Applying Primary Ambient Extraction for Immersive Spatial Audio Reproduction

Jianjun He, and Woon-Seng Gan
Digital Signal Processing Lab, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore.
E-mail: jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg

***Abstract*—Spatial audio reproduction is essential to create a natural listening experience for digital media. Majority of the legacy audio contents are in channel-based format, which is very particular on the desired playback system. Considering the diversity of today's playback systems, the quality of reproduced sound scenes degrades significantly when mismatches between the audio content and the playback system occur. An active sound control approach is required to take the playback system into consideration. Primary ambient extraction (PAE) is an emerging technique that can be employed to solve this pressing issue and achieve an efficient, flexible, and immersive spatial audio reproduction. In this paper, we will review the recent advancements in PAE. A unified framework to extend existing PAE approaches is proposed to improve the robustness of PAE when dealing with complex audio signals in practical situations. Various practical issues on implementing PAE in spatial audio applications are discussed. Objective and subjective evaluations are conducted to validate the feasibility of applying PAE in spatial audio reproduction.**

I. INTRODUCTION

Sound is an inherent part of our everyday lives for information, communication and interaction [1]. The role of 3D sound, or spatial sound, in high stress applications, like flight navigation and communication systems, is indisputable [2]. Spatial sound has also been proven to be beneficial in personal route guidance for visually impaired people [3] and in medical therapy for patients [4], [5]. Last but not least, the ever growing market of consumer electronics calls for spatial audio reproduction for digital media, virtual reality (e.g., Oculus Rift), augmented reality (e.g., Microsoft HoloLens).

Considering the variety of applications, spatial audio reproduction of digital media (especially the movies and games) has gained significant popularity over the recent years [6]. As introduced by Begault [1], the process of sound reproduction can generally be considered as a source-medium-receiver model [7]. The source for spatial audio reproduction refers to the audio content, which could be represented in channel-based, object-based, and transform-domain-based formats [8], [9]. The medium refers to the audio playback systems, which could generally be classified into loudspeaker systems, headphones, or even hybrid or special systems. Finally, our ears play the role of the receiver that captures the incoming sound, while performing an individualized filtering of the sound since no two human ears are the same. Creating natural immersive spatial sound requires consistence among the source, medium and receiver, as illustrated in Fig. 1. In this paper, we focus on the mismatch problems between the source and medium.

Despite the growing interest in object-based audio [6], such as Dolby Atmos [10], DTS X [11], MPEG-H [8], most existing audio content is still in channel-based formats (such as stereo and multichannel signals). The channel-based audio, being specific in its playback configuration, does not support flexible playback configurations in domestic or personal listening circumstances [6]. Considering the wide diversity of today's playback systems [11], it becomes necessary to process the audio signals such that the reproduction of the audio content is not only compatible with various playback systems but also able to achieve the best quality (especially spatial quality [12]) with the actual playback system [13].

Depending on the actual playback system, the challenges in spatial audio reproduction can be broadly categorized into two main types: loudspeaker playback and headphone playback [14]. The challenge in loudspeaker playback lies in the mismatch in the number [15] and even the type [16]-[18] between the intended and the actual loudspeaker system. Conventional techniques to solve this challenge are often referred to as audio remixing (i.e., down-mix and up-mix), for example, "Left only, Right only (LoRo)", matrix-based mixing surround sound systems, etc [15], [19]-[21]. These audio remixing techniques basically compute the loudspeaker signals as the weighted sums of the input signals. For headphone playback, the challenge arises when the audio content is not tailored for headphone playback (usually intended for loudspeaker playback). Virtualization is often regarded as the technique to solve this challenge [1], where virtualization of loudspeakers is achieved by binaural rendering technique that convolves the channel-based signals with head-related impulse responses (HRIRs) of the corresponding loudspeaker positions. These conventional techniques in spatial audio reproduction are capable of solving the compatibility issue, but the spatial quality of the reproduced sound scene is usually limited [19], [22]-[24]. To improve the spatial quality of the sound reproduction, MPEG Surround and related techniques were introduced, which usually employ the one-channel down-mixed signal and the subband spatial cues. It is found that such techniques better suit the reproduction of the distinct directional source signals as compared to the diffuse signals [23], [26].

To further improve the quality of the reproduced sound scene, the perception of the sound scenes is modeled as a

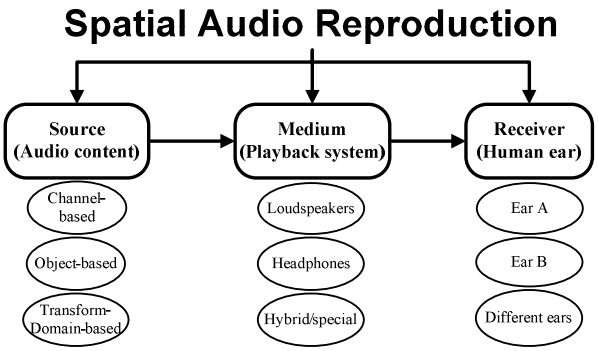

Fig. 1 Source-medium-receiver view of spatial audio reproduction

combination of the foreground sound and background sound [27], which are often referred to as primary (or direct) and ambient (or diffuse) components, respectively [28]-[30], [7]. The primary components consist of point-like directional sound sources, whereas the ambient components are made up of diffuse environmental sound, such as the reverberation, applause, or nature sound like waterfall [26], [31]. Due to the perceptual differences between the primary and ambient components, different rendering schemes should be applied to the primary and ambient components for optimal spatial audio reproduction of sound scenes [26], [32]. However, channel-based audio provides only the mixed signals [33], which necessitate the process of extracting primary and ambient components from the mixed signals (a.k.a., PAE).

As a spatial audio processing tool [15], [28], [26], [23], [7], [38], PAE has also been incorporated into spatial audio coding systems, such as spatial audio scene coding [28], [33], and directional audio coding [34]. Essentially, PAE serves as a front-end to facilitate flexible, efficient, and immersive spatial audio reproduction, as illustrated in Fig. 2. First, by decomposing the primary and ambient components of the sound scene, PAE enables the sound reproduction format to be independent of the input format, hence increasing the flexibility of spatial audio reproduction [33], [35]. Second, PAE based reproduction of sound scenes does not require the individual sound objects as in object-based format (which is the most flexible), but is able to recreate perceptually similar sound scenes, hence maintaining the efficiency of spatial audio reproduction [29]. Last but not least, PAE extracts the two key components of the sound scenes, namely, directional and diffuse sound components. These components are highly useful in recreating an immersive listening experience of the sound scene [28], [36]-[39].

Fig. 3 illustrates the PAE based spatial audio reproduction system, where the primary and ambient components undergo different rendering schemes [40]. The rendering schemes differ for loudspeaker or headphone playback [31], [36], [41]. For loudspeaker playback, the primary components are reproduced using vector base amplitude panning [42] or vector base intensity panning [43], [44] to reproduce the accurate direction of the sound sources. The ambient components, on the other hand, are further decorrelated and

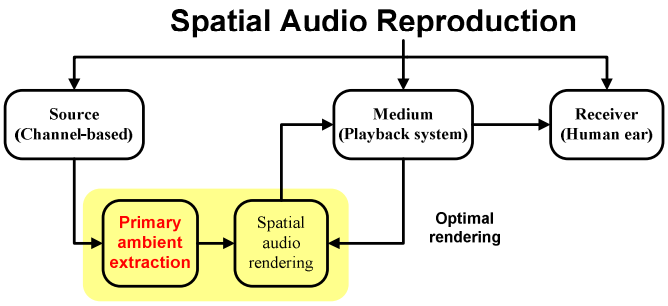

Fig. 2 Achieving consistency in source and medium in spatial audio reproduction

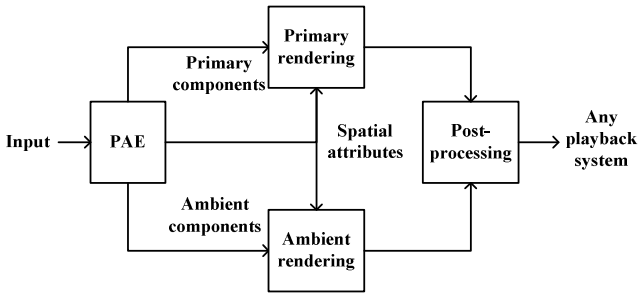

Fig. 3 Block diagram of PAE based spatial audio reproduction

distributed to all the loudspeaker channels to create an envelopment effect of the sound environment [28], [45]. The PAE based loudspeaker system is particularly suitable for active sound control using a novel hybrid loudspeaker consisting of conventional loudspeakers and directional loudspeaker [16]-[18], [83], especially considering the recent advances of directional loudspeaker [80]-[83]. For headphone playback, PAE based virtualization applies binaural rendering to the extracted primary components, creating accurate virtual sound sources in the desired directions [28], [7], [46]. Similar to the loudspeaker playback case, the ambient components are decorrelated using artificial reverberation [23], [28], [37], [32] to create a more natural sound environment.

The remainder of this paper is structured as follows. Section II presents a detailed review of prior work in PAE. In Section III, a unified framework that extends basic PAE approaches to handle complex input signals is proposed. Evaluation framework and experimental results are discussed in Section IV. In Section V, we conclude this paper.

II. Prior Work in Primary Ambient Extraction

In this section, we will summarize existing works on PAE. As discussed above, the target audio content of PAE is channel-based signals. On this note, we classify the PAE approaches based on the number of channels in the input signals: single channel, stereo, and multichannel. From another perspective, the complexity of the audio scenes affects the performance of PAE drastically. Based on the existing PAE work, the complexity of audio scenes can generally be classified into three levels, namely, basic, medium, and complex. The basic complexity level refers to the audio scene where there is

Table I An overview of recent work in PAE

No. of channels	Complexity of audio scenes		
	Basic (single source, only amplitude panning)	Medium (single source)	Complex (multiple sources)
Stereo	Time frequency masking: [53], [31], [49], [34] PCA: [54]-[58], [49], [26], [17]-[19], [46], [29] Least-squares: [45], [38], [36], [41], [29], [59] Ambient spectrum estimation: [60], [61] Others: [22], [32], [62]	LMS: [37] Shifted PCA: [63] Time shifting: [64]	PCA: [65], [40], [66]
Multichannel	PCA: [26] Others: [48], [67], [18], [68]	ICA and time-frequency masking: [69] Pairwise correlations: [70] Others: [27]	ICA: [69]
Single		NMF: [72] Neural network: [73]	

usually one dominant source in the primary components, with its direction created using only amplitude panning techniques. More specific conditions for the basic level will be detailed in subsection A. The medium complexity level requires only the condition of one dominant sources, without restricting how its direction (can be amplitude panning, delay, or HRIR, etc.) can be created. In the complex audio scene level, we consider multiple dominant sources in the primary components. The number of dominant sources in this case is also usually limited to 2-3 since it is impractical for listeners to concentrate on too many sources at one time. Instead, listeners would simply consider those sources as ambient components. Note that those PAE approaches that claimed to work in multiple sources using subband techniques, but did not present a detailed study, will not be classified in the complex level category. From these two perspectives, we shall classify the existing PAE approaches into several different categories, as summarized in Table I. With a glance of this table, it is observed that most of the PAE works are mainly focused on the stereo signals, due to the large amount of stereo content. There are some works carried out for multichannel signals, while very few works is on single channel signals. This is probably due to the limited information in single channel signals. Next, we will review the PAE work in each category.

A. Basic Stereo Signal Model

PAE aims to separate the primary component from the ambient component based on their perceptual spatial features. The perceptual spatial features can be characterized by the inter-channel relationships, including inter-channel time difference (ICTD), inter-channel level difference (ICLD), and inter-channel cross-correlation coefficient (ICC) [47]. Since the number of primary sources is usually unknown and might be varying, a common practice in spatial audio processing is to convert the signals into time-frequency domain using short-time Fourier transform (STFT) [31], [26], [48], [34], [45], [49], [25] or subband via filter banks like hybrid quadrature mirror filter banks [50]. For each frequency band or subband, it is generally assumed that the primary component of the input signal is composed of only one dominant source [31], [26], [45], [49]. Denoting the *b*th subband of input stereo signals at time index *m* as $\mathbf{x}_c[m,b]=\left[x_c(mN,b),\ldots,x_c(mN+N-1,b)\right]^T$, $c\in\{1,2\}$, where *N* is the length of one frame. PAE is carried out in each subband of each frame independently, and the extracted primary and ambient components are combined via inverse STFT or synthesis filter banks. The stereo signal model is expressed as:

$$\mathbf{x}_c[m,b]=\mathbf{p}_c[m,b]+\mathbf{a}_c[m,b], \tag{1}$$

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

The stereo signal model assumes the primary and ambient components in the two channels to be correlated and uncorrelated, respectively. Correlated primary component in the stereo signal can allow amplitude difference and time difference [51], i.e., $p_1(n)=kp_0(n+\tau_0)$. where k is referred to as the primary panning factor (PPF) and τ_0 is the ICTD. In this signal model, we only consider the primary component to be amplitude panned by k [26], [45], [49]. This amplitude panned primary component is commonly found in stereo recordings using pan pot stereo and coincident techniques as well as sound mixes using panning [51]. For an ambient component that consists of environmental sound, it is usually considered to be uncorrelated with the primary component [37], [52]. The ambient component in the two channels is also assumed to be uncorrelated and relatively balanced in terms of power, considering the diffuseness of ambient component. To quantify the power difference between the primary and ambient components, we introduce the primary power ratio (PPR) γ, which is defined as the ratio of total primary power to total signal power in two channels. Given any stereo input signal that fulfills the above conditions, we can derive the PPF and PPR of the stereo signal using the auto-correlations r_{00}, r_{11} and cross-correlation r_{01},

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

$$\gamma=\frac{2r_{01}+(r_{11}-r_{00})k}{(r_{11}+r_{00})k}. \tag{3}$$

Previous studies have shown that PPF and PPR are useful parameters for the extraction and the evaluation of the performance of the PAE approaches [29].

B. Stereo Signals

PAE for stereo signals in the basic complexity category can be classified into four types: (i) time frequency masking, (ii) principal component analysis (PCA), (iii) least-squares (LS), (iv) ambient spectrum estimation, and others.

One of the earliest works in primary or ambient extraction was from Avendano and Jot in 2002 [53]. In this work, a time-frequency masking was constructed to extract ambient components $\hat{A}_c$ from stereo signals X_c, as

$$\hat{A}_c(m,l) = X_c(m,l)\Psi_A(m,l), \quad (4)$$

where $0 \le \Psi_A(m,l) \le 1$ is the real-valued ambient mask at time-frequency bin (m,l). The time-frequency regions that present high coherence will have stronger primary components and low coherence time-frequency regions can be attributed to stronger ambient components [31]. Thus, they derived the ambient mask using a nonlinear function of the inter-channel coherence. Following works on time-frequency masking derive the ambient mask based on the characteristic that ambient components have equal level in the two channels of the stereo signal [49] or diffuseness measured from recordings [34].

Principal component analysis has been the most widely studied PAE approach [54]-[58], [49], [25], [17]-[19], [46], [29]. The key idea behind PCA for PAE is to extract the principal component with the largest variance as the primary components (as the name suggests). Variants of PCA include the modified PCA that ensures uncorrelated ambience extraction [56], enhanced post-scaling to restore the correct primary-to-ambient energy ratio [57] and correct power of primary and ambient components [58]. In our earlier work [29], we derive a simplified solution for PCA and conducted a comprehensive objective evaluation of PCA, which leads us to the applications of PCA in PAE.

Least-squares is another type of commonly used PAE approach [45], [38], [36], [41], [29], [59]. Based on the basic stereo signal model, least-squares derives the estimated primary and ambient components by minimizing the mean-square-error (MSE) [45]. Serval variants of least-squares have been proposed and studied in our earlier work [29]. Furthermore, other least-squares variants were introduced to improve the spatial quality of the extracted primary and ambient components [36], [59]. A linear estimation based framework was proposed in [29], which generalizes the PCA and LS based PAE approaches. Under the linear estimation framework, the extracted primary and ambient components are estimated as a weighted sum of the input signals, which can be expressed as:

$$\begin{bmatrix}\hat{\mathbf{p}}_0 & \hat{\mathbf{p}}_1 & \hat{\mathbf{a}}_0 & \hat{\mathbf{a}}_1\end{bmatrix} = \begin{bmatrix}\mathbf{x}_0 & \mathbf{x}_1\end{bmatrix}\begin{bmatrix} w_{\mathrm{P0,0}} & w_{\mathrm{P1,0}} & w_{\mathrm{A0,0}} & w_{\mathrm{A1,0}} \\ w_{\mathrm{P0,1}} & w_{\mathrm{P1,1}} & w_{\mathrm{A0,1}} & w_{\mathrm{A1,1}} \end{bmatrix}. \quad (5)$$

Under this framework, different PAE approaches can be derived based on different objectives related to the performance of the extraction. Details on the linear estimation based PAE can be found in [29].

One problem with the linear estimation based PAE approaches is that the extracted components suffer from error due to uncorrelated ambient components, because of the nature of summing input signals directly. To solve this problem, a new framework based on ambient spectrum estimation was introduced recently [60], [61]. The ambient spectrum estimation framework for PAE was derived based on the assumption that the diffuse ambient components have identical magnitude but varied phase. Such an additional relation is used to simplify the PAE problem into the problem of estimating only the ambient phase [60], or ambient magnitude [61]. Subsequently, the ambient phase or magnitude can be estimated using an optimization technique with a sparsity constraint. Furthermore, an approximate solution to the ambient spectrum estimation problem (known as ambient phase estimation using input signal phase APEX) greatly reduce the computational cost while yielding close performance.

Other PAE approaches that fall into this category include [22] that derives an out-of-phase signal as ambient components; [32] that considers ambient components as the sum of a common component and an independent component; and [62] that classifies various signal models for extraction.

In order to handle stereo signals that consist of primary components, whose directions are created using time/phase differences (i.e., medium complexity), several works can be found in the literature. Usher and Benesty proposed an adaptive approach using normalized least-mean-squares (NLMS) to extract reverberation from stereo microphone recordings [37]. However, this adaptive approach cannot always yield a good performance in a short time. By contrast, we proposed time-shifted PCA [63] and extended time shifting technique [64] to solve this problem, which is much simpler and yields a better performance. The time shifting technique in PAE will be discussed in Section III.

With respect to stereo signals with multiple sources, there is not much work conducted so far. One prior work by Dong *et al.* applied PCA in polar coordinates to reduce the coding noise of stereo signals for multiple source case [65]. However, the extraction performance was not studied. To fill in this gap, we conducted two works that studied PCA with different frequency partitioning methods in frequency domain [40], and PCA with multiple time shifts in time domain [66], which will be discussed in Section III.

C. Multichannel Signals

Besides the extensive study on PAE for stereo signals, PAE on multichannel signals is less well studied [48], [67], [18], [68]. PCA was originally proposed to work for multichannel signals with only one dominant amplitude-panned source in [26]. . For other multichannel signals with one dominant source, independent component analysis (ICA) can be combined with time-frequency masking to extract the dominant sources [69]. Another approach that was extended from [31], achieves primary ambient extraction using a system of pairwise correlation [70]. Recently, Stefanakis *et al.*

Fig. 4 Framework of preprocessing and postprocessing on PAE

introduced W-disjoint orthogonality (WDO) and PCA based foreground suppression techniques in multichannel microphone recordings [27]. In the case of multiple sources in multichannel signals, blind source separation techniques can be employed for the purpose of primary ambient extraction. When the number of dominant sources is no more than the number of channels (as it is the case for PAE), ICA is a common technique [71]. Compared to stereo signals, PAE with multichannel signals is in fact easier to solve since there are more information available. Moreover, PAE approaches based on stereo signals can also be extended to multichannel signals, which will be presented in Section III.

D. Single Channel Signals

Since there are no inter-channel cues in single channel signals, how primary and ambient components can be defined and characterized becomes a critical problem. Nevertheless, two works shed some light on solving such a problem. In [72], it is considered that ambient components exhibit a less repetitive and constructive spectra structure than primary components. Therefore, when applying non-negative matrix factorization (NMF) on the single channel signal, primary components are better explained and factorized, and the residue can thus be considered as ambient components. However, the NMF method suffers from high computational complexity and latency. To avoid this problem, Uhle and Paul introduced a supervised learning approach for ambient extraction from single channel signals [73], where a neural network is trained to obtain an ambient spectra mask. Subjective listening tests in [73] validated the improved perceptual quality of the up-mix systems employing these PAE approaches.

III. A UNIFIED FRAMEWORK FOR PRIMARY AMBIENT EXTRACTION IN PRACTICE

In order to deal with complex signals in PAE, two types of methods can be considered. The first type concerns the direct design of PAE approaches based on the new signal model. However, it is less flexible as different approaches needs to be developed for different complex cases and there might not be a promising solution. On this note, we can consider to extend the basic PAE approaches that are proposed under the basic stereo signal model, to handle complex input signals. By using preprocessing techniques and associated post-processing, the complex signals can be maximally transformed into signals that closely match the conditions in the basic signal model, and hence basic PAE approaches can be re-used. Compared to the direct method, this method is more flexible and effective. The unified framework of PAE with preprocessing and post-processing for complex input signals is illustrated in Fig. 4. In this section, we will describe five types of preprocessing and post-processing techniques.

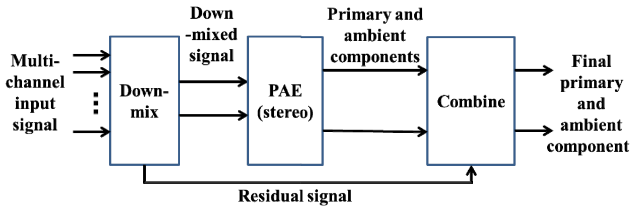

Fig. 5 Down-mix based multichannel PAE

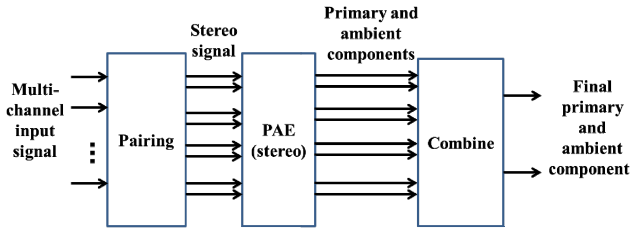

Fig. 6 Multichannel PAE based on pairwise stereo

A. Preprocessing for Multichannel Signals

Since most work in PAE are focused on stereo signals and there is still much content in multichannel format, it is also important to develop PAE approaches for multichannel signals. Direct multichannel PAE takes the multichannel signal as the direct input of PAE, e.g., PCA [26], least-squares [45], and linear system of pairwise correlations [70]. However, considering the extensive study of PAE on stereo signals, reusing the existing stereo PAE approaches is thus preferred. In this section, we shall discuss two ideas on how to extend stereo PAE to multichannel PAE using down-mix and pairing techniques.

The most straightforward way to extend stereo PAE to multichannel PAE is via down-mixing the multichannel signal into the stereo signal. The block diagram for down-mix based multichannel PAE is shown in Fig. 5. Stereo based PAE is applied to the down-mixed signal. However, due to the existence of various down-mix methods [74], we also consider a combine process where the residual signal from the down-mix process is combined with the PAE output to obtain the final primary and ambient components. Take 5.1 surround sound as an example. The down-mix methods [21] include: LoRo, LoRo excluding the center channel (because usually the center channel contains the vocal of a movie/game sound track and does not involve in the spatial effects), LoRo excluding rear channels (considered as only ambience), and PAE independently applied on front left and right signals and rear signals together with the residual center channel signal. Other down-mix methods (e.g., binaural down-mix, varying the weights on the down-mixing matrix) could also be considered. Selecting the best down-mix method and stereo PAE method for multichannel PAE is content- and application- dependent. More comprehensive study and evaluation need to be carried out.

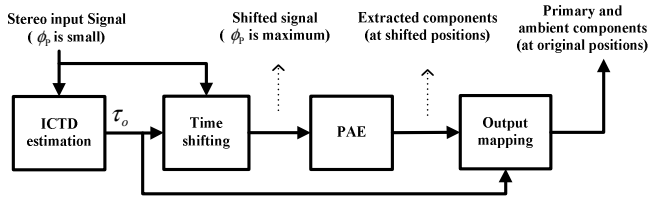

Fig. 7 Block diagram of time shifting as a preprocessing in PAE. ϕ_P is the primary correlation at zero lag.

Compared to down-mix, a more general way to solve multichannel PAE problem is to apply pairwise stereo PAE. As shown in Fig. 6, different pairs of stereo signals are selected from the multichannel input signal and PAE is applied independently to these pairs of stereo signals. Finally, the output from each PAE process is combined to form the final primary and ambient components. One critical problem in this approach is how to select the pairs of stereo signals from multichannel signals. For a total number of *M* channels, there are $M(M-1)/2$ options in total. One possible solution is to apply PAE for all possible pairs. However, in addition to the high computational cost, not every pair fits the stereo signal model. Considering the pairwise amplitude panning nature of multichannel signals, pairing every two neighboring channels (based on the actual layout) is thus more desirable. In this case, only *M* pairs of stereo signals are constructed.

B. Preprocessing for PPF and PPR

In some PAE approaches (e.g., ambient spectrum estimation, it is usually assumed that $k \geq 1$, so that the derivation can be greatly simplified. In order to deal with the case with $k < 1$, we can preprocess the input signal by simply switching the channels. After the extraction, a corresponding switch on the channels of the output is also required.

On the other hand, sound scenes are usually very complex and the two model parameters (PPF and PPR) may change drastically from one frame to another. Such variations could incur undesirable distortions, especially when the frame length is rather short. A good practice in PAE implementation involves applying a forgetting factor in the computation of auto- and cross- correlations or directly in the PPF and PPR, so that the output is smoothened [48], [49]. However, one drawback is that some transient sound events could not be well reproduced.

C. Preprocessing for time differences

In order to deal with stereo input signals with partially correlated primary components at zero lag (mainly due to the ICTD of the primary components), the time shifting technique was proposed in [64]. The block diagram of the time shifting as a preprocessing in PAE is shown in Fig. 7. First, the stereo input signal is time-shifted according to the estimated ICTD of the primary component. Subsequently, conventional PAE approaches (e.g., PCA) can be applied to the shifted signal and extract primary and ambient components at shifted positions. Finally, the time indices of extracted primary and ambient components are mapped to their original positions based on the same ICTD. According to the experimental results in [64], the time shifting technique could reduce the extraction error and extract the spatial cues more accurately.

Furthermore, only nonnegative primary correlation (i.e., $\phi_P \geq 0$) is accounted in the time shifting based PAE. In order to handle the possible negative primary correlation, another preprocessing can be employed before time shifting. That is, multiple one of the two-channel signals by -1. Correspondingly, the same channel in the output shall also be multiplied by -1.

D. Preprocessing for Multiple Sources

Though one dominant source in the primary components is found to be quite common in PAE, it is still possible to encounter the cases with multiple dominant (usually up to three) sources in some movies and games. To handle such cases, we investigated two types of preprocessing techniques to adapt the PAE approaches. The first technique considers subband decomposition of the full-band input signal [28], [29] and then performs PAE in each subband. For this approach, the partitioning of the frequency bins into subbands is found to be critical, where an adaptive top-down partitioning method introduced in [40] outperforms other methods. The other way is a multi-shift technique that involves multiple instances of time shifting, performs extraction for each shifted signals, and combines the extracted components from all shifting versions using different weights. The weighting methods based on ICC is found to yield the most robust performance [66].

E. Post-processing

As discussed in this section, any preprocessing techniques are coupled with a corresponding post-processing technique, so that the final output of the extracted primary and ambient components matches with the input signals. Besides, we can also consider other post-processing techniques to further enhance the performance of the extraction output, especially the spatial performance. For example, the diffuseness of the extracted ambient components in many approaches is not so good [29]. Therefore, decorrelation [75] and post-scaling techniques [45], [58] are applied to further enhance the ambient extraction.

IV. Evaluation, Experiments and Discussions

With the complex input signals transformed into "simpler" signals that match the basic stereo signal model, basic stereo based PAE approaches as shown in Table I can be re-used. In this section, we consider four commonly used PAE approaches as discussed in Section II.B: Masking [49], PCA [26], LS [45], and APEX [61], and evaluate their timbre and spatial quality using objective measures and subjective listening tests.

A. Performance Measures

An evaluation framework for PAE was initially proposed in [29]. In general, we are concerned with the extraction

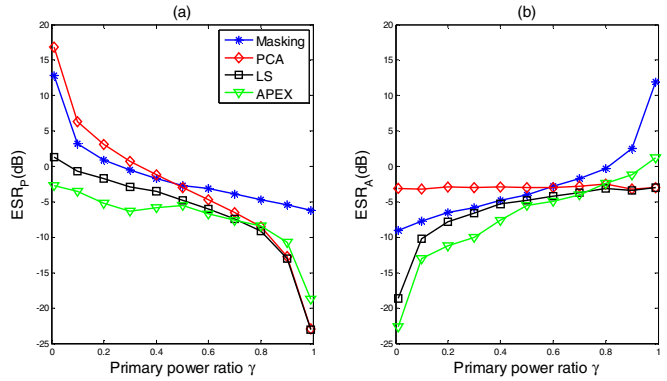

Fig. 8 ESR of (a) extracted primary components and (b) extracted ambient components.

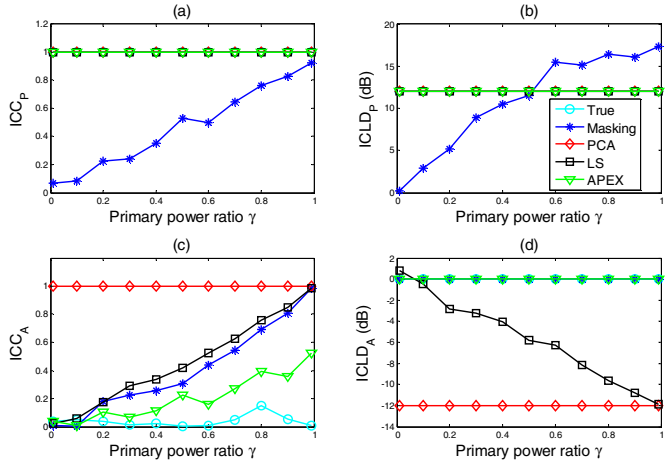

Fig. 9 Spatial performance of (a), (b) the extracted primary and (c), (d) ambient components.

accuracy (timbre quality) and spatial accuracy (spatial quality) in PAE. The overall extraction accuracy of PAE is quantified by error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components, where lower ESR indicates better extraction of these components. The ESR for the primary and ambient components are computed as

$$\mathrm{ESR_P} = 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\|\hat{\mathbf{p}}_c - \mathbf{p}_c\|_2^2}{\|\mathbf{p}_c\|_2^2}\right\},$$
$$\mathrm{ESR_A} = 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\|\hat{\mathbf{a}}_c - \mathbf{a}_c\|_2^2}{\|\mathbf{a}_c\|_2^2}\right\}. \quad (6)$$

On the other hand, spatial accuracy is measured using the inter-channel cues, i.e., ICC, ICTD and ICLD. In these approaches, there is no ICTD involved in the primary components according to the basic mixing model and hence it is not evaluated. Correct localization of the primary components requires $\mathrm{ICC_P}$ to be 1 and $\mathrm{ICLD_P}$ close to its true values. The spatial accuracy of the ambient component is evaluated in terms of its diffuseness, where a more diffuse ambient component requires both $\mathrm{ICC_A}$ and $\mathrm{ICLD_A}$ (in dB) to be closer to 0.

B. Objective Evaluation

Simulations are conducted to evaluate these PAE approaches. The stereo mixed signals employed in the experiments are synthesized in the following way. One frame (4096 samples, sampling rate: 44.1 kHz) of speech signal is selected as the primary component, which is amplitude panned to channel 1 with a panning factor k = 2. A wave lapping sound recorded at the beach is selected as the ambient component, which is decorrelated using all-pass filters with random phase [76]. The stereo signal is obtained by mixing the primary and ambient components based on different γ values ranging from 0 to 1 with an interval of 0.1.

The ESR of the four PAE approaches with respect to different values of γ is illustrated in Fig. 8. Generally, the performance of all these PAE approaches varies with γ. As γ increases, $\mathrm{ESR_P}$ decreases while $\mathrm{ESR_A}$ increases (except $\mathrm{ESR_A}$ of PCA). Comparing these four approaches, we found that the recently proposed APEX approach yields the best overall performance, especially for lower γ. LS always outperforms PCA and Masking, as also found out in [29]. The Masking approach performs better than PCA when γ is low, and but is the worst when γ becomes high.

The spatial accuracy of PAE is shown in Fig. 9. For the extracted primary components, all approaches but Masking yield very accurate $\mathrm{ICC_P}$ and $\mathrm{ICLD_P}$. This is the outcome of taking into account the inter-channel relations in these three approaches compared to Masking. The spatial performance of the extracted ambient components differs from the primary components. As shown in Fig. 9(c), the lowest and highest $\mathrm{ICC_A}$ are achieved with true ambient components and ambient components extracted by PCA, respectively. The APEX approach outperforms the other three PAE approaches. For $\mathrm{ICLD_A}$, we observed that only APEX and Masking could extract ambient components with equal level in the two channels, whereas PCA and LS perform much worse.

C. Subjective Evaluation

Lastly, subjective tests were carried out to evaluate the perceptual performance of these four PAE approaches. A total of 17 subjects, between 20-30 years old, participated in the listening tests. None of the subjects reported any hearing issues. The tests were conducted in a quiet listening room in Nanyang Technological University, Singapore. An Audio Technica MTH-A30 headphone was used. The stimuli used in this test were synthesized using amplitude panned (k = 2) primary components (speech, music, and bee sound) and decorrelated ambient components (forest, canteen, and waterfall sound) based on two values of primary power ratio ($\gamma = 0.3,\ 0.7$) for the duration of 2-4 seconds. Both the extraction accuracy and spatial accuracy were examined. The testing procedure was based on MUSHRA [77], [78], where a more specific anchor (i.e., the mixture) is used instead of the low-passed anchor, according to recent revision of MUSHRA as discussed in [78]. The MATLAB GUI was modified based on the one used in [79]. Subjects were asked to listen to the clean reference sound and tested sounds obtained from

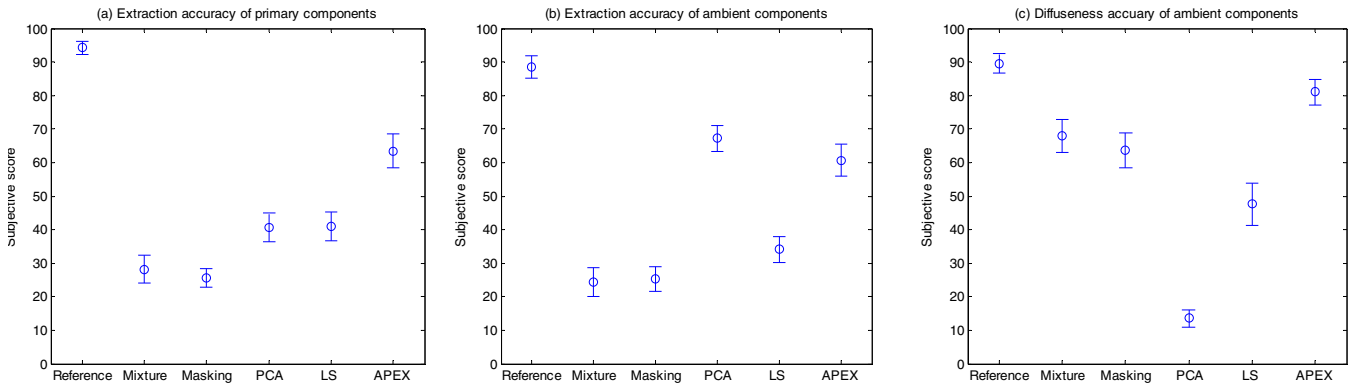

Fig. 10 Subjective score of the PAE approaches.

different PAE approaches, and give a score of 0-100 as the response, where 0-20, 21-40, 41-60, 61-80, and 81-100 represents a bad, poor, fair, good, and excellent quality, respectively. Finally, we analyzed the subjects' responses for the hidden reference (i.e., clean primary or ambient components), mixture, and the four PAE approaches. The mean values with 95% confidence intervals of the subjective scores of the extraction and spatial accuracy for the tested PAE approaches are illustrated in Fig. 10. Note that for each PAE approach, we combine the subjective scores of different test stimuli and different values of primary power ratio, so as to represent the overall performance of these PAE approaches.

Despite the relatively large variations among the subjective scores that are probably due to the different scales employed by the subjects and the differences among the stimuli, we observe the following trends. On one hand, APES outperforms other PAE approaches in extracting more accurate primary components, as shown in Fig. 10(a). In Fig. 10(b), APEX, though slightly worse off than PCA, still produces considerable accuracy in ambient extraction. The good perceptual performance of ambient components extracted from PCA lies in the very low amount of primary leakage [29]. On the other hand, we found that the spatial performance were also affected by the undesired leakage signals as compared to the clean reference, as found in the mixtures that preserve all spatial cues, but were rated lower than the reference. With respect to the diffuseness of the ambient components, APEX performs the best while PCA performs quite poorly. On this note, we find PCA sacrifices on the diffuseness of the extracted ambient components for the sake of a better perceptual extraction performance. Furthermore, ANOVA results indicate the significant differences among these PAE approaches.

Comparing the subjective evaluation results with the objective evaluation results obtained with ESR, we found that they are consistent in general. For the perceptual performance that cannot be explained by ESR alone, we shall include more specific objective performance measures, such as those introduced in [29]. Meanwhile, rather than only focusing on the overall perceptual performance, subjective listening tests shall also be conducted to evaluate the more specific perceptual performance. These results would help us gain more insights on the performance of PAE approaches.

V. CONCLUSIONS

In this paper, we investigated primary ambient extraction approaches for immersive spatial audio reproduction. An in-depth review of the prior work in PAE was conducted, where we realized that most PAE approaches were studied for signals under the basic stereo signal model. In order to improve the practical application of PAE, we proposed a unified framework that extends the conventional PAE approaches to deal with complex signals. The proposed framework employs preprocessing and post-processing techniques, including down-mixing and pairing for multichannel input signals, smoothening the model parameters, time shifting for partially correlated primary components, subband decomposition and multi-shift for multiple dominant sources, and etc. On the other hand, we evaluate four typical PAE approaches using objective measures and subjective listening tests. These evaluation results provide us more insights on applying PAE in practical spatial audio reproduction applications.

REFERENCES

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

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

[3] J. M. Loomis, J. R. Marston, R. G. Golledge, and R. L. Klatzky, "Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance," *J. Vis. Impair blind*, vol. 99, no. 4, pp. 219-232, Jan. 2005.

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

[5] T. Särkämö, E. Pihko, et al., "Music and speech listening enhance the recovery of early sensory processing after stroke," *J. Cognitive Neuroscience*, vol. 22, no. 12, pp. 2716-2727, 2010.

[6] ITU, "Report ITU-R BS.2159-4: Multichannel sound technology in home and broadcasting applications," 2012.
[7] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones: integration of signal processing techniques," *IEEE Signal Processing Magazine*, vol. 32. no. 2, pp. 100-113, Mar. 2015.
[8] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio – the new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, Dec. 2014.
[9] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank,and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proc. of the IEEE* , vol.101, no.9, pp.1920-938, Sept. 2013.
[10] Dolby Atmos-Next Generation Audio for Cinema (White Paper). 2013.http://www.dolby.com/uploadedFiles/Assets/US/Doc/Professional/Dolby-Atmos-Next-Generation-Audio-for-Cinema.pdf
[11] J. M. Jot, and Z. Fejzo, "Beyond surround sound - creation, coding and reproduction of 3-D audio soundtracks," *in 131st Audio Eng. Soc. Conv.*, New York, NY, Oct. 2011.
[12] F. Rumsey, "Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm," J. Audio Eng. Soc., vol. 50, no. 9, pp. 651–666, Sep. 2002.
[13] F. Rumsey, "Spatial audio: eighty years after Blumlein," J. Audio Eng. Soc., vol. 59, no. 1/2, pp. 57–62, Jan./Feb. 2011.
[14] F. Rumsey, "Spatial audio processing: upmix, downmix, shake it all about," *J. Audio Eng. Soc.,* vol. 61, no. 6, pp. 474–478, Jun. 2013.
[15] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001.
[16] W. S. Gan, E. L. Tan, and S. M. Kuo, "Audio projection: directional sound and its application in immersive communication," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 43-57, Jan. 2011.
[17] E. L. Tan, and W. S. Gan, "Reproduction of immersive sound using directional and conventional loudspeakers," *Proc. Acoust. 2012,* Hong Kong, Apr. 2012.
[18] E. L. Tan, W. S. Gan, and C. H. Chen, "Spatial sound reproduction using conventional and parametric loudspeakers," in *Proc. APSIPA ASC*, Hollywood, CA, 2012.
[19] M. R. Bai and G. Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *IEEE Trans. Consumer Electron.*, vol. 53, no. 3, pp. 1011-1019, Aug. 2007.
[20] M. A. Gerzon, "Optimal reproduction matricies for multispeaker stereo," *J. Audio Eng. Soc.,* vol. 40, no. 7/8, pp. 571–589, Jul./Aug. 1992.
[21] Rec. ITU-R BS.775, Multi-Channel Stereophonic Sound System with or without Accompanying Picture, ITU, 1993.
[22] J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, Lang. Process.,* vol.16, no. 8, pp. 1503-1511, Nov. 2008.
[23] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and other Applications.* Hoboken, NJ: Wiley, 2007.
[24] S. K. Zielinski, and F. Rumsey, "Effects of down-mix algorithms on quality of surround sound," *J. Audio Eng. Soc.,* vol. 51, no. 9, pp. 780–798, Sep. 2003.
[25] C. Faller, and F. Baumgarte, "Binaural cue coding-Part II: schemes and applications," *IEEE Trans. Speech Audio Process.*, vol.11, no.6, pp.520,531, Nov. 2003
[26] M. M. Goodwin and J. M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. ICASSP,* Hawaii, 2007, pp. 9-12.
[27] N. Stefanakis, and A. Mouchtaris, "Foreground suppression for capturing and reproduction of crowded acoustic environments," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 51-55.
[28] M. M. Goodwin and J. M. Jot, "Spatial audio scene coding," in *Proc. 125th Audio Eng. Soc. Conv.,* San Francisco, 2008.
[29] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process*., vol. 22, no. 2, pp. 505-517, Feb. 2014.
[30] K. Kowalczyk, O. Thiergart, M. Taseska, G. D. Galdo, V. Pulkki, and E. Habets, "Parametric spatial sound processing," *IEEE Signal Process. Magazine*, vol. 32, no. 2, Mar 2015, pp. 31- 42.
[31] 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.
[32] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th Audio Eng. Soc. Conv.,* London, UK, 2010.
[33] J. M. Jot, J. Merimaa, M. M. Goodwin, A. Krishnaswamy, and J. Laroche, "Spatial audio scene coding in a universal two-channel 3-D stereo format," in *123rd Audio Eng. Soc. Conv.,* New York, NY, Oct. 2007.
[34] V. Pulkki, "Spatial Sound Reproduction with directional audio coding," *J. Audio Eng. Soc.,,* vol. 55, no. 6, pp. 503-516, Jun. 2007.
[35] F. Rumsey, "Time-frequency processing for spatial audio," *J. Audio Eng. Soc.,* vol. 58, no. 7/8, pp. 655–659, Jul./Aug. 2010.
[36] S. W. Jeon, Y. C. Park, S. Lee, and D. Youn, "Robust representation of spatial sound in stereo-to-multichannel upmix," in *Proc. 128th Audio Eng. Soc. Conv.,* London, UK, 2010.
[37] J. Usher and J. Benesty, "Enhancement of spatial sound quality: a new reverberation-extraction audio upmixer," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 15, no. 7, pp. 2141-2150, Sep. 2007.
[38] C. Faller, "Matrix surround revisited," *in Audio Eng. Soc. 30th Int. Conf.,* Saariselka, Finland, Mar. 2007.
[39] P. Kleczkowski, A. Krol, and P. Malecki, "Multichannel sound reproduction quality improves with angular separation of direct and reflected sounds," *J. Audio Eng. Soc.*, Vol. 63, No. 6, pp. 427-442, Jun. 2015.
[40] J. He, W. S. Gan, and E. L. Tan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2892-2896.
[41] C. Faller and J. Breebaart, "Binaural reproduction of stereo signals using upmixing and diffuse rendering," in *Proc. 131th Audio Eng. Soc. Conv.,* New York, 2011.
[42] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.
[43] M. Goodwin and J.-M. Jot. A frequency-domain framework for spatial audio coding based on universal spatial cues. *120th Conv. of the Audio Eng. Soc.*, May 2006.
[44] J. M. Jot, V. Larcher, and J. M. Pernaux, "A Comparative Study of 3-D Audio Encoding and Rendering Techniques," in *Proc. 16th Audio Eng. Soc. Int. Conf.*, Rovaniemi, Finland, 1999.
[45] C. Faller, "Multiple-loudspeaker playback of stereo signals", *J. Audio Eng. Soc.,* vol. 54, no. 11, pp. 1051-1064, Nov. 2006.
[46] T. Lee, Y. Baek, Y. C. Park, and D. H. Youn, "Stereo upmix-based binaural auralization for mobile devices," *IEEE Trans. Consum. Electron*, vol. 60, no. 3, pp.411-419, Aug. 2014.
[47] F. Baumgarte, and C. Faller, "Binaural cue coding-Part I: psychoacoustic fundamentals and design principles," *IEEE*

Trans. Speech Audio Process., vol.11, no.6, pp.509-519, Nov. 2003

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

[49] 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.

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

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

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

[53] C. Avendano, and J. M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," *Proc. ICASSP,* vol.2, no., pp. 13-17 May 2002.

[54] 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.

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

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

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

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

[59] C. Uhle, and E. A. P. Habets, "Direct-ambient decomposition using parametric wiener filtering with spatial cue control," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 36-40.

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

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

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

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

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

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

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

[67] A. Walther, and C. Faller, "Direct-ambient decomposition and upmix of surround signals," in *Proc. IWASPAA*, New Paltz, NY, Oct. 2011, pp. 277-280.

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

[69] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time–frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 14, no. 6, pp. 2165–2173, Nov. 2006.

[70] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," in *Proc. 133rd Audio Eng. Soc. Conv.,* San Francisco, 2012.

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

[72] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using non-negative matrix factorization", in *Proc. 30th Audio Eng. Soc. Int. Conf.*, Saariselka, Finland, 2007.

[73] C. Uhle, and C. Paul, "A supervised learning approach to ambience extraction," *Proc. DAFx* Espoo, Finland, 2008.

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

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

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

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

[78] J. Liebetrau, F. Nagel, N. Zacharov, et al., "Revision of Rec. ITU-R BS. 1534," *in Proc. 137th AES conv.*, LA, Oct, 2014.

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

[80] C. Shi and Y. Kajikawa, "Ultrasound-to-ultrasound Volterra filter identification of the parametric array loudspeaker," *Proc. 20th Int. Conf. Digital Sig. Process.*, Singapore, Jul. 2015.

[81] C. Shi and Y. Kajikawa, "Identification of the parametric array loudspeaker with a Volterra filter using the sparse NLMS algorithm," *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[82] C. Shi and Y. Kajikawa, "A convolution model for computing the far-field directivity of a parametric loudspeaker array," *J. Acoust. Soc. Amer.*, vol. 137, no. 2, pp. 777-784, Feb. 2015.

[83] C. Shi, E. L. Tan, and W. S. Gan, "Hybrid immersive three-dimensional sound reproduction system with steerable parametric loudspeakers," *Proc. 21st Int. Congr. Acoust., Montreal*, Canada, Jun. 2013.