $\hat{\mathbf{a}}_L, \hat{\mathbf{a}}_R$ are the extracted primary and ambient components.

In [37], Faller introduced a least-squares (LS) approach to estimate the primary and ambient components for surround sound up-mixing. For stereo signal, we show that both PCA and LS originated from linear estimation approaches. Experimental studies in [29] show that PCA and LS based PAE produce superior extraction results than single-channel time-frequency masks, especially in extracting the primary components. However, the relationship and differences between PCA and LS based PAE remain unclear, hence further studies are required. Other techniques like factor analysis [38] and independent component analysis [25] are also applied in PAE. In addition to stereo signals, PAE can also be extended to deal with multichannel signals [39].

Motivated by the fact that the primary and ambient components in stereo signal model are usually linearly mixed [10], the problem of PAE for stereo audio signals can be solved using linear estimation based approaches. For the signal model, the extracted primary or ambient component is expressed as a weighted sum of the left and right channels of the stereo signal. Using this paradigm, we derive the optimal weightings for four approaches, namely, PCA, LS, minimum leakage LS (MLLS), and minimum distortion LS (MDLS). This is followed by a comparative study on the extraction results and performance evaluation of these approaches.

On the other hand, we found that the performance of these linear estimation based PAE approaches (PCA is used as an example) is severely degraded in complex cases when the assumptions for the signal model are unmet. For example, when the primary component is not completely correlated at zero lag, significant error can be found in the extracted primary component, as well as inter-channel time difference (ICTD) and inter-channel level difference (ICLD) of the extracted primary component [35]. The differences in the ICTD and ICLD cues between the extracted primary component and true primary component can lead to erroneous sound localization. To address this problem in ambient extraction, a normalized least-mean-square approach was proposed in [26]. Our analysis in [35] argues that the degraded performance is caused by the ignored ICTD and the resulting low correlation of the primary components. Therefore, we applied a time-shifting operation prior to the PCA decomposition (or other approaches) in PAE to increase the correlation of primary component of the stereo signal.

The rest of this report is organized as follows. In Chapter 2, we review the stereo signal model, and the key assumptions of the stereo signal model. Subsequently, a linear estimation based framework on PAE and three groups of performance measures [29] are presented in Chapter 3. Chapter 4 discusses several approaches applied in PAE. In Chapter 5, the performance degradation of PAE using PCA (as an example) in primary-complex case is first discussed. To alleviate the problem, shifted PCA based PAE is introduced. Theoretical analysis in these two chapters is validated using experiments, respectively. Finally, we suggest a few possible future works for PAE in Chapter 6 and conclude this work in Chapter 7.

1.3 Contributions

The main original contributions of this report are:

1. Based on the current PCA and LS approaches, the linear estimation framework for PAE is formulated.

2. An objective evaluation system with three groups of measures to quantify the extraction performance of PAE is introduced.

3. Minimum leakage LS and minimum distortion LS have been derived in PAE, and a comprehensive comparison among PCA, LS, and the proposed approaches is performed.

4. The remaining problems of current PAE approaches when applied in more practical cases are investigated. For the most common primary-complex case, the shifted PCA based PAE is introduced and further verified.

2. STEREO SIGNAL MODEL

In this chapter, we first introduce the basic stereo signal model and its key assumptions. In general, a stereo signal model [10] consists of two parts: (i) a directional component referred as the primary component; and (ii) a diffused component referred as the ambient component. Denoting the time-domain left and right input signals as $\mathbf{x}_{\mathrm{L}}, \mathbf{x}_{\mathrm{R}}$, we can express the stereo signal model as:

$$\mathbf{x}_{\mathrm{L}} = \mathbf{p}_{\mathrm{L}} + \mathbf{a}_{\mathrm{L}},\ \mathbf{x}_{\mathrm{R}} = \mathbf{p}_{\mathrm{R}} + \mathbf{a}_{\mathrm{R}}. \tag{1}$$

This stereo signal model assumes that the primary and ambient components in the left and right channels are correlated and uncorrelated, respectively. The correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as follows:

$$\phi_{ij}(\tau) = r_{ij}(\tau) \Big/ \sqrt{r_{ii}(\tau) r_{jj}(\tau)}, \tag{2}$$

where $r_{ij}(\tau)$ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ at lag τ. Two signals are considered correlated when $\max\left|\phi_{ij}(\tau)\right| = 1$; uncorrelated when $\max\left|\phi_{ij}(\tau)\right| = 0$; and partially correlated when $0 < \max\left|\phi_{ij}(\tau)\right| < 1$.

Correlated primary component satisfies one of the following conditions [2]: (i) amplitude panned, i.e., $\mathbf{p}_{\mathrm{R}} = k\mathbf{p}_{\mathrm{L}}$, where k is the primary panning factor (PPF); (ii) time shifted, i.e., $p_{\mathrm{R}}(n) = p_{\mathrm{L}}(n+\tau)$, where $p_{\mathrm{R}}(n)$ is the nth sample in $\mathbf{p}_{\mathrm{R}}$ and τ is the inter-channel time difference (ICTD); and (iii) amplitude panned and time shifted, i.e., $p_{\mathrm{R}}(n) = kp_{\mathrm{L}}(n+\tau)$. In [10], the correlated primary component is assumed to be amplitude panned and uncorrelated with ambient component. Considering the diffuseness of ambience, power of the ambience is relatively balanced in the two channels of the stereo signal. To quantify the power difference between the primary and ambient components, we introduce the primary power ratio (PPR) γ, which is defined as the ratio of total primary power to total signal power in two channels:

$$\gamma = P_{\mathrm{pt}} \Big/ \left(P_{\mathrm{pt}} + P_{\mathrm{at}}\right), \tag{3}$$

where P_{pt} and P_{at} denote the total power of the primary and ambient components in two channels, respectively. From (3), it is clear that PPR ranges from zero to one. Summarizing the assumptions for the stereo signal model, we have

$$\mathbf{p}_{\mathrm{R}} = k\mathbf{p}_{\mathrm{L}}, \mathbf{a}_{\mathrm{L}} \perp \mathbf{a}_{\mathrm{R}}, \mathbf{p}_{\mathrm{L}}(\mathbf{p}_{\mathrm{R}}) \perp \mathbf{p}_{\mathrm{L}}(\mathbf{p}_{\mathrm{R}}), \tag{4}$$

$$P_{\mathbf{p}_{\mathrm{R}}} = k^2 P_{\mathbf{p}_{\mathrm{L}}} = k^2 P_{\mathrm{p}},\ P_{\mathbf{a}_{\mathrm{L}}} = P_{\mathbf{a}_{\mathrm{R}}} = P_{\mathrm{a}}, \tag{5}$$

where $P_{(.)}$ denotes the signal power and $\perp$ represents two signals are uncorrelated. Given any stereo input signal that fulfills the above conditions, we can relate the auto-correlation and cross-correlation at zero-lag to the power of these components as

$$r_{\mathrm{LL}} = \mathbf{x}_{\mathrm{L}}{}^{T}\mathbf{x}_{\mathrm{L}} = NP_{\mathbf{x}_{\mathrm{L}}} = N\left(P_{\mathbf{p}_{\mathrm{L}}} + P_{\mathbf{a}_{\mathrm{L}}}\right), \tag{6}$$

$$r_{\mathrm{RR}} = \mathbf{x}_{\mathrm{R}}{}^{T}\mathbf{x}_{\mathrm{R}} = NP_{\mathbf{x}_{\mathrm{R}}} = N\left(k^2 P_{\mathbf{p}_{\mathrm{L}}} + P_{\mathbf{a}_{\mathrm{L}}}\right), \tag{7}$$

$$r_{\mathrm{LR}} = \mathbf{x}_{\mathrm{L}}{}^{T}\mathbf{x}_{\mathrm{R}} = \mathbf{p}_{\mathrm{L}}{}^{T}\mathbf{p}_{\mathrm{R}} = NkP_{\mathbf{p}_{\mathrm{L}}}, \tag{8}$$

where T is the transpose operator and N is the number of samples in one frame of the stereo signal. From (6)-(8), the PPF and PPR of the stereo signal model are

$$k = \frac{r_{\mathrm{RR}} - r_{\mathrm{LL}}}{2r_{\mathrm{LR}}} + \sqrt{\left(\frac{r_{\mathrm{RR}} - r_{\mathrm{LL}}}{2r_{\mathrm{LR}}}\right)^2 + 1}, \tag{9}$$

$$\gamma = \frac{2r_{\mathrm{LR}} + \left(r_{\mathrm{RR}} - r_{\mathrm{LL}}\right)k}{\left(r_{\mathrm{RR}} + r_{\mathrm{LL}}\right)k}. \tag{10}$$

The primary component is panned to the right channel for $k > 1$ and to the left channel for $k < 1$. In spatial audio, the PPF is closely related to inter-channel level difference (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 γ increases. In following chapters, we shall see that PPF and PPR are useful parameters to compute the extraction results of primary and ambient components, as well as to evaluate the performance of the approaches in PAE.

3. LINEAR ESTIMATION MODEL AND PERFORMANCE MEASURES

In this chapter, we first investigate the linear estimation model applied in PAE. Three groups of performance measures are also introduced to provide a comprehensive evaluation of the extraction performance of various PAE approaches.

3.1 Linear estimation model in PAE

In this report, we examine the blind extraction of primary and ambient components from stereo input signal. Inspired by the mixing signal model given in (1), we address the PAE problem using a linear estimation model, where the primary and ambient components in both channels are estimated as a weighted sum of the two channels of the stereo signal. The primary and ambient components are expressed as

$$\begin{aligned} \hat{\mathbf{p}}_{\mathrm{L}} &= w_{\mathrm{LpL}}\mathbf{x}_{\mathrm{L}} + w_{\mathrm{LpR}}\mathbf{x}_{\mathrm{R}},\ \hat{\mathbf{p}}_{\mathrm{R}} = w_{\mathrm{RpL}}\mathbf{x}_{\mathrm{L}} + w_{\mathrm{RpR}}\mathbf{x}_{\mathrm{R}}, \\ \hat{\mathbf{a}}_{\mathrm{L}} &= w_{\mathrm{LaL}}\mathbf{x}_{\mathrm{L}} + w_{\mathrm{LaR}}\mathbf{x}_{\mathrm{R}},\ \hat{\mathbf{a}}_{\mathrm{R}} = w_{\mathrm{RaL}}\mathbf{x}_{\mathrm{L}} + w_{\mathrm{RaR}}\mathbf{x}_{\mathrm{R}}, \end{aligned} \tag{11}$$

where $\hat{\mathbf{p}}_{\mathrm{L}}, \hat{\mathbf{p}}_{\mathrm{R}}$ and $\hat{\mathbf{a}}_{\mathrm{L}}, \hat{\mathbf{a}}_{\mathrm{R}}$ are the estimated primary and ambient components in the left and right channels, respectively; and $w_{(.)}$ is the estimated weight of the extracted component in a specific channel for a specific estimated component, where the first, middle, and last letter of the subscript denote the left ("L") or right ("R") channel of the extracted component, the primary ("*p*") or ambient ("*a*") component being extracted, and the channel of the input signal, respectively. By rewriting (11) in matrix form, we have

$$\begin{bmatrix} \hat{\mathbf{p}}_{\mathrm{L}}{}^{T} \\ \hat{\mathbf{p}}_{\mathrm{R}}{}^{T} \\ \hat{\mathbf{a}}_{\mathrm{L}}{}^{T} \\ \hat{\mathbf{a}}_{\mathrm{R}}{}^{T} \end{bmatrix} = \begin{bmatrix} w_{\mathrm{LpL}} & w_{\mathrm{LpR}} \\ w_{\mathrm{RpL}} & w_{\mathrm{RpR}} \\ w_{\mathrm{LaL}} & w_{\mathrm{LaR}} \\ w_{\mathrm{RaL}} & w_{\mathrm{RaR}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\mathrm{L}}{}^{T} \\ \mathbf{x}_{\mathrm{R}}{}^{T} \end{bmatrix} = \mathbf{W} \begin{bmatrix} \mathbf{x}_{\mathrm{L}}{}^{T} \\ \mathbf{x}_{\mathrm{R}}{}^{T} \end{bmatrix}. \tag{12}$$

With this formulation, the PAE problem now becomes the estimation of weighting matrix **W**, where the first and last two rows of **W** are the extraction weights for the primary and ambient components, respectively. The weights in **W** are crucial to the extraction performance. Hence, before deriving the optimal weights, we first introduce three groups of measures to evaluate the performance of the linear estimation based PAE approaches. The first two groups measure the extraction accuracy of the extracted components whereas the third group quantifies the accuracy of the spatial cues in two channels.

3.2 Performance measures for primary components

In the first group of measures, we shall introduce a few performance scores to quantify the extraction error. By considering the widely used error measure between the extracted primary component in channel *i* ("*L*" or "*R*") $\hat{\mathbf{p}}_i$ and the true component $\mathbf{p}_i$, we have

$$\boldsymbol{\varepsilon} = \hat{\mathbf{p}}_i - \mathbf{p}_i. \tag{13}$$

Based on (13), we compute the error-to-signal ratio (ESR), which is defined as the ratio between the power of extraction error and the power of the true primary component:

$$\mathrm{ESR}_{\mathrm{P}i} = P_{\boldsymbol{\varepsilon}} / P_{\mathbf{p}_i}. \tag{14}$$

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

Based on (11), the general form of $\hat{\mathbf{p}}_i$ can be expressed as

$$\hat{\mathbf{p}}_i = w_{i\mathrm{pL}}\mathbf{x}_{\mathrm{L}} + w_{i\mathrm{pR}}\mathbf{x}_{\mathrm{R}}, \tag{15}$$

where $w_{i\mathrm{pL}}$ and $w_{i\mathrm{pR}}$ are the elements in the first and second row of **W**, respectively, depending on the channel represented by *i*. Based on the assumptions stated in (4) and substituting (1) into (15), we have

$$\begin{aligned} \hat{\mathbf{p}}_i &= \left(w_{i\mathrm{pL}}\mathbf{p}_{\mathrm{L}} + w_{i\mathrm{pR}}\mathbf{p}_{\mathrm{R}}\right) + \left(w_{i\mathrm{pL}}\mathbf{a}_{\mathrm{L}} + w_{i\mathrm{pR}}\mathbf{a}_{\mathrm{R}}\right) \\ &= w_{i\mathrm{p}}\mathbf{p}_i + \left(w_{i\mathrm{pL}}\mathbf{a}_{\mathrm{L}} + w_{i\mathrm{pR}}\mathbf{a}_{\mathrm{R}}\right) \\ &= \mathbf{p}_i + \left(w_{i\mathrm{p}} - 1\right)\mathbf{p}_i + \left(w_{i\mathrm{pL}}\mathbf{a}_{\mathrm{L}} + w_{i\mathrm{pR}}\mathbf{a}_{\mathrm{R}}\right), \end{aligned} \tag{16}$$

where $w_{i\mathrm{p}}$ is the weight of $\mathbf{p}_i$ in the extracted component $\hat{\mathbf{p}}_i$, and can be expressed as

$$w_{ip} = \begin{cases} w_{ipL} + k w_{ipR}, & i = \mathrm{L} \\ k^{-1} w_{ipL} + w_{ipR}, & i = \mathrm{R} \end{cases} \tag{17}$$

Substituting (16) into (13), the extraction error $\boldsymbol{\varepsilon}$ becomes

$$\boldsymbol{\varepsilon} = \left(w_{ip} - 1\right)\mathbf{p}_i + \left(w_{ipL}\mathbf{a}_L + w_{ipR}\mathbf{a}_R\right) = \boldsymbol{\varepsilon}_{Ds} + \boldsymbol{\varepsilon}_{Lk}, \tag{18}$$

where $\boldsymbol{\varepsilon}_{Ds} = \left(w_{ip} - 1\right)\mathbf{p}_i$ and $\boldsymbol{\varepsilon}_{Lk} = w_{ipL}\mathbf{a}_L + w_{ipR}\mathbf{a}_R$ are the distortion and leakage comprising the extraction error, respectively. The distortion comes from the extraction weight w_{ip}, which is fluctuating in different blocks and causes variation in sound level or timbre. We consider the primary component $\mathbf{p}_i$ to be completely extracted and hence distortionless if $w_{ip} = 1$. On the other hand, the leakage $\boldsymbol{\varepsilon}_{Lk}$ of the extracted primary component originates from true ambient components $\mathbf{a}_L$ and $\mathbf{a}_R$ from the stereo signal. These two error components are subsequently quantified by the ratio of their power to the power of true components of the stereo signal, as the distortion-to-signal ratio (DSR) [27] and the leakage-to-signal ratio (LSR), respectively:

$$\begin{aligned} \mathrm{DSR}_{Pi} &= P_{\boldsymbol{\varepsilon}_{Ds}} / P_{\mathbf{p}_i} = P_{\left(w_{ip}-1\right)\mathbf{p}_i} / P_{\mathbf{p}_i} = \left(w_{ip} - 1\right)^2, \\ \mathrm{LSR}_{Pi} &= P_{\boldsymbol{\varepsilon}_{Lk}} / P_{\mathbf{p}_i} = P_{w_{ipL}\mathbf{a}_L + w_{ipR}\mathbf{a}_R} / P_{\mathbf{p}_i}. \end{aligned} \tag{19}$$

In the second group of measures, we compute the extraction similarity between the extracted primary component and the true primary component using the cross correlation coefficient at zero lag. We referred to this measure as the extraction-to-true signal correlation (ETSC) [29]:

$$\mathrm{ETSC}_{Pi} = \mathbf{p}_i^T \hat{\mathbf{p}}_i \Big/ \sqrt{\mathbf{p}_i^T \mathbf{p}_i * \hat{\mathbf{p}}_i^T \hat{\mathbf{p}}_i}. \tag{20}$$

Higher correlation indicates higher similarity and better extraction of $\mathbf{p}_i$. By combining the measures of extraction error and extraction similarity in these two groups, a comprehensive evaluation of the extraction accuracy for the primary components can be obtained.

In the third group of measures, we consider the spatial cues of the extracted primary components in the stereo signal. There are three widely used spatial cues for the primary components, namely, the inter-channel cross-correlation coefficient (ICC), ICTD, and ICLD. These cues are used to evaluate the sound localization accuracy of the extracted primary components [2], [12]. There has been many studies to estimate ICTD after the coincidence model proposed by Jeffress (see [40]-[41] and references therein). Based on the classical model proposed by Jeffress [40], the ICC at different time lags is first calculated and the lag number corresponds to the maximum ICC would be the estimated ICTD. Inter-channel level difference is obtained by taking the difference of the power between the right and left channel signals. It shall be noted that the influence of different frequencies on these cues is ignored since no frequency dependency is taken into account in the stereo signal model and the linear estimation model.

3.3 Performance measures for ambient components

Similar performance measures as the primary components can be obtained to quantify the performance of ambience extraction. The extraction error is quantified in the first group. First, we obtain the extraction error between the extracted ambient component $\hat{\mathbf{a}}_i$ and the true ambient component $\mathbf{a}_i$:

$$\boldsymbol{\varepsilon} = \hat{\mathbf{a}}_i - \mathbf{a}_i. \tag{21}$$

Similarly, ESR for the ambient component can be computed:

$$\mathrm{ESR}_{Ai} = P_{\boldsymbol{\varepsilon}} / P_{\mathbf{a}_i}. \tag{22}$$

Based on (11), the general form of $\hat{\mathbf{a}}_i$ is expressed as

$$\hat{\mathbf{a}}_i = w_{iaL}\mathbf{x}_L + w_{iaR}\mathbf{x}_R, \tag{23}$$

where w_{iaL} and w_{iaR} are the elements in the third or fourth row of $\mathbf{W}$. Substituting the signal model in (1) and assumptions in (4) into (23), we arrive at

$$\begin{aligned} \hat{\mathbf{a}}_i &= \left(w_{iaL}\mathbf{p}_L + w_{iaR}\mathbf{p}_R\right) + \left(w_{iaL}\mathbf{a}_L + w_{iaR}\mathbf{a}_R\right) = w_{iai}\mathbf{a}_i + w_{iaj}\mathbf{a}_j + \left(w_{iaL}\mathbf{p}_L + w_{iaR}\mathbf{p}_R\right) \\ &= \mathbf{a}_i + \left(w_{iai} - 1\right)\mathbf{a}_i + w_{iaj}\mathbf{a}_j + \left(w_{iaL}\mathbf{p}_L + w_{iaR}\mathbf{p}_R\right), \end{aligned} \tag{24}$$

where *j* represents the counterpart channel of *i*. Substituting (24) into (21), the extraction error $\boldsymbol{\varepsilon}$ is written as

$$\boldsymbol{\varepsilon} = \left(w_{iai} - 1\right)\mathbf{a}_i + w_{iaj}\mathbf{a}_j + \left(w_{iaL}\mathbf{p}_L + w_{iaR}\mathbf{p}_R\right) = \boldsymbol{\varepsilon}_{Ds} + \boldsymbol{\varepsilon}_{In} + \boldsymbol{\varepsilon}_{Lk}, \tag{25}$$

where the three parts in $\boldsymbol{\varepsilon}$: $\boldsymbol{\varepsilon}_{Ds} = \left(w_{iai} - 1\right)\mathbf{a}_i, \boldsymbol{\varepsilon}_{In} = w_{iaj}\mathbf{a}_j$, and $\boldsymbol{\varepsilon}_{Lk} = w_{iaL}\mathbf{p}_L + w_{iaR}\mathbf{p}_R$ are the distortion, interference, and leakage, respectively. The distortion comes from the extraction weight w_{iai}. We consider the ambient component $\mathbf{a}_i$ to be completely

extracted and hence distortionless if $w_{iai}=1$. Interference $\boldsymbol{\varepsilon}_{\mathrm{In}}$ is produced by the uncorrelated ambient in the other channel $\mathbf{a}_j$, and the leakage $\boldsymbol{\varepsilon}_{\mathrm{Lk}}$ for the extracted ambient component originates from true primary components $\mathbf{p}_{\mathrm{L}}$ and $\mathbf{p}_{\mathrm{R}}$. Again, these three error components are quantified by the ratio of their power to the power of true ambient signal, as DSR, interference-to-signal ratio (ISR), and LSR, respectively:

$$\begin{aligned}
\mathrm{DSR}_{\mathrm{A}i} &= P_{\boldsymbol{\varepsilon}_{\mathrm{Ds}}} / P_{\mathbf{a}_i} = P_{(w_{iai}-1)\mathbf{a}_i} / P_{\mathbf{a}_i} = (w_{iai}-1)^2, \\
\mathrm{ISR}_{\mathrm{A}i} &= P_{\boldsymbol{\varepsilon}_{\mathrm{In}}} / P_{\mathbf{a}_i} = P_{w_{iaj}\mathbf{a}_j} / P_{\mathbf{a}_i} = {w_{iaj}}^2, \\
\mathrm{LSR}_{\mathrm{A}i} &= P_{\boldsymbol{\varepsilon}_{\mathrm{Lk}}} / P_{\mathbf{a}_i} = P_{w_{iaL}\mathbf{p}_{\mathrm{L}}+w_{iaR}\mathbf{p}_{\mathrm{R}}} / P_{\mathbf{p}_i}.
\end{aligned} \tag{26}$$

A similar decomposition of the error components with their corresponding measures can be found in [42]. Comparing our measures of extraction error for primary components and ambient components, we find that no interference is found in the extracted primary components due to the complete correlation of the primary components. All measures between primary and ambient components extraction are very similar, except ISR. Error-to-signal ratio measures the overall error of the extracted component, and these additional three measures would provide more information on the evaluation of the extraction performance. In particular, LSR corresponds to the perceptual difference between the primary and ambient components. Since leakage is generally much more noticeable and undesirable than interference or distortion in the extracted signals, LSR is generally the most critical measure among the three. Nevertheless, more consideration should be placed on DSR when sound level or timbre is crucial.

In the second group of measures, we also use ETSC to determine the extraction similarity between the extracted ambient component and the true ambient component:

$$\mathrm{ETSC}_{\mathrm{A}i} = \mathbf{a}_i^{\,T}\hat{\mathbf{a}}_i \Big/ \sqrt{\mathbf{a}_i^{\,T}\mathbf{a}_i * \hat{\mathbf{a}}_i^{\,T}\hat{\mathbf{a}}_i}\,. \tag{27}$$

Higher correlation indicates higher similarity and better extraction of $\mathbf{a}_i$. By combining the measures of extraction error and extraction similarity in these two groups, a comprehensive evaluation of the extraction accuracy for ambient components can be obtained.

As the ambient component is assumed to be uncorrelated in both channels, we consider ICC and ICLD for the extracted ambient components in the third group of measures. ICC is a good measure of the diffuseness of the stereo signal [43]. Since the ambience is diffused and relatively balanced in the two channels of the stereo signal, a better extraction of the ambience is obtained when the ICC is lower and the ICLD is closer to one.

4. PRIMARY-AMBIENT EXTRACTION BASED ON LINEAR ESTIMATION

Following the discussion in Chapter 3, we shall derive the optimal weights in **W** for PAE of stereo signals. In this chapter, we assume that the input signal satisfies all the assumptions discussed in Chapter 2. Next, we will discuss four approaches, namely, PCA, LS, MLLS, and MDLS, with different objectives to estimate the weights for the extraction of the primary and ambient components. Finally, we compare the performance of these four PAE approaches.

4.1 Primary-ambient extraction using PCA

Principal component analysis is a widely used method in multivariate analysis [36]. The central idea of PCA is to linearly transform the data set into orthogonal principal components with descending variances. PCA was introduced to solve the PAE problem in [30]. A geometric representation of PCA based PAE is illustrated in Fig. 4.1. Based on the stereo signal model, PAE using PCA decomposition can be mathematically described as [29]:

$$\mathbf{u}_0 = \arg\max_{\mathbf{u}_0}\left(\left\|\mathbf{u}_0^T\mathbf{x}_L\right\|^2 + \left\|\mathbf{u}_0^T\mathbf{x}_R\right\|^2\right),\ \mathbf{u}_1 = \arg\min_{\mathbf{u}_1}\left(\left\|\mathbf{u}_1^T\mathbf{x}_L\right\|^2 + \left\|\mathbf{u}_1^T\mathbf{x}_R\right\|^2\right),\ \text{s.t. } \mathbf{u}_0 \perp \mathbf{u}_1, \tag{28}$$

where $\mathbf{u}_0, \mathbf{u}_1$ are the primary and ambient basis vectors that maximize and minimize the total projection energy of the input signal vectors, respectively. A closed-form solution of (28) can be obtained by eigenvalue decomposition of the input covariance matrix [10].

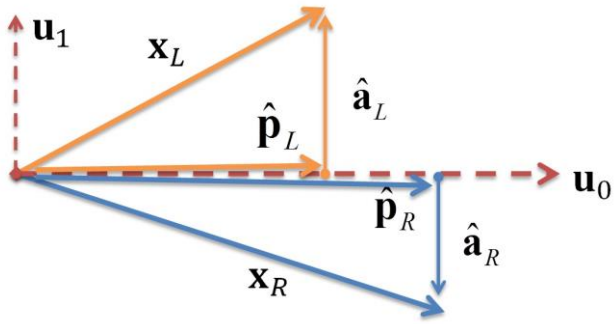

Figure 4.1 A geometric representation of PAE using PCA.

In general, primary component possesses more power than ambience, i.e., $\gamma > 0.5$. Hence, it is common practice to relate the larger eigenvalue to the primary component and the smaller eigenvalue to the ambience. In the following, we shall observe that this assumption is redundant. First, we find the larger eigenvalue and its corresponding primary basis vector

$$\lambda_0 = 0.5\left[r_{LL} + r_{RR} + \sqrt{\left(r_{LL} - r_{RR}\right)^2 + 4r_{LR}^{\ 2}}\right], \tag{29}$$

$$\mathbf{u}_0 = r_{LR}\mathbf{x}_L + \left(\lambda_0 - r_{LL}\right)\mathbf{x}_R. \tag{30}$$

Next, we compute the extracted primary components as

$$\hat{\mathbf{p}}_L = \frac{\mathbf{u}_0^{\ H}\mathbf{x}_L}{\mathbf{u}_0^{\ H}\mathbf{u}_0}\mathbf{u}_0,\ \hat{\mathbf{p}}_R = \frac{\mathbf{u}_0^{\ H}\mathbf{x}_R}{\mathbf{u}_0^{\ H}\mathbf{u}_0}\mathbf{u}_0. \tag{31}$$

Using (6)-(10), the expressions for the extracted primary components can be simplified to

$$\hat{\mathbf{p}}_L(\text{PCA}) = \frac{1}{1+k^2}\left(\mathbf{x}_L + k\mathbf{x}_R\right),\ \hat{\mathbf{p}}_R(\text{PCA}) = k\hat{\mathbf{p}}_L(\text{PCA}) = \frac{k}{1+k^2}\left(\mathbf{x}_L + k\mathbf{x}_R\right). \tag{32}$$

Similarly, the extracted ambient components are obtained as

$$\hat{\mathbf{a}}_L(\text{PCA}) = \frac{k}{1+k^2}\left(k\mathbf{x}_L - \mathbf{x}_R\right),\ \hat{\mathbf{a}}_R(\text{PCA}) = -\frac{1}{k}\hat{\mathbf{a}}_L(\text{PCA}) = -\frac{1}{1+k^2}\left(k\mathbf{x}_L - \mathbf{x}_R\right). \tag{33}$$

From (32)-(33), we observe that the weights for the extracted primary and ambient components are solely dependent on PPF, and the primary and ambient components extracted by PCA are not affected by PPR. The primary components are correlated and scaled by *k* between the right and left channels, while the ambience is negatively correlated and panned to the opposite direction of the

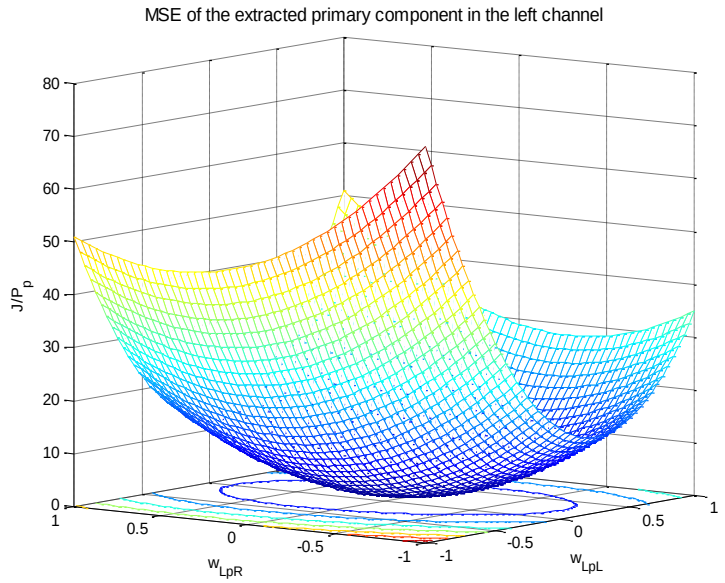

Figure 4.2 An example of MSE performance surface of $\hat{\mathbf{p}}_{\mathrm{L}}$ ($k=5,\gamma=0.5$).

primary components, as indicated by the scaling factor $-1/k$. Rewriting the expressions of the extracted primary and ambient components in (31)-(32) using the true primary and ambient components, we have

$$\hat{\mathbf{p}}_{\mathrm{L}}(\mathrm{PCA})=\mathbf{p}_{\mathrm{L}}+\frac{1}{1+k^2}(\mathbf{a}_{\mathrm{L}}+k\mathbf{a}_{\mathrm{R}}),\ \hat{\mathbf{p}}_{\mathrm{R}}(\mathrm{PCA})=\mathbf{p}_{\mathrm{R}}+\frac{k}{1+k^2}(\mathbf{a}_{\mathrm{L}}+k\mathbf{a}_{\mathrm{R}}), \tag{34}$$

$$\hat{\mathbf{a}}_{\mathrm{L}}(\mathrm{PCA})=\frac{k^2}{1+k^2}\mathbf{a}_{\mathrm{L}}-\frac{k}{1+k^2}\mathbf{a}_{\mathrm{R}},\ \hat{\mathbf{a}}_{\mathrm{R}}(\mathrm{PCA})=\frac{1}{1+k^2}\mathbf{a}_{\mathrm{R}}-\frac{k}{1+k^2}\mathbf{a}_{\mathrm{L}}. \tag{35}$$

No primary components are found in (35), indicating that the expressions for the extracted primary and ambient components in (34) and (35) are not inter-changeable. Thereby, the basis vector with larger variance always corresponds to the primary component regardless of the power difference between the primary and ambient components. This observation shows that the assumption that primary component power is always higher than ambience power is redundant. However, if this assumption is not satisfied in the stereo signal, the extraction error is more likely to increase, as can be seen from (34).

4.2 Primary-ambient extraction using LS

Least-squares is frequently used to approximate solutions of 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 component [37]. Considering the extraction of the primary component in the left channel:

$$\hat{\mathbf{p}}_{\mathrm{L}}=w_{\mathrm{LpL}}\mathbf{x}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{x}_{\mathrm{R}}=(w_{\mathrm{LpL}}+kw_{\mathrm{LpR}})\mathbf{p}_{\mathrm{L}}+w_{\mathrm{LpL}}\mathbf{a}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{a}_{\mathrm{R}}. \tag{36}$$

The extraction error can then be written as

$$\boldsymbol{\varepsilon}_{\mathbf{p}_{\mathrm{L}}}=\hat{\mathbf{p}}_{\mathrm{L}}-\mathbf{p}_{\mathrm{L}}=(w_{\mathrm{LpL}}+kw_{\mathrm{LpR}}-1)\mathbf{p}_{\mathrm{L}}+w_{\mathrm{LpL}}\mathbf{a}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{a}_{\mathrm{R}}, \tag{37}$$

and the MSE is

$$J=E\left[\boldsymbol{\varepsilon}_{\mathbf{p}_{\mathrm{L}}}{}^{T}\boldsymbol{\varepsilon}_{\mathbf{p}_{\mathrm{L}}}\right]. \tag{38}$$

By substituting the assumptions and relationships in the signal model stated in (3)-(5) and (37) into (38), the MSE becomes

$$J=P_{\mathrm{p}}\left\{\left[1+(k^2+1)\frac{1-\gamma}{2\gamma}\right]w_{LpL}{}^2+\left[k^2+(k^2+1)\frac{1-\gamma}{2\gamma}\right]w_{\mathrm{LpR}}{}^2-2w_{\mathrm{LpL}}-2kw_{\mathrm{LpR}}+2kw_{\mathrm{LpL}}w_{\mathrm{LpR}}+1\right\}. \tag{39}$$

An example of the two-dimensional MSE function for $\hat{\mathbf{p}}_{\mathrm{L}}$ is shown in Fig. 4.2. The bottom of the bowl shape curve corresponds to the minimum MSE and optimal weights. Hence, the optimal weights can be easily obtained by taking the gradients of *J* with respect to $w_{\mathrm{LpL}}, w_{\mathrm{LpR}}$ and equating their results to zero. The optimal weights for LS are

$$w_{\mathrm{LpL}}(\mathrm{LS})=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}, w_{\mathrm{LpR}}(\mathrm{LS})=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}. \tag{40}$$

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

$$\hat{\mathbf{p}}_{\mathrm{L}}(\mathrm{LS})=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}),\ \hat{\mathbf{p}}_{\mathrm{R}}(\mathrm{LS})=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}), \tag{41}$$

$$\begin{aligned}\hat{\mathbf{a}}_{\mathrm{L}}(\mathrm{LS})&=\frac{1+k^2+(k^2-1)\gamma}{1+\gamma}\frac{1}{1+k^2}\mathbf{x}_{\mathrm{L}}-\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\mathbf{x}_{\mathrm{R}},\\ \hat{\mathbf{a}}_{\mathrm{R}}(\mathrm{LS})&=-\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\mathbf{x}_{\mathrm{L}}+\frac{1+k^2+(1-k^2)\gamma}{1+\gamma}\frac{1}{1+k^2}\mathbf{x}_{\mathrm{R}}.\end{aligned} \tag{42}$$

From (41)-(42), we observe that the weights for the extracted primary and ambient components are not only dependent on PPF, but also related to PPR. However, the panning relationship in the extracted primary component is not found in the extracted ambience.

4.3 Primary-ambient extraction using MLLS

As discussed in Chapter 3, three types of error may be found in the extracted components, namely, the distortion, the interference, and the leakage. As both the distortion and interference come from the true component in the two channels, they may exhibit some similarity with the true component, and making them less likely to be detected by the listener when mixed in the extracted component. Conversely, the leakage is the most undesirable as the leakage is usually the most perceptually noticeable among the three errors. Having this in mind, priority should be given to the minimization of the leakage in the extraction process. The relative leakage power can be quantified by the leakage-to-extracted signal ratio (LeSR), and is given as

$$\mathrm{LeSR}_{\mathrm{P}i}=P_{\boldsymbol{\varepsilon}_{\mathrm{L}k}}/P_{\hat{\mathbf{p}}_i},\ \ \mathrm{LeSR}_{\mathrm{A}i}=P_{\boldsymbol{\varepsilon}_{\mathrm{L}k}}/P_{\hat{\mathbf{a}}_i}. \tag{43}$$

By minimizing ESR using the constraint that extracted component has the minimum relative leakage, the minimum leakage least squares is obtained. Again, considering the extracted primary component in the left channel, we have

$$\hat{\mathbf{p}}_{\mathrm{L}}=w_{\mathrm{LpL}}\mathbf{x}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{x}_{\mathrm{R}}=(w_{\mathrm{LpL}}+kw_{\mathrm{LpR}})\mathbf{p}_{\mathrm{L}}+w_{\mathrm{LpL}}\mathbf{a}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{a}_{\mathrm{R}}, \tag{44}$$

and the leakage comes from the ambient component. Using (43)-(44), the $\mathrm{LeSR}_{\mathrm{PL}}$ is computed:

$$\mathrm{LeSR}_{\mathrm{PL}}=\frac{(w_{\mathrm{LpL}}{}^2+w_{\mathrm{LpR}}{}^2)P_{\mathrm{a}}}{(w_{\mathrm{LpL}}+kw_{\mathrm{LpR}})^2P_{\mathrm{p}}+(w_{\mathrm{LpL}}{}^2+w_{\mathrm{LpR}}{}^2)P_{\mathrm{a}}}. \tag{45}$$

Minimizing $\mathrm{LSIR}_{\mathrm{PL}}$ with respect to $w_{\mathrm{LpL}}, w_{\mathrm{LpR}}$, we find

$$w_{\mathrm{LpR}}=kw_{\mathrm{LpL}}. \tag{46}$$

Next, we find the extraction error:

$$\boldsymbol{\varepsilon}=\hat{\mathbf{p}}_{\mathrm{L}}-\mathbf{p}_{\mathrm{L}}=(w_{\mathrm{LpL}}+kw_{\mathrm{LpR}}-1)\mathbf{p}_{\mathrm{L}}+w_{\mathrm{LpL}}\mathbf{a}_{\mathrm{L}}+w_{\mathrm{LpR}}\mathbf{a}_{\mathrm{R}}. \tag{47}$$

By substituting (46) into (47), the error becomes

$$\boldsymbol{\varepsilon}=\hat{\mathbf{p}}_{\mathrm{L}}-\mathbf{p}_{\mathrm{L}}=\left[(1+k^2)w_{\mathrm{LpL}}-1\right]\mathbf{p}_{\mathrm{L}}+w_{\mathrm{LpL}}\mathbf{a}_{\mathrm{L}}+kw_{\mathrm{LpL}}\mathbf{a}_{\mathrm{R}}. \tag{48}$$

Using (16), the $\mathrm{ESR}_{\mathrm{PL}}$ is expressed as

$$\mathrm{ESR}_{\mathrm{PL}}=\frac{\left[(1+k^2)w_{\mathrm{LpL}}-1\right]^2P_{\mathrm{p}}+(w_{\mathrm{LpL}}{}^2+k^2w_{\mathrm{LpL}}{}^2)P_{\mathrm{a}}}{P_{\mathrm{p}}}. \tag{49}$$

By minimizing $\mathrm{ESR}_{\mathrm{PL}}$, we arrive to

$$w_{\mathrm{LpL}}(\mathrm{MLLS})=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}. \tag{50}$$

Substituting (50) into (46), we then have

$$w_{\mathrm{LpR}}(\mathrm{MLLS})=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}. \tag{51}$$

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

$$\hat{\mathbf{p}}_{\mathrm{L}}(\mathrm{MLLS})=\frac{2\gamma}{1+\gamma}\frac{1}{1+k^2}(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}), \tag{52}$$

and using the same steps as discussed above, the remaining components are given as

$$\hat{\mathbf{p}}_{\mathrm{R}}\left(\mathrm{MLLS}\right)=\frac{2\gamma}{1+\gamma}\frac{k}{1+k^2}\left(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}\right), \tag{53}$$

$$\hat{\mathbf{a}}_{\mathrm{L}}\left(\mathrm{MLLS}\right)=\frac{k}{1+k^2}\left(k\mathbf{x}_{\mathrm{L}}-\mathbf{x}_{\mathrm{R}}\right),\ \hat{\mathbf{a}}_{\mathrm{R}}\left(\mathrm{MLLS}\right)=-\frac{1}{1+k^2}\left(k\mathbf{x}_{\mathrm{L}}-\mathbf{x}_{\mathrm{R}}\right). \tag{54}$$

Comparing the extraction results of MLLS with those using PCA and LS, we notice that MLLS combines the extracted primary component from LS and the extracted ambient component from PCA. This is because by having the same minimum leakage in the extracted primary component as PCA, only LS possess the minimum error. On the other hand, the severe primary leakage problem in ambience extracted using LS is not found in PCA. Thus, it is clear that MLLS extracts both the primary and ambient components with minimal leakage and error.

4.4 Primary-ambient extraction using MDLS

Inspired by the popular minimum variance distortionless response (MVDR) filter [44], the minimum distortion least squares can be applied in PAE by minimizing the extraction error ESR using the constraint that the extracted component is undistorted. Mathematically, we can express it as follows

$$\min_{\mathbf{w}}\mathrm{ESR}\ \text{s.t.}\ w_{ip}\left(\text{or } w_{iai}\right)=1. \tag{55}$$

With (55), the solution to (48) for each extracting component is easily found to be

$$\hat{\mathbf{p}}_{\mathrm{L}}\left(\mathrm{MDLS}\right)=\frac{1}{1+k^2}\left(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}\right),\ \hat{\mathbf{p}}_{\mathrm{R}}\left(\mathrm{MDLS}\right)=\frac{k}{1+k^2}\left(\mathbf{x}_{\mathrm{L}}+k\mathbf{x}_{\mathrm{R}}\right). \tag{56}$$

$$\hat{\mathbf{a}}_{\mathrm{L}}\left(\mathrm{MDLS}\right)=\mathbf{x}_{\mathrm{L}}-\frac{2k\gamma}{\left(k^2-1\right)\gamma+k^2+1}\mathbf{x}_{\mathrm{R}},\hat{\mathbf{a}}_{\mathrm{R}}\left(\mathrm{MDLS}\right)=-\frac{2k\gamma}{\left(1-k^2\right)\gamma+k^2+1}\mathbf{x}_{\mathrm{L}}+\mathbf{x}_{\mathrm{R}}. \tag{57}$$

4.5 Comparison and discussion

In this subsection, we shall compare the relationships and differences of the extraction results as well as the performance among the four PAE approaches, namely, PCA, LS, MLLS, and MDLS.

Comparing the expressions for the extracted primary components in (32), (41), (52)-(53), and (56), we observe that the primary component extraction results of MDLS and MLLS are identical with the results of PCA and LS, respectively. This indicates that PCA extracts the primary component with minimum distortion and error, while LS leads to minimum leakage and error in the extracted primary component. The primary components extracted using PCA and LS are both panned by PPF between the two channels, revealing that the PPF is faithfully retained in the extracted primary components. A scaling difference is noticed between the primary components extracted by PCA and by LS, i.e.,

$$\hat{\mathbf{p}}_{\mathrm{L(R)}}\left(\mathrm{LS}\right)=\alpha_{\mathrm{p}}\hat{\mathbf{p}}_{\mathrm{L(R)}}\left(\mathrm{PCA}\right), \tag{58}$$

where the scaling factor $\alpha_{\mathrm{p}}=2\gamma/\left(1+\gamma\right)$ is a function of PPR. Since $\gamma\in[0,1]$, α_{p} is smaller or equal to one suggests that the primary component extracted by LS has lower power than the component extracted using PCA for all values of PPR, except PPR=1.

Considering the ambient extraction of PCA, LS, MLLS, and MDLS described by (33), (42), (54), and (57), respectively, MLLS has the same performance as PCA which extracts the ambient component with minimum leakage from the primary component. The difference between the ambient components extracted by MDLS and LS is also a scaling factor, which is given as

$$\hat{\mathbf{a}}_{\mathrm{L(R)}}\left(\mathrm{LS}\right)=\alpha_{\mathrm{a,L(R)}}\hat{\mathbf{a}}_{\mathrm{L(R)}}\left(\mathrm{MDLS}\right). \tag{59}$$

As compared to (58), the scaling factor is different for the left and right channels.

In the following section, a comparative analysis of PAE using PCA, LS, MLLS, and MDLS is presented. Here, we summarize the results of the performance measures obtained with the left channel in Table 4-1. For the stereo signal model, it is noted that all the measures for the left and right channel in PAE are symmetrical, and therefore, the measures for the right channel can be obtained by replacing the PPF with its reciprocal. It is clear from Table 4-1 that the performance of PCA, LS, MLLS, and MDLS are highly dependent on PPR and PPF. A higher PPR in PAE indicates higher accuracy in extracting the primary component as the errors ($\mathrm{ESR_{PL}}$, $\mathrm{LSR_{PL}}$) become smaller and the extracted primary component is more similar to the true primary component (as shown by $\mathrm{ETSC_{PL}}$). Generally speaking, $\mathrm{ESR_{PL}}$ in LS is observed to be lower than that in PCA since $\left(1-\gamma\right)/2\gamma\le\left(1-\gamma\right)/\left(1+\gamma\right)$, and similar observations are also found with $\mathrm{LSR_{PL}}$. The distortion measure $\mathrm{DSR_{PL}}=0$ indicates that primary component extracted using PCA and MDLS is free of distortion, while the distortion in LS and MLLS increases as PPR decreases. Hence, LS and MLLS extract primary components with minimum leakage and error at the expense of introducing some distortion in the extracted primary components. All the approaches discussed in this section extract primary component without interference. For the spatial cues

Table 4-1 Evaluation results for PCA, LS, MLLS, and MDLS in PAE (left channel)

Measure	Primary component		Ambient component		
	PCA/ MDLS	LS/ MLLS	PCA/ MLLS	LS	MDLS
ESR	$\frac{1-\gamma}{2\gamma}$	$\frac{1-\gamma}{1+\gamma}$	$\frac{1}{1+k^2}$	$\frac{1}{1+k^2}\frac{2\gamma}{1+\gamma}$	$\frac{2\gamma}{(k^2-1)\gamma+k^2+1}$
LSR	$\frac{1-\gamma}{2\gamma}$	$\frac{1-\gamma}{2\gamma}\left(\frac{2\gamma}{1+\gamma}\right)^2$	0	$\frac{1}{1+k^2}\frac{2\gamma(1-\gamma)}{(1+\gamma)^2}$	$\frac{(1+k^2)(1-\gamma)2\gamma}{\left((1+k^2)(1+\gamma)-2\gamma\right)^2}$
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		$\left(\frac{k}{1+k^2}\right)^2$	$\left(\frac{k}{1+k^2}\frac{2\gamma}{1+\gamma}\right)^2$	$\left(\frac{2k\gamma}{(1+k^2)(1+\gamma)-2\gamma}\right)^2$
ETSC	$\sqrt{\frac{2\gamma}{1+\gamma}}$		$\sqrt{\frac{k^2}{1+k^2}}$	$\sqrt{\frac{k^2}{1+k^2}+\frac{1-\gamma}{(1+k^2)(1+\gamma)}}$	
ICC(ICTD)	1(0)		1	$\frac{2k\gamma}{\sqrt{(1+k^2)^2-(1-k^2)^2\gamma^2}}$	
ICLD	k^2		$\frac{1}{k^2}$	$\frac{1}{k^2}\frac{1+\gamma+k^2(1-\gamma)}{1+\gamma+\frac{1}{k^2}(1-\gamma)}$	$\frac{1}{k^2}\frac{1-\gamma+k^2(1+\gamma)}{1-\gamma+\frac{1}{k^2}(1+\gamma)}$

(ICC_P, $ICLD_P$) of the primary component, all four approaches are capable of preserving the spatial information in the extracted primary components.

Significant differences in the measures of the extracted ambient components are found among PCA, LS, MLLS, and MDLS. Considering the extraction accuracy of the ambient component, we observe that ESR_{AL} in LS is the smallest. The measure LSR_{AL} = 0 in PCA indicates that no primary components are leaked into the extracted ambient component. By contrast, certain amount of primary leakage is found in ambient component extracted using LS or MDLS. Among the four approaches, only MDLS extracts ambience without distortion, and PCA or MLLS has the worst performance in this respect. Based on $ETSC_{AL}$, the ambient component extracted by LS and MDLS tends to be more similar to the true ambient than PCA and MLLS. Although none of the approaches produce uncorrelated and balanced ambience, the best performance on this regard is achieved using LS.

By assuming the input signal satisfies all the assumptions discussed in Chapter 2, PPF and PPR can be correctly estimated using (9) and (10), respectively. We found the four approaches perform well in extracting the primary component from the stereo signal. All of these approaches produce zero interference in the extracted primary component; smaller ESR and LSR are found in LS and MLLS, while PCA and MDLS extract the primary component free from distortion. Inferior performance is found for ambient extraction using these four approaches. While LS produces slightly better results in most of the performance measures, the extracted ambient component using PCA and MLLS exhibits zero primary leakage, and only the ambience extracted using MDLS is undistorted. Due to the limitation of the linear estimation, it is impossible to obtain a perfectly uncorrelated and balanced ambience using linear estimated based PAE approaches. Therefore, some post-processing techniques such as decorrelation [45] and post-scaling [37] should be applied to further enhance the extracted ambience.

4.6 Experimental results

A series of simulations are conducted to validate the theoretical analysis presented in earlier sections. In these simulations, a speech signal is selected as the primary component and uncorrelated white Gaussian noise with equal variance in two channels is synthesized as ambience. To simulate different panning to the left, middle, and right, the primary component is amplitude panned to the right channel by letting k = 0.2, 1, and 5, respectively. Subsequently, the input stereo signals are obtained by linearly mixing

the primary and ambient components based on PPR, ranging from zero to one. Four approaches discussed in this chapter are employed to extract primary and ambient components from the synthesized stereo signals. Finally, the performance of PAE using these approaches is evaluated based on the performance measures introduced in Chapter 3. Furthermore, some practical recommendations in the application of these approaches are also discussed based on the evaluation and comparison.

The simulation results of PAE using PCA, LS, MLLS, and MDLS are shown in Figs. 4.3-4.6. The plots in Fig. 4.3 revealed that the estimated PPR and PPF are almost the same as the true PPR and PPF. Since the extraction results of primary component are identical for PCA and LS with MDLS and MLLS, respectively, only the performance results of primary component extraction using PCA and LS are shown in the following. The extraction accuracy of primary component in both channels (k=5) is compared in Fig. 4.4. As PPR increases, the extraction error given by $\mathrm{ESR}_{\mathrm{P}i}$ and $\mathrm{LSR}_{\mathrm{P}i}$ reduces, and the extraction similarity given by $\mathrm{ETSC}_{\mathrm{P}i}$ increases in both channels. While most of the values of $\mathrm{LSR}_{\mathrm{P}i}$ and $\mathrm{ETSC}_{\mathrm{P}i}$ in PCA and LS are identical, it is observed that $\mathrm{ESR}_{\mathrm{P}i}$ in LS is relatively lower than that in PCA, which indicates that LS is superior to PCA in extracting the primary component in terms of MSE between the extracted primary component and true primary component. On the other hand, the distortion of primary component extracted using LS is high when PPR is low, while no distortion is found with PCA.

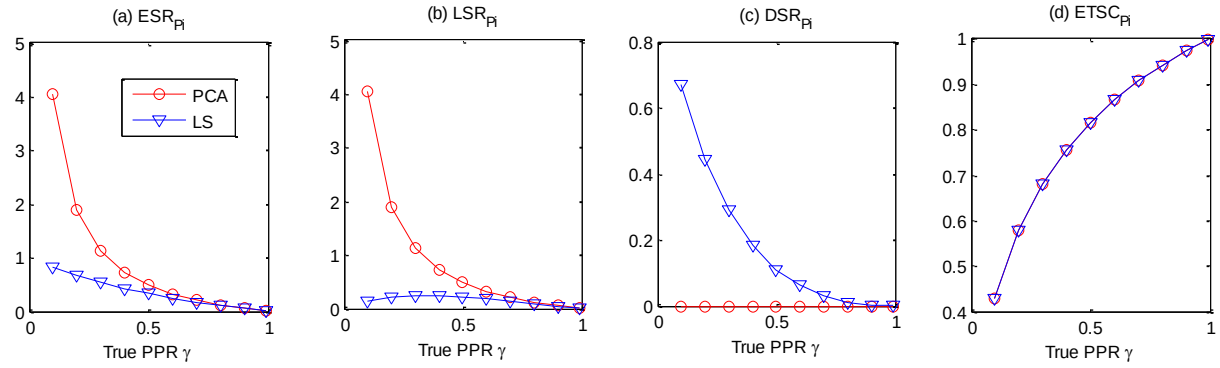

Figure 4.3 Comparison of PCA (or MDLS) and LS (or MLLS) in primary component extraction (k=5). Legend in (a) applies to all plots.

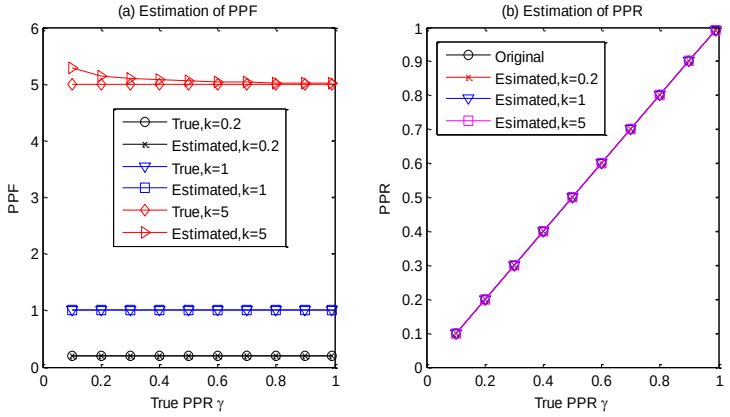

Figure 4.4 Estimation of PPF and PPR.

The performance for the ambience extraction in both channels (k=5) is illustrated in Fig. 4.5. Unlike the primary component extraction, the performance of ambient extraction in two channels has significant variation in the four approaches. Since ambience is relatively stronger in the left channel as compared to primary component, the performance of ambience extraction in the left channel is better than that in the right channel. As compared with that in the right channel, the performance of ambience extraction in the left channel is quite similar among the four approaches. It is found that LS has lower extraction error (as shown by $\mathrm{ESR}_{\mathrm{AR}}$) and higher extraction similarity (as shown by $\mathrm{ETSC}_{\mathrm{AR}}$), PCA (or MLLS), and MDLS can completely remove the leakage and distortion, respectively. The extraction error including leakage and interference of the extracted ambience in the left channel using MDLS is obviously much higher than those of the other methods. Figure 4.6 presents the results of the localization parameters. It is

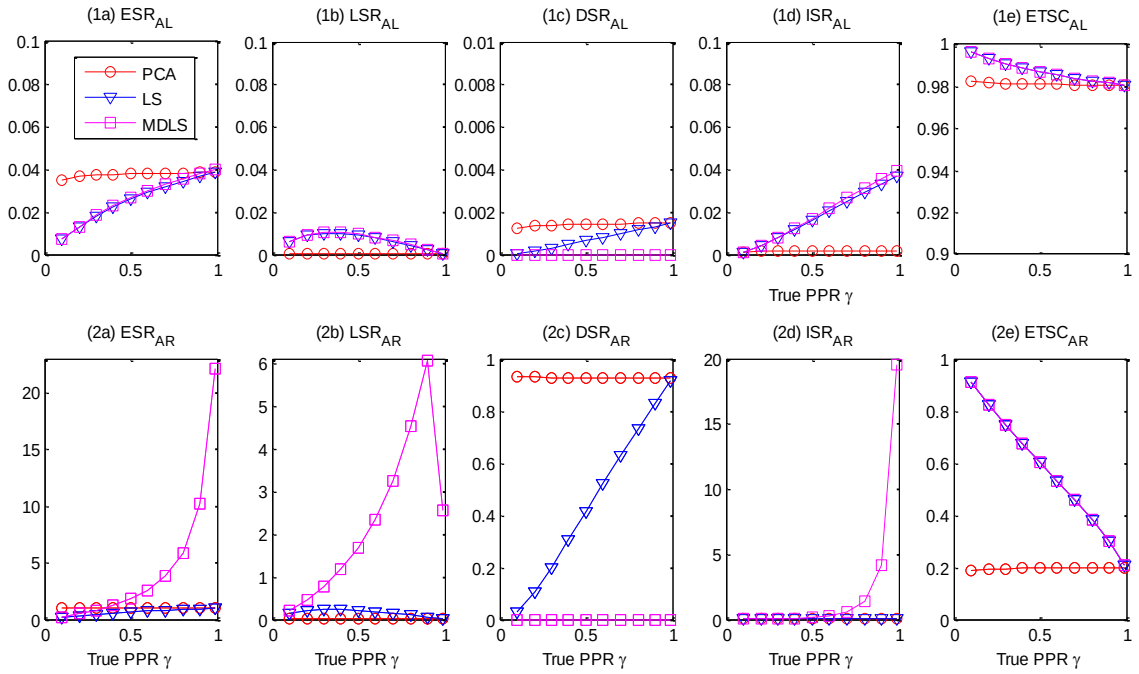

Figure 4.5 Comparison of PCA (MLLS), LS and MDLS in ambient extraction (k=5). Legend in (1a) applies to all plots.

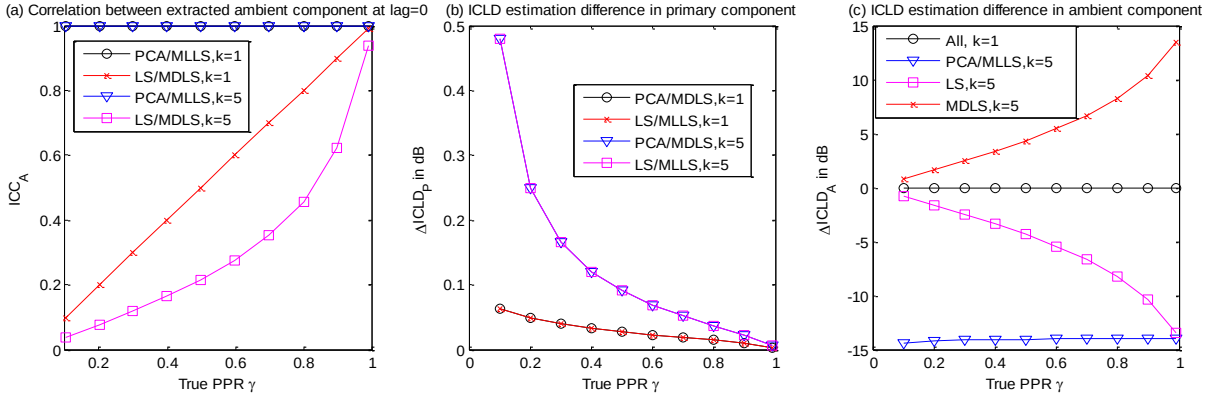

Figure 4.6 Comparison of spatial cues in PAE using PCA, LS, MLLS, and MDLS.

found that the ICC, ICTD, and ICLD of the primary component extracted by PCA and LS are very close to the true values. However, none of these approaches are able to extract uncorrelated and balanced ambience when $k \neq 1$.

From (59), the difference between PCA and LS in extracted primary component is given by a scaling factor. This scaling factor is solely dependent on PPR, which is determined by the power difference between true primary and ambient components in every block as the input signal is usually divided into small blocks before PAE. In the case of stationary primary and ambient components, the scaling factor is almost constant and leading to similar PAE performance in PCA and LS. When considering the non-stationary primary component such as speech, this operation leads to noticeable difference between the extracted components using PCA and LS. An example to illustrate the variation of the scaling factor in a signal is shown in Fig. 4.7. It is observed that the scaling factor is fluctuating according to the power difference between primary and ambient components. The scaling factor rises close to one when primary component power is comparably stronger than the relatively constant ambient power. Likewise, the scaling factor drops to zero when the primary component becomes weak. This indicates that PCA and LS behave similarly when primary component is dominant, and only LS extracts weak primary component at the ambient-dominant periods of the signal. As a result, LS has low ESR_{pi} but the extracted primary component may have some discontinuity and more distortion. A tradeoff approach to reduce the distortion in the primary component extracted by LS is to slow down the fast changes of the scaling factor using a forgetting factor in the recursive computation of scaling factor.

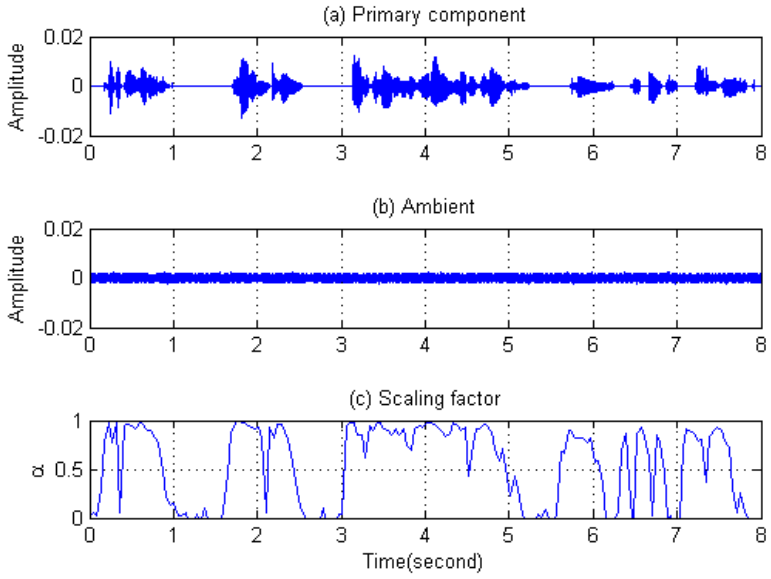

Figure 4.7 Comparison of PCA and LS in primary component extraction.

Several guidelines on the applications of these approaches for PAE can be drawn from our previous analysis and discussions. The selection of the best PAE approach is generally based on the post-processing techniques and playback systems, associated with the specific audio application, as well as audio content and user preference. In some spatial audio enhancement systems where the extracted primary or ambient component are summed with 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 preferred as the leakage becomes the most important performance consideration. MLLS is also suggested when different rendering and playback techniques are employed on the extracted primary and ambient components. In the case where both primary and ambient components are extracted, processed, and finally mixed together, then the extraction error becomes more critical and LS is recommended. In cases like spatial audio coding and interactive audio in gaming, where the primary component is usually more important than ambient component, PCA would be a better choice. For hi-fidelity applications where timbre is of high importance, such as in musical applications, MDLS is more suitable.

5. PRIMARY-AMBIENT EXTRACTION IN PRIMARY-COMPLEX CASE

It is unlikely for any stereo input signal to fulfill all the assumptions of the stereo signal model discussed in Chapter 2. Several new cases can be defined by relaxing one or more of the assumptions of the stereo signal model. In this chapter, we focus our discussion on one of the most frequent case which has a partially correlated primary component at zero lag. For the rest of this chapter, we shall refer this case as the primary-complex case. In this chapter, performance of the PCA based PAE is evaluated in primary-complex case. Next, time shifted PCA based approach is introduced to improve the extraction performance.

5.1 Primary-ambient extraction in primary-complex case

Based on the signal model in (1), by considering a partially correlated primary component, the correlation coefficient of the primary component becomes

$$\phi_p = \frac{\mathbf{p}_L{}^H \mathbf{p}_R}{\sqrt{\mathbf{p}_L{}^H \mathbf{p}_L \mathbf{p}_R{}^H \mathbf{p}_R}} < 1, \mathbf{a}_L \perp \mathbf{a}_R, \mathbf{p}_L(\mathbf{p}_R) \perp \mathbf{a}_L(\mathbf{a}_R), \tag{60}$$

$$P_{\mathbf{p}_R} = k^2 P_{\mathbf{p}_L} = k^2 P_{\mathbf{p}}, P_{\mathbf{a}_L} = P_{\mathbf{a}_R} = P_a, \gamma \in (0.5,1). \tag{61}$$

The zero-lag auto-correlation and cross-correlation of the input signal are

$$r_{LL} = P_{\mathbf{p}} + P_a, r_{RR} = k^2 P_{\mathbf{p}} + P_a, r_{LR} = \phi_p k P_{\mathbf{p}}, \tag{62}$$

and hence the estimated PPF and PPR are

$$\hat{k}_{pc} = k = \phi_p \frac{r_{RR} - r_{LL}}{2r_{LR}} + \sqrt{\left(\phi_p \frac{r_{RR} - r_{LL}}{2r_{LR}}\right)^2 + 1}, \tag{63}$$

$$\hat{\gamma}_{pc} = \gamma = \frac{2r_{LR} + \phi_p (r_{RR} - r_{LL})\hat{k}_{pc}}{\phi_p (r_{RR} + r_{LL})\hat{k}_{pc}}. \tag{64}$$

Good estimates of PPF and PPR require *a priori* knowledge of the primary correlation ϕ_p. However, the primary correlation may not be available at all times. To find a solution to this problem, we shall revisit the solutions that are obtained in the ideal case. By substituting (62) in (9)-(10), we can find the ratio difference between the estimated PPF, PPR and true PPF, PPR, respectively as

$$\frac{\hat{k}_{ic}}{k} = \frac{k^2 - 1}{2\phi_p k^2} + \sqrt{\left(\frac{k^2 - 1}{2\phi_p k^2}\right)^2 + \frac{1}{k^2}}, \tag{65}$$

$$\frac{\hat{\gamma}_{ic}}{\gamma} = \frac{\sqrt{\left(k^2 - 1\right)^2 + 4\phi_p{}^2 k^2}}{k^2 + 1}. \tag{66}$$

With reference to (65) and (66), the ratio differences of PPF and PPR estimated in the primary-complex with respect to the correlation of the primary component are illustrated in Fig. 5.1. It is observed that PPF is wrongly estimated when $k \neq 1$, and PPR is often underestimated. Furthermore, the accuracy of the estimation of PPF and PPR increases as the correlation of the primary component increases. It is interesting to note when the panning is weaker, the estimated PPF is closer to the true values whereas estimated PPR, on the contrary, becomes less accurate. The inaccuracy of PPF and PPR results in the ICLD of the primary component to be incorrect and lower extraction performance.

Next, we analyze the performance of PCA based PAE in primary-complex case. The analysis can be directly extended to other linear estimation based approaches. Note that we mainly focus on the degradation of the extraction performance caused by the partially correlated primary component rather than the estimation error in PPF and PPR. For this reason, the estimation error of PPF and PPR is ignored in this analysis. First, we rewrite (32)-(33) using the true primary and ambient components:

$$\hat{\mathbf{p}}_L(\mathrm{PCA}) = \mathbf{p}_L + \frac{-k}{1+k^2}(k\mathbf{n}_L - \mathbf{n}_R) + \frac{1}{1+k^2}(\mathbf{a}_L + k\mathbf{a}_R), \tag{67}$$

$$\hat{\mathbf{p}}_R(\mathrm{PCA}) = \mathbf{p}_R + \frac{1}{1+k^2}(k\mathbf{n}_L - \mathbf{n}_R) + \frac{k}{1+k^2}(\mathbf{a}_L + k\mathbf{a}_R), \tag{68}$$

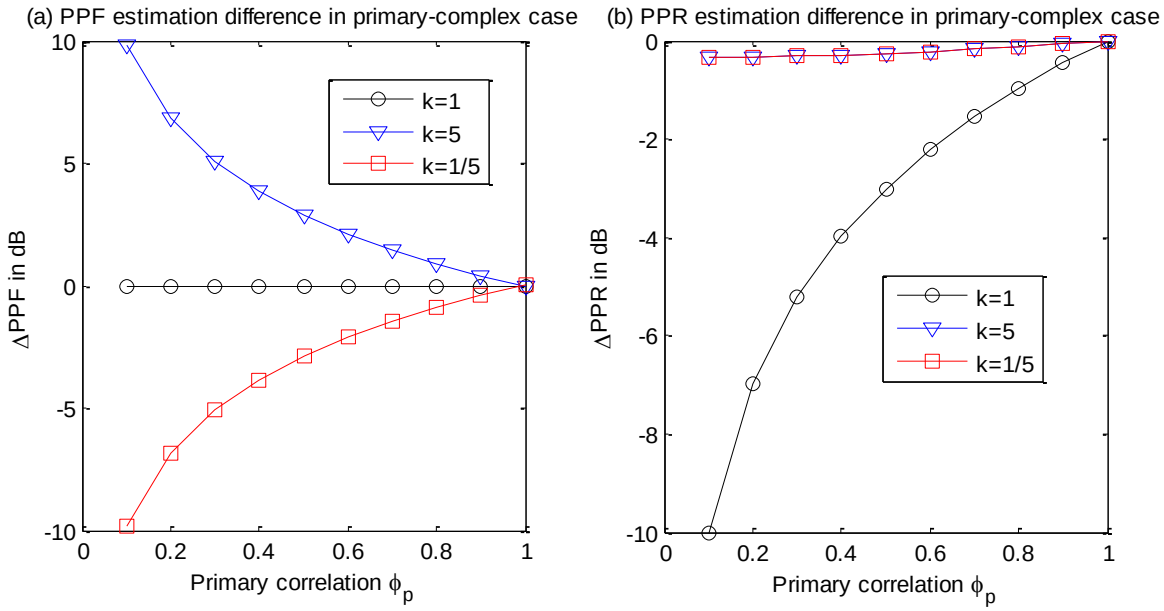

Figure 5.1 Estimation of PPF and PPR in primary-complex case.

$$\hat{\mathbf{a}}_L(\mathrm{PCA}) = \frac{k^2}{1+k^2}\mathbf{a}_L + \frac{k}{1+k^2}(k\mathbf{n}_L - \mathbf{n}_R) + \frac{-k}{1+k^2}\mathbf{a}_R, \tag{69}$$

$$\hat{\mathbf{a}}_R(\mathrm{PCA}) = \frac{1}{1+k^2}\mathbf{a}_R + \frac{-1}{1+k^2}(k\mathbf{n}_L - \mathbf{n}_R) + \frac{-k}{1+k^2}\mathbf{a}_L, \tag{70}$$

where interference signals of the primary components $\mathbf{n}_L \perp \mathbf{n}_R$ are the uncorrelated components decomposed from the input primary components $\mathbf{p}_L, \mathbf{p}_R$. These interferences, caused by the partial correlation of the primary component, incur additional error as compared to the extraction results using PCA in the ideal case (where primary component is correlated at zero lag and there is no interference). Additional leakages are found in the extracted ambient component, which are the interferences from the primary components.

The amount of primary interference power is related to primary correlation. In the following, we evaluate the performance of PCA based PAE in primary-complex case. Since there is only a panning factor difference between primary component in the left and right channel, we can easily compute the measures for the two channels by using the PPF and its reciprocal, respectively. Here, we summarize the results of the performance measures of the PCA based PAE in primary-complex case obtained with the left channel in Table 5-1.

It is clear from Table 5-1 that the accuracy measures (ESR and ETSC) are affected by the correlation of the primary components ϕ_p. Within the three specific error measures, ISR for the extracted primary component and LSR for the extracted ambient component are affected by ϕ_p, and hence cause the degradation of the extraction performance. To illustrate how the extraction accuracy is influenced by ϕ_p, six measures in Table 5-1 having PPF = 3 and PPR at [0.2, 0.5, 0.8] are plotted in Fig. 5.2. Our observations on the extraction accuracy are as follows. In the ideal case where $\phi_p = 1$, the primary and ambient components are extracted with relative little error and high similarity to original primary component. As ϕ_p decreases, the extracted primary and ambient components are found to possess more errors and lower similarity for all values of PPR. It is also found from 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 extraction. The ICLD of the primary component can only be retained when PPF k is accurately estimated. Therefore, it is concluded that the performance of PCA based PAE is degraded when primary component becomes partially correlated at zero lag.

5.2 Time shifted PCA in primary-ambient extraction

Many primary components in most stereo audio signals are time-shifted and amplitude panned and such primary components have low correlation at zero lag. In the previous subsection, it is shown that the performance of PAE is considerably degraded by the low correlation of primary component at zero lag. Moreover, the ICTD of the primary component is completely lost after the extraction

Table 5-1 Performance of PCA based PAE in primary-complex case.

Measures	Primary component	Ambient component
ESR	$\frac{1-\gamma}{2\gamma}+\frac{2k^4}{\left(1+k^2\right)^2}\left(1-\phi_p\right)$	$\frac{2k^4}{\left(1+k^2\right)^3}\frac{2\gamma}{1-\gamma}\left(1-\phi_p\right)+\frac{1}{1+k^2}$
LSR	$\frac{1-\gamma}{2\gamma}$	$\frac{2k^4}{\left(1+k^2\right)^3}\frac{2\gamma}{1-\gamma}\left(1-\phi_p\right)$
ISR	$\frac{2k^4}{\left(1+k^2\right)^2}\left(1-\phi_p\right)$	$\frac{k^2}{\left(1+k^2\right)^2}$
DSR	0	$\left(1+k^2\right)^{-2}$
ETSC	$\sqrt{\frac{4\gamma}{\left[2+\left(1+k^2\right)\beta\right]\left(1+\beta+\gamma\right)}}$	$\sqrt{\left(1-\frac{\gamma\beta}{1+\beta-\gamma}\right)\frac{k^2}{1+k^2}}$
ICLD	k^2	k^{-2}
ICC(ICTD)	1(0)	1

where $\beta=\sqrt{1+\left(\frac{2k}{1+k^2}\right)^2\left(\frac{1}{\phi_p{}^2}-1\right)}-1.$

process. To overcome these issues, we propose a novel time-shifted PCA (SPCA) based PAE. Some preliminary results are reported in [35]. A block diagram of the proposed SPCA based PAE is shown in Fig. 5.3. This approach aims to retain the ICTD in the extracted primary component, and to enhance the performance of PAE as the primary component becomes more correlated. In SPCA based PAE, the stereo input signal is first time-shifted according to the estimated ICTD of the primary component before the PAE operation. Subsequently, the extracted primary and ambient samples are then time-shifted back using the estimated ICTD. Suppose the input signal is shifted in the right channel and denote the ICTD as τ, we can express the expressions for each sample in the extracted components as

$$\hat{p}_L(\mathrm{j})=\frac{1}{1+k^2}\left[x_L(j)+kx_R(j-\tau)\right],\ \hat{p}_R(\mathrm{j}-\tau)=\frac{k}{1+k^2}\left[x_L(j)+kx_R(j-\tau)\right]. \tag{71}$$

$$\hat{a}_L(\mathrm{j})=\frac{k}{1+k^2}\left[kx_L(j)-x_R(j-\tau)\right],\ \hat{a}_R(\mathrm{j}-\tau)=-\frac{1}{1+k^2}\left[kx_L(j)-x_R(j-\tau)\right]. \tag{72}$$

Based on Jeffress's model [40], the lag number corresponds to the maximum ICC at various lags is considered as the estimated ICTD. Note that the conventional PAE is a special case of shifted PAE when ICTD is zero. Actually, the ICC of the stereo input signal is used as a representative of ICC of the primary component to estimate the ICTD of the primary component. The validity lies in the uncorrelated property of the ambient, which shall still be uncorrelated even if it is shifted. Thus, we can compute the shifted correlation coefficient of the mixed signal using the shifted correlations and represent it as a function of primary correlation for every lag:

$$\phi_x(\tau)=\sqrt{\frac{P_{\mathbf{p}_L}P_{\mathbf{p}_R}}{\left(P_{\mathbf{p}_L}+P_a\right)\left(P_{\mathbf{p}_R}+P_a\right)}}\phi_p(\tau)=\mathrm{A}\phi_p(\tau). \tag{73}$$

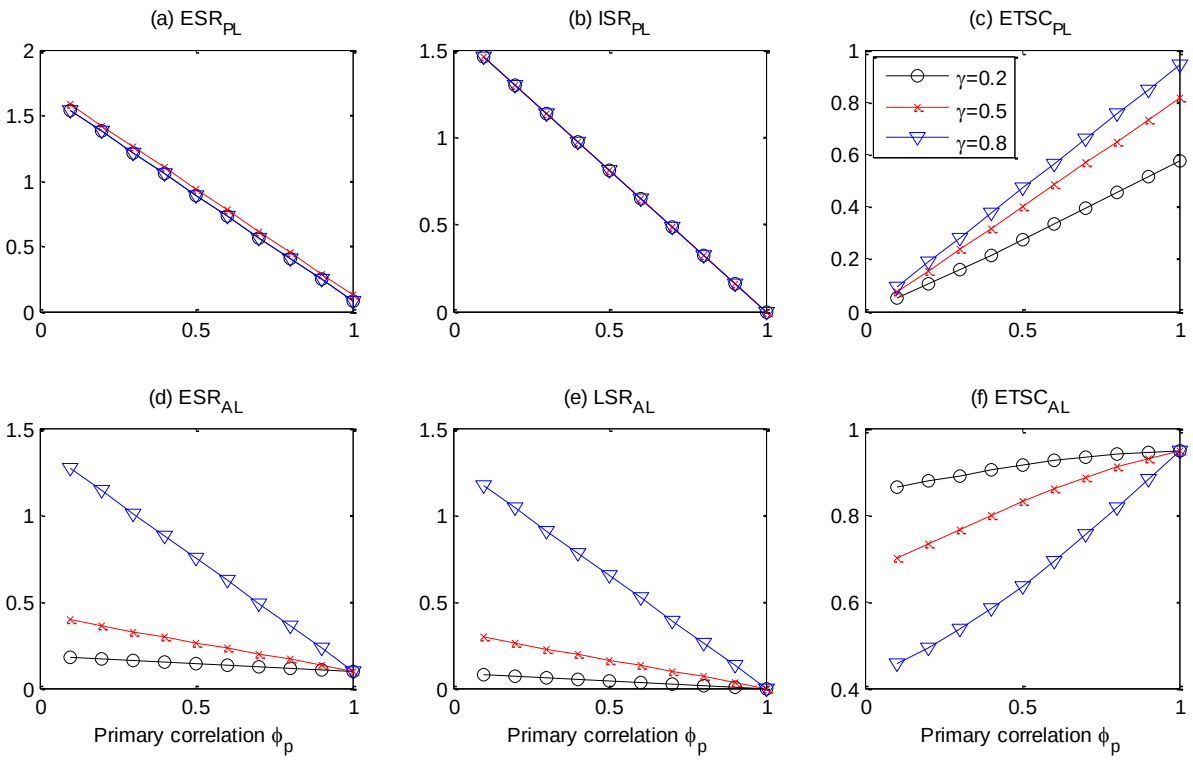

Figure 5.2 Performance of PCA based PAE in primary-complex case with varying ϕ_p according to the results in Table 5-1 (k = 3). (a) ESR_{PL}; (b) ISR_{PL}; (c) $ETSC_{PL}$; (d) ESR_{AL}; (e) LSR_{AL}; (f) $ETSC_{AL}$. Legend in (c) applies to all plots.

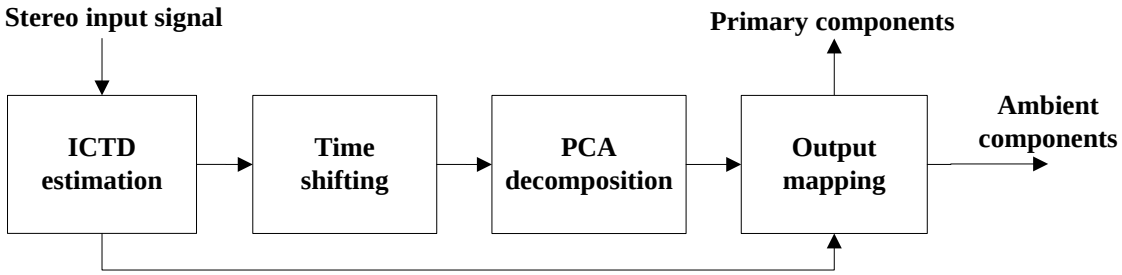

Figure 5.3 Block diagram of shifted PCA based primary component extraction.

Since the power of the primary or ambient component does not change much after time-shifting, A is usually a constant. Therefore, the lag number τ_o that corresponds to the maximum ICC of the mixed signal $\phi_x(\tau)$ would be the lag number for the ICC of the primary component $\phi_p(\tau)$ to achieve its maximum correlation. Hence, the ICTD of the primary component is estimated as this lag number.

Time-shifting of the input stereo signal can be a very effective pre-processing technique to retain the ICTD in the primary component, and enhance the PAE performance for the primary- complex case, assuming that the ICTD is correctly estimated. From Lord Rayleigh's Duplex theory, it is well known that ICTD and ICLD together account for the localization of the sound. Hence, we can inter-adjust and inter-calibrate the estimation of ICTD and ICLD for different scenarios. A detailed study on the estimation of ICTD and ICLD in complex cases is discussed in [46].

When the lags shifted in two successive blocks are not the same, proper mapping strategy are required in output mapping for the extracted primary or ambient components. Normally we will encounter overlapped samples when ICTD in the successive block is smaller and gapped samples when ICTD becomes larger, as shown in Fig. 5.4(a) and 5.4(b), respectively. To maintain the ICTD, a straightforward mapping method is to leave the gapped samples to zero but averaging the overlapped samples in a cross-fading manner. However, switching artifacts are perceived in the gapped fields but hardly found in the overlapped fields after some testing. Therefore, all the successive blocks shall be overlapped to avoid the switching noise. To achieve this, the hop size of the

block shall be set smaller than the block size such that no gapped samples can be found even when the ICTD increase is maximum. An illustration of the proposed overlapped output mapping strategy for the extracted components is depicted in Fig. 5.5. We shall find that however the ICTD changes, the two successive blocked are ensured to be non-gapped.

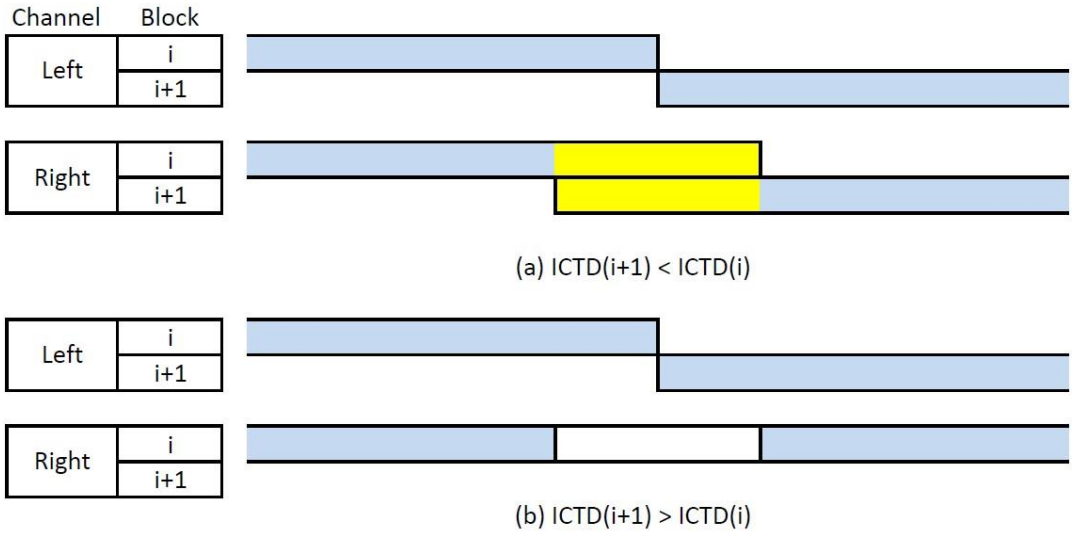

Figure 5.4 Conventional output mapping strategy.

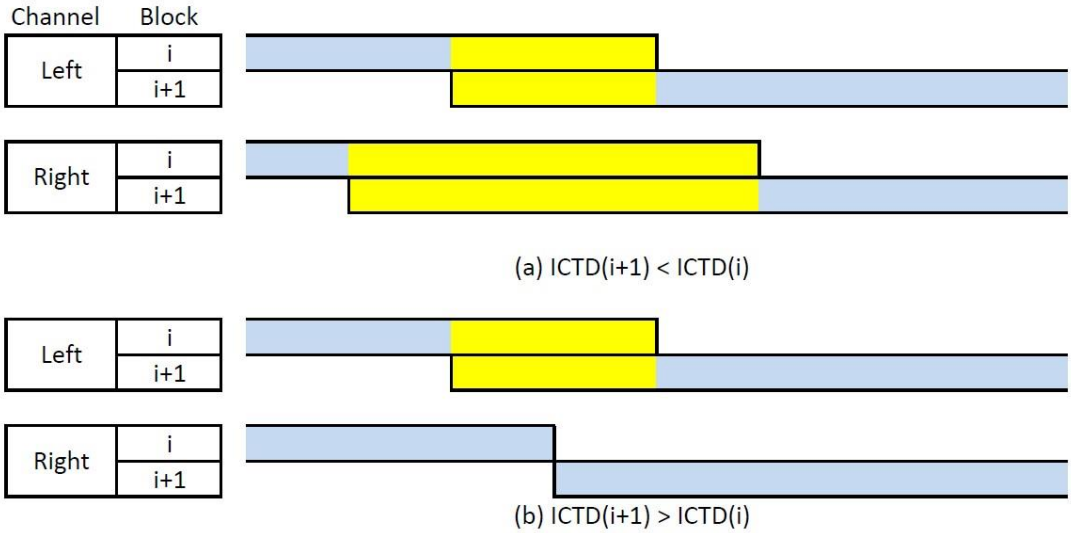

Figure 5.5 Proposed overlapped output mapping strategy.

5.3 Experimental results

To validate the performance of the proposed SPCA based PAE, a number of simulation and subjective listening experiments are conducted. The results for the shifted PCA shall be discussed in the following. Some other test results can be found in [47]. Again, a speech is selected as the primary component which is amplitude panned by a factor of 3 and time shifted by 40 time units, both to the right channel, while ambience is uncorrelated and balanced white Gaussian noise. Subsequently, the primary and ambient components are linearly mixed based on PPR, ranging from 0 to 1. PCA and shifted PCA are employed to extract primary and ambient components from the synthesized stereo signal. Note that the normalized correlation of the tested primary component at zero lag is 0.1676, which is increased to one after shifting the mixed signal according to estimated ICTD. The unity correlation implies that the primary component is completely correlated in shifted PCA.

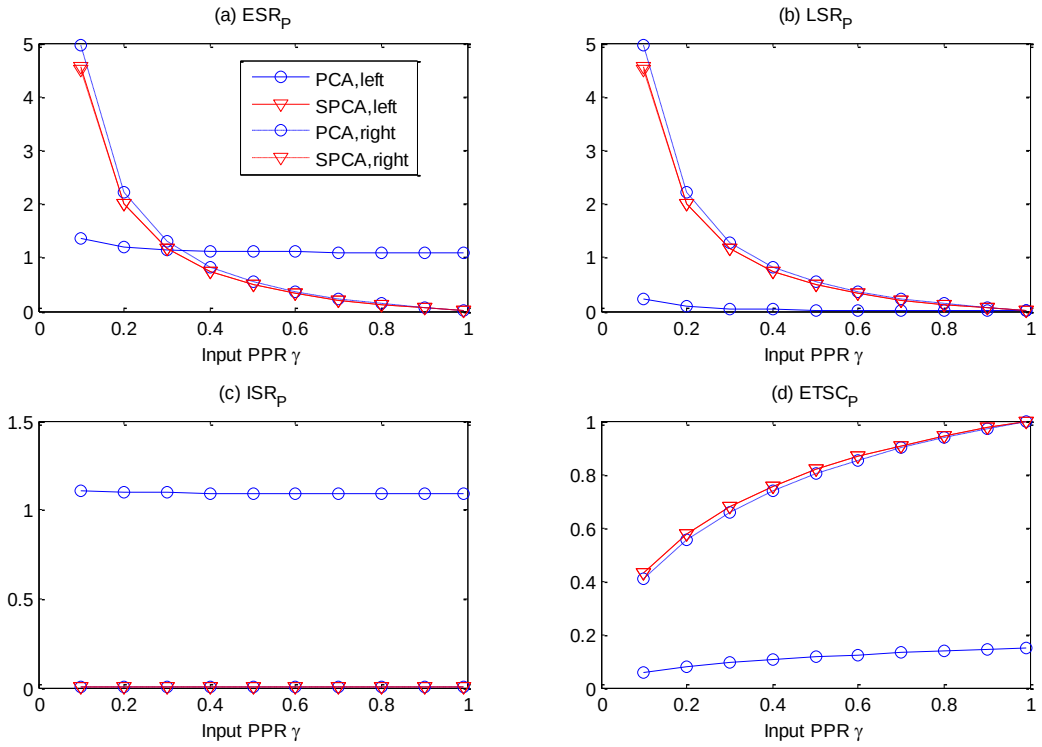

Figure 5.6 Performance comparison of primary component extraction using PCA and SPCA. (a) ESR, (b) LSR, (c) ISR, (d) ETSC. The legend in (a) applies to all plots.

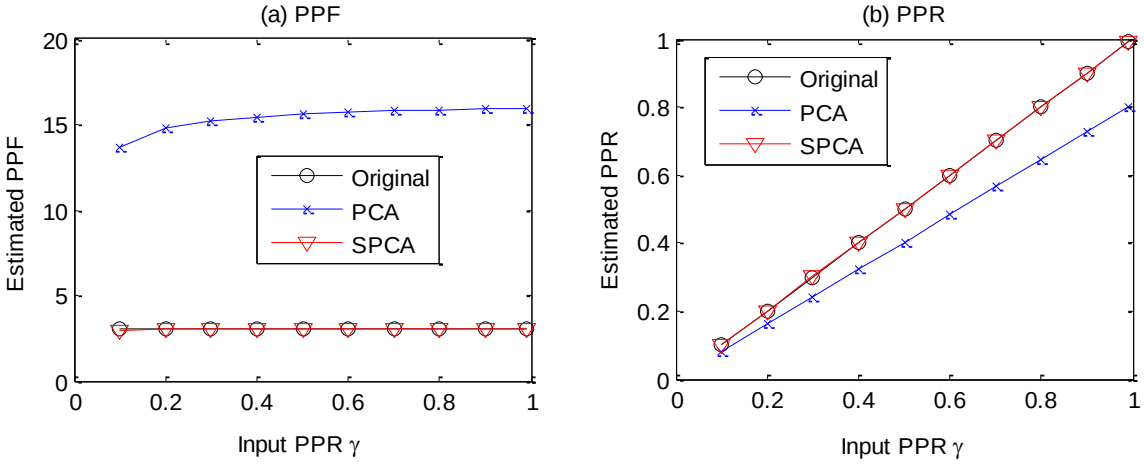

Figure 5.7 Comparison of the estimation of PPF and PPR between PCA and SPCA in primary-complex case.

The performance measures of PCA and shifted PCA are shown in Figs. 5.6-5.9. In Fig. 5.6, there is significant error in the estimation of PPF and PPR in PCA, which is correctly extracted in the shifted PCA. Figure 5.7 summarizes the performance measure of primary component extraction using PCA and shifted PCA. Although the measures for the right channel are highly similar between PCA and shifted PCA, significant reduction of ESR and a higher ETSC are found in the left channel for shifted PCA when PPR is high ($\gamma \geq 0.5$). The improvement of shifted PCA in primary component extraction lies in the removal of interference error as shown in Fig. 5.8(c), though its ambient leakage in the left channel is slightly more than the case in PCA, as shown in Fig. 5.8(b).

Based on the observations of Fig. 5.9, shifted PCA extracts the ambience with less error (ESR) and higher similarity (ETSC) than PCA in the left channel, while in the right channel, the performance of shift PCA is very close to PCA. As can be found in Fig. 5.9 (a)-(c), the improvement lies in the reduction of primary leakage (LSR) though the interference (ISR) increases a bit. In terms of

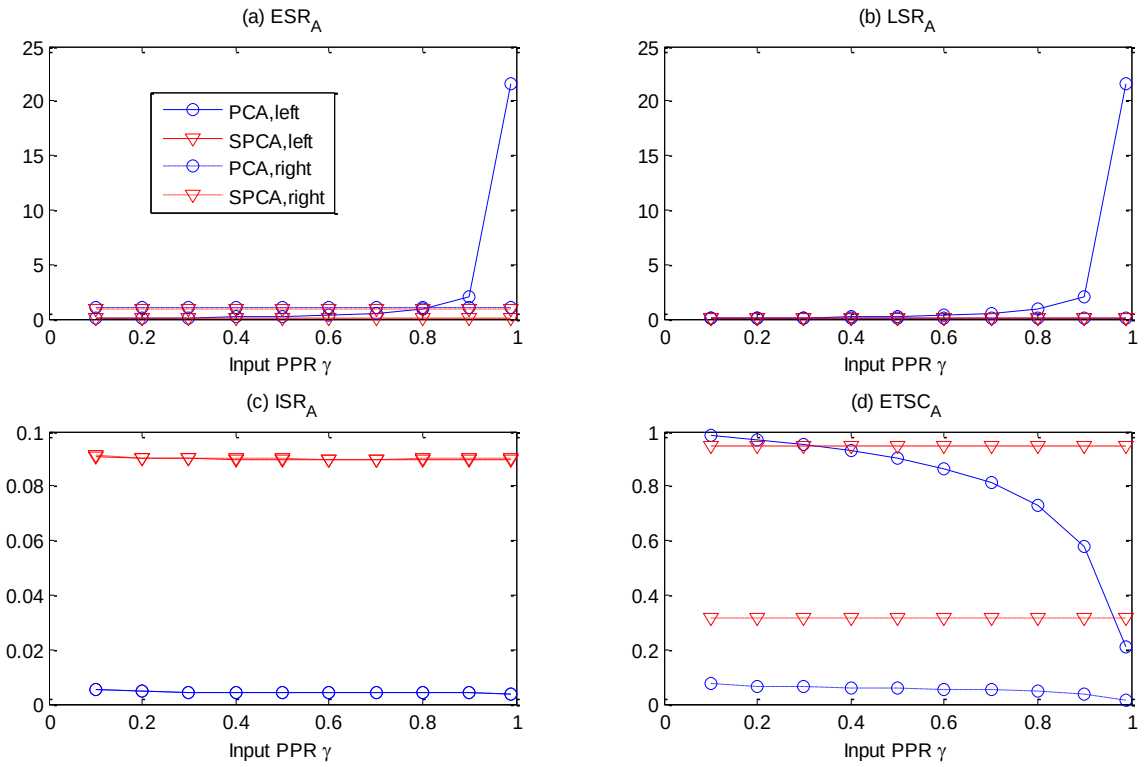

Figure 5.8 Performance comparison of ambient extraction using PCA and SPCA. (a) ESR, (b) LSR, (c) ISR, (d) ETSC. The legend in (a) applies to all plots.

the spatial feature of the extracted primary component, shifted PCA outperforms PCA by extracting primary component having ICTD and ICLD values closer to the true primary component, as shown in the Fig. 5.9. As in the ideal case, both approaches fail to extract an uncorrelated and balanced ambience. This drawback lies in the limitation of PCA, which is mathematically a linear transformation. Therefore, some post-processing techniques like decorrelation [41] and post-scaling [20] can be applied to further render the extracted ambience.

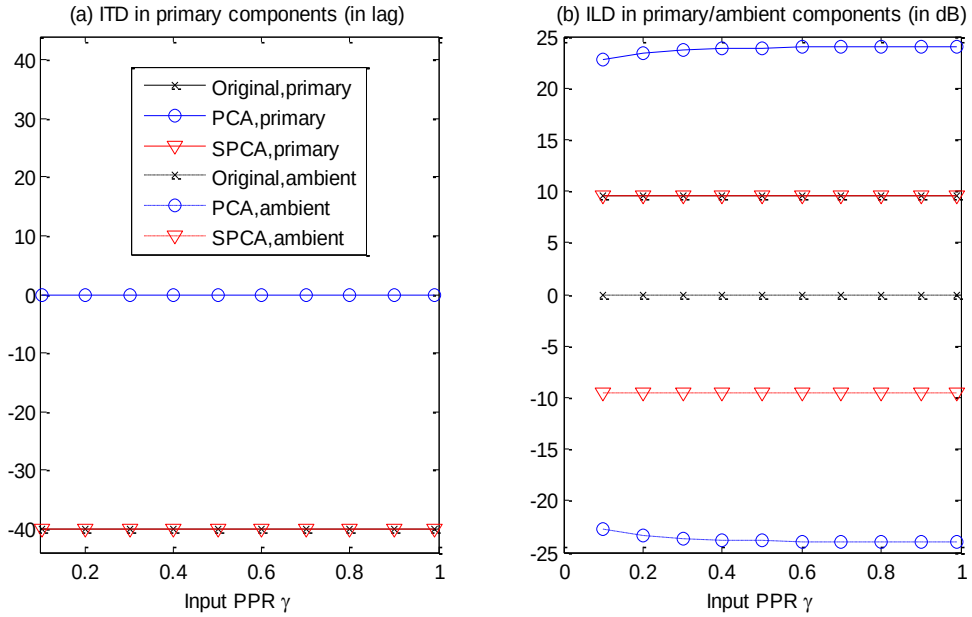

Figure 5.9 Performance comparison of localization parameters in PAE using PCA and SPCA. Legend in (a) applies to all plots.

In another experiment, we tested the extraction of a panning matchbox sound that is mixed with uncorrelated noise. The plots in Fig. 5.10 present the short-time cross-correlation of the original primary component, mixed signal, and primary component extracted by PCA and primary component extracted by shifted PCA. The position of the peaks on the mesh of these plots represents the direction of the primary components. The distinct panning pattern of original primary component in Fig. 5.10(a) is lost after mixing with uncorrelated ambience, as shown in Fig. 5.10(c). Comparing the correlation of extracted primary component using PCA and shifted PCA, we can easily verify that shifted PCA preserves the spatial imagery of the primary component from the mixed stereo signal, which is not observed in PCA as shown in Fig. 5.10(b). In addition to these objective measures, our informal subjective listening experiments also revealed that shifted PCA performs better than PCA in PAE in terms of extraction accuracy and better localization of the primary components.

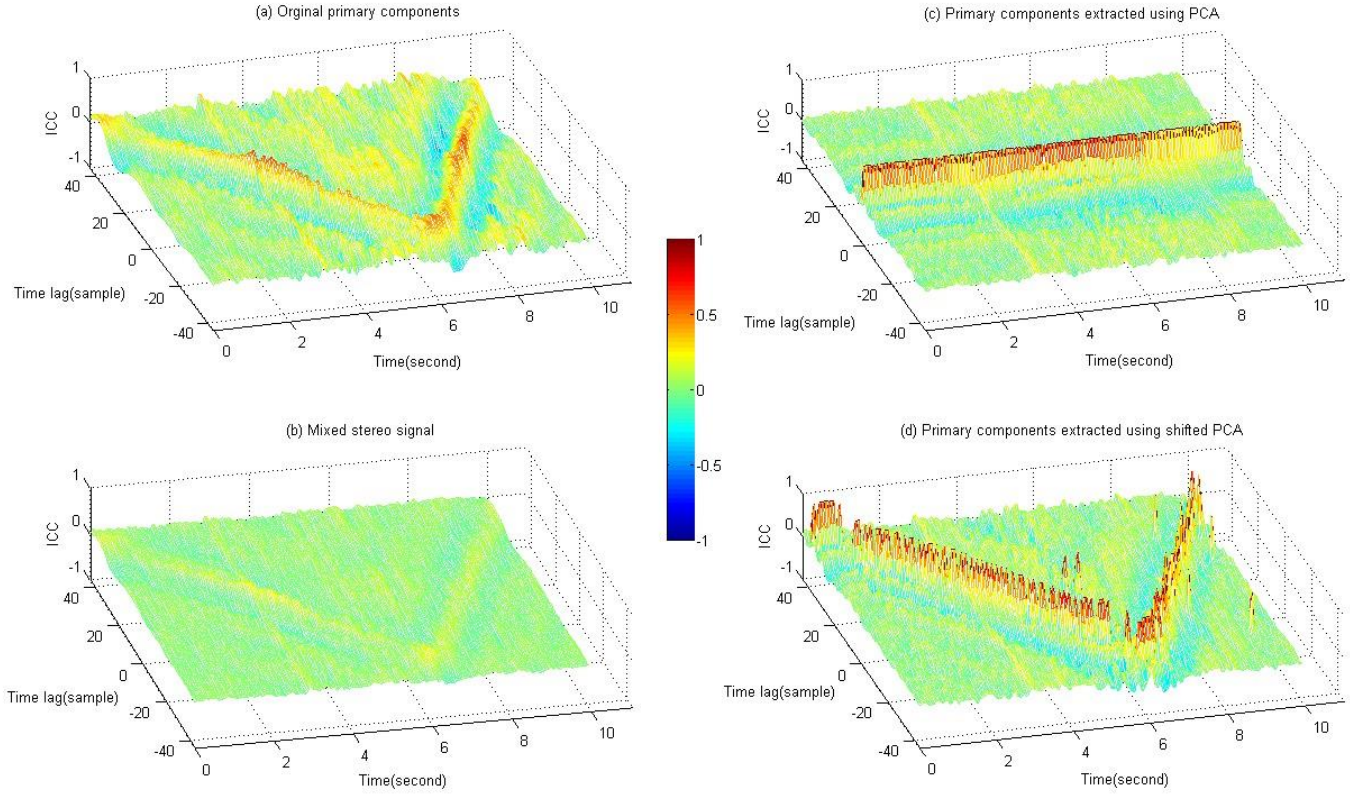

Figure 5.10 Short-time interaural cross-correlation function. (a) Original primary component; (b) stereo signal with mixed primary and ambient components; (c) extracted primary component using PCA; (d) extracted primary component using shifted PCA. Block size is 4000 samples with 50% overlap.

6. FUTURE WORK

It can be seen from current literature that primary-ambient extraction plays an important role in spatial audio to create the ultimate immersive audio experience. To extract the primary and ambient components, it requires rather a clear yet mathematically describable definition of these components. On this regard, a signal model is often introduced so that mathematical solutions can be derived and the extraction performance can further be evaluated. The signal model should be as close to the real signals (recordings, movie audio, gaming audio, etc.) as possible, since mismatch between the input signal and the signal model will lead to significant degradation of the extraction performance (as can be seen in our discussion in Chapter 5). Basically, two directions of the further research can be from the perspectives of the signal model and the real input signals.

On one hand, we are working on extending and generalizing the signal model so that the model can better match practical cases. In current work, linear estimation based PAE approaches are derived based on the basic stereo signal model discussed in Chapter 2. By taking the time difference of the primary components into consideration, we develop the time-shifted PAE approaches (e.g. SPCA), as discussed in Chapter 5. However, in many applications of spatial audio, concurrent sound sources from different directions are also frequently encountered. In such a case, SPCA which corresponds to one direction is therefore problematic. One widely used solution to resolve this problem is to apply the PAE approaches in subband, assuming that one subband accounts for one direction only. The subband approach may inherently distort the original timbre of the primary components. From another perspective, we may extend the SPCA (and other time shifted PAE approaches) from one single shift to multiples shifts, to account for multiple sound sources defined in the primary components and their associated ICTDs. This effort can lead to the development of a multi-shift PCA (MSPCA) to extract the primary and ambient components. The general structure of the MSPCA is shown in Fig. 6.1. First, several ICTDs are estimated from the stereo input signal. Next, the input signal is time-shifted according to the estimated ICTDs. For every ICTD, PCA is applied to obtain the extracted primary components from the shifted signals. Finally, the extracted primary components are accordingly mapped, weighted and summed to obtain the final output of the extracted primary components. It shall be noted that the weightings can be selected based on the significance level of each shifted version of the input signal, and the reduction of the extraction error. Optimal filtering and even adaptive filtering based approaches are of great interest in future work on this regard. Furthermore, incorporating acoustic modeling and psychoacoustic testing would help improve PAE and create a better spatial audio for human.

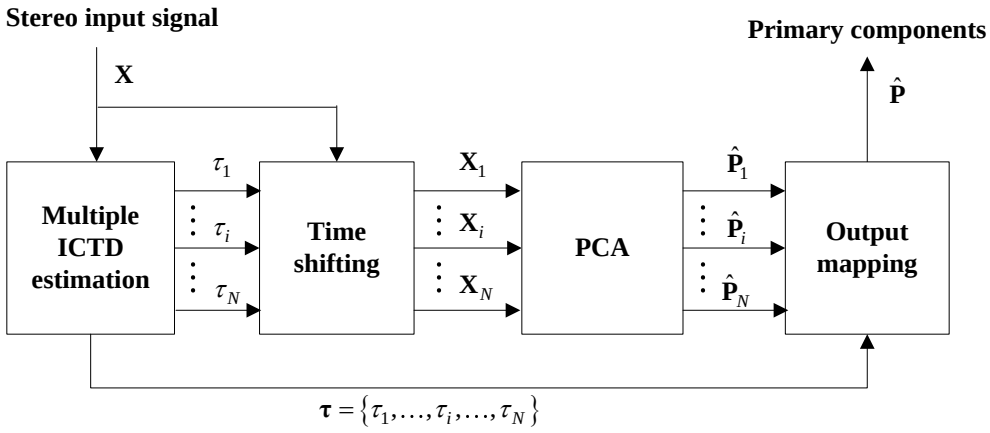

Figure 6.1 Typical structure of MSPCA (MSPCA-T). Input signal $\mathbf{X} = \{\mathbf{x}_L, \mathbf{x}_R\}$; τ_i is the *i*th estimated ICTD; $\mathbf{X}_i$ and $\hat{\mathbf{P}}_i$ are the corresponding shifted signal and extracted primary components, respectively. The final output of the extracted primary components is denoted by $\hat{\mathbf{P}}$.

On the other hand, gaining more information on the input signal will provide us a better understanding of the primary and ambient components and a more accurate stereo model can be developed. This requires further study of the audio features [48], [49] of the input signal as well as the statistical or mathematical interpretation of the primary and ambient components, which may change in different circumstances [50]. For examples, in stereo signal case, the primary and ambient components are discriminated based on correlation between the left and right channels. In the case of multichannel signals, this needs to be adjusted and requires more consideration [39]. However, this definition will not hold for single channel case as there is no inter-channel relationship. On this note, intra-channel differences are to be studied and it requires some different decomposition like non-negative matrix factorization [38]. With different cases identified, appropriate approaches can be applied in PAE for arbitrary input.

From the relation of PAE problem with other research problems, we may gain some more insights. For example, independent component analysis (ICA) can be used to separate independent primary components from the ambience. When only the extraction of primary components is of interest, some noise reduction approaches might be applied. In the case ambience of greater interest such as the extraction of reverberation, the method to model the reverberation can be employed to extract ambience. Sound localization techniques, though are only interested in the direction of the sound sources, can also be used hand in hand with PAE. Furthermore, acoustic modeling and psychoacoustic testing would help improve PAE and create a better spatial audio for human.

In summary, we can see that to design a robust PAE algorithm, there is still much research to be conducted. A brief illustration of the PAE processing is shown in Fig. 6.2. Generally, a detection and classification process is required to gain enough knowledge about the blind input signal. After that, some pre-processing techniques are applied. In the pre-processing stage, scaling, time shifting, grouping, and even transformation techniques are often found to be useful. Subsequently, appropriate PAE approaches shall be applied where some self-examination and adjustment can be included. For a complete spatial audio system, post-processing techniques are usually essential depending on the application. Normally, accurate localization of the sound events and realistic rendering of the sound environment are basic requirements of a spatial audio system. Additional requirements may be introduced by various applications. For example, a better front-back discrimination and externalization shall be the target of 3D headphone system. For loudspeaker system like i3D, crosstalk cancellation and sound level adjustment should be considered.

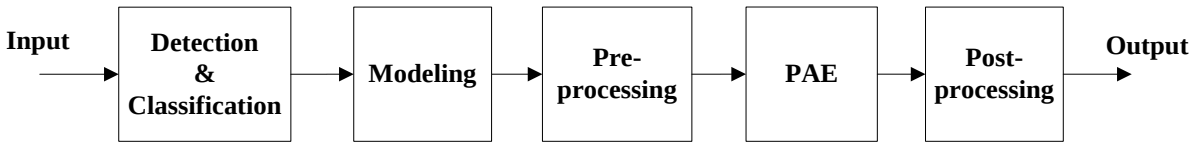

Figure 6.2 Block diagram of a complete and robust PAE based 3D audio system.

7. CONCLUSIONS

In this report, we presented a brief review of the primary-ambient extraction approaches applied in spatial audio. We found that PAE are essential to create an immersive 3D audio experience. In our current work, we revisited the problem of PAE of stereo signals using linear estimation approaches. Based on the stereo signal model, we formulated the PAE problem as a weighting matrix estimation framework. Under this framework, we discussed three groups of performance measures and derived the optimal solutions for PCA, LS, MLLS, MDLS, and ALS, as well as their relationship and differences among these approaches. In particular, PCA was found to be identical with MDLS in terms of extracting primary components with minimum distortion; and LS is identical with MLLS in its ability to extract primary component with minimum extraction error. The ambient leakage in the extracted primary components is minimized for all four approaches. The only difference between extracted primary components derived from PCA and LS is a scaling factor, which is solely related to PPR. All the approaches discussed in this report yield better primary component extraction when PPR is higher, which is on the contrary for ambience extraction. In ambience extraction, PCA (and also MLLS), LS, and MDLS minimize the relative leakage, error, and distortion, respectively. Based on our discussions in this report, different PAE approaches can be applied depending on the characteristics and requirements of the application.

While accurate extraction of primary and ambient components using PCA was found in the ideal case. The conventional PAE approaches exhibited severe performance degradation in primary-complex case. The estimations for PPF and PPR were derived with the additional priori knowledge of primary correlation. However, the primary correlation is usually unavailable. The proposed shifted PAE method overcomes the problem by appropriately time-shifting input signals prior to the conventional PCA based PAE. This technique corrects the ICTD and ICLD of the extracted primary component, and reduces the extraction error as well as increasing the extraction similarity. These theoretical results are found to be consistent with our simulation results and informal subjective listening experiments. Moreover, the shifted PCA based PAE is simple, efficient, and yet effective.

Finally, we discussed the future research work that shall be continued in the context of designing a better and more robust PAE method. We also find audio recognition, classification, pre-processing, and post-processing techniques are essential in producing a better spatial audio.

REFERENCES

[1] D. S. Brungart, 3D sound for virtual reality and multimedia, Academic Press Professional, Cambridge, MA, USA, 2000.

[2] J. Blauert, Spatial hearing: the psychophysics of human sound localization. Cambridge, MA: MIT Press, 1997.

[3] Mills, W. (1972). Auditory localization. In J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory*. New York: Academic Press.

[4] W.G.Gardner and K.D.Martin, "HRTF Measurements of a KEMAR Dummy Head Microphone" *MIT Media Lab Perceptual Computing Section*, Tech. Rep.280 (1994).

[5] J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. on audio, speech and language process.*, vol.16, no. 8, Nov. 2008.

[6] E. L. Tan, W. S. Gan, and C. H. Chen, "Spatial sound reproduction using conventional and parametric loudspeakers," *in Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific* , Hollywood, CA, Dec. 2012.

[7] H.Møller, M.F.Sørensen, C.B.Jensen, and D.Hammershøi, "Binaural technique: do we need individual recordings ?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451-469, 1996.

[8] E. M. Wenzel, M. Field, 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, 1993.

[9] K. Sunder, E. L. Tan, and W. S. Gan, "On the study of frontal-emitter headphone to improve 3D audio playback", in 133rd *Audio Eng. Soc. Conv.*, San Francisco, Oct. 2012.

[10] M. M. Goodwin and J. M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," *in IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Hawaii, Apr. 2007.

[11] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching", in *128th Audio Eng. Soc. Conv.*, London, UK, May. 2010.

[12] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001.

[13] J. Breebaart and C. Faller, *Spatial audio processing: MPEG surround and other applications*. Chichester, UK: John Wiley & Sons, 2007.

[14] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503-516, Jun. 2007.

[15] M. M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," *in 123rd Audio Eng. Soc. Conv.*, New York, Oct. 2007.

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

[17] S. Y. Park, S. Lee, and D. Youn, "Robust representation of spatial sound in stereo-to-multichannel upmix," *in 128th Audio Eng. Soc. Conv.*, London, UK, May 2010.

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

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

[20] 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.

[21] M. M. Goodwin and J. M. Jot, "Spatial audio scene coding," *in 125th Audio Eng. Soc. Conv.*, San Francisco, Oct. 2008.

[22] M. A. Gerzon, "General metatheory of auditory localization," *in 92nd Audio Eng. Soc. Conv.*, Vienna, Austria, Mar. 1992.

[23] V. Pulkki, "Virtual source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[24] 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.

[25] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York: John Wiley & Sons, 2001.

[26] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio mixer," *IEEE Tran. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2141-2150, Sept. 2007.

[27] 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.

[28] M. Goodwin, "Geometric signal decompositions for spatial audio enhancement," in *IEEE Intl. Conf. on Acoust., Speech, and Signal Process.*, Las Vegas, April 2008.

[29] J. Merimaa, M. M. Goodwin, J. M. Jot, "Correlation-based ambience extraction from stereo recordings", *in 123rd Audio Eng. Soc. Conv.*, New York, Oct. 2007.

[30] 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.
[31] Y. H. Baek, S. W. Jeon, Y. C. Park, and S. Lee, "Efficient primary-ambient decomposition algorithm for audio upmix," *in 133rd Audio Eng. Soc. Conv.*, San Francisco, Oct. 2012.
[32] M. Briand, D. Virette and N. Martin, "Parametric representation of multichannel audio based on principal component analysis," in *120th Audio Eng. Soc. Conv.*, Paris, May 2006.
[33] J. Se-Woon, H. Dongil, S. Jeongil, P. Young-Cheol, and Y. Dae-Hee, "Enhancement of principal to ambient energy ratio for PCA-based parametric audio coding," *in IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Dallas, March 2010.
[34] D. Shi, R. Hu, W. Tu, X. Zheng, J. Jiang, and S. Wang, "Enhanced principal component using polar coordinate PCA for stereo audio coding," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, Melbourne, VIC, July 2012.
[35] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," *in IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Vancouver, Canada, May 2013.
[36] I. Jolliffe, *Principal component analysis, 2nd ed*.. New York: Springer-Verlag, 2002.
[37] C. Faller, "Multiple-loudspeaker playback of stereo signals", *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051-1064, Nov. 2006.
[38] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using non-negative matrix factorization", in *30th Audio Eng. Soc. Int. Conf.*, Saariselka, Finland, Mar. 2007.
[39] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," in *133rd Audio Eng. Soc. Conv.*, San Francisco, Oct. 2012.
[40] A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35-39, Feb. 1948.
[41] 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.
[42] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation" *IEEE Tran. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.
[43] Y. Ando, and P. Cariani, *Auditory and Visual Sensation*. New York: Springer, 2009
[44] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[45] C. Faller, "Parametric multichannel audio coding: synthesis of coherence cues," *IEEE Tran. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 299-310, Jan. 2006.
[46] J. Woodruff and D. L. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Tran. Audio, Speech Lang. Process.*, vol. 20, no. 5, pp. 1503-1512, Jul. 2006.
[47] J. He. (2012 Nov. 30). A comparative testing of cue extraction using PCA and shifted PCA [Online]. Available: http://eeeweba.ntu.edu.sg/dsplab/audiobeam/jhe007/.
[48] Y. Wang, Z. Liu, and J. Huang, "Multimeida content analysis using both audio and visual cues," *IEEE Sig. Process. Mag.*, vol. 17, no. 6, pp. 12-36, Nov. 2000.
[49] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Tran. Audio, Speech Lang. Process.*, vol. 10, no. 7, pp. 504-516, Oct. 2002.
[50] A. Härmä, "Classification of time-frequency regions in stereo audio," *J. Audio Eng. Soc.*, vol. 59, no. 10, pp. 707-720, Oct. 2011.