Manuscript accepted in IEEE/ACM Trans. Audio, Speech, Lang. Process.

Primary-Ambient Extraction Using Ambient Spectrum Estimation for Immersive Spatial Audio Reproduction

Jianjun He, *Student Member, IEEE*, Woon-Seng Gan, *Senior Member, IEEE*, and Ee-Leng Tan

***Abstract*— The diversity of today's playback systems requires a flexible, efficient, and immersive reproduction of sound scenes in digital media. Spatial audio reproduction based on primary-ambient extraction (PAE) fulfills this objective, where accurate extraction of primary and ambient components from sound mixtures in channel-based audio is crucial. Severe extraction error was found in existing PAE approaches when dealing with sound mixtures that contain a relatively strong ambient component, a commonly encountered case in the sound scenes of digital media. In this paper, we propose a novel ambient spectrum estimation (ASE) framework to improve the performance of PAE. The ASE framework exploits the equal magnitude of the uncorrelated ambient components in two channels of a stereo signal, and reformulates the PAE problem into the problem of estimating either ambient phase or magnitude. In particular, we take advantage of the sparse characteristic of the primary components to derive sparse solutions for ASE based PAE, together with an approximate solution that can significantly reduce the computational cost. Our objective and subjective experimental results demonstrate that the proposed ASE approaches significantly outperform existing approaches, especially when the ambient component is relatively strong.**

***Index Terms*—Primary-ambient extraction (PAE), spatial audio, ambient spectrum estimation (ASE), sparsity, computational efficiency.**

I. Introduction

SPATIAL audio reproduction of digital media (such as movies and video games) has gained significant popularity over the recent years [1]. The reproduction methods generally differ in the formats of audio content. Despite the growing interest in object-based audio formats [1], such as Dolby Atmos [2], DTS multi-dimensional audio (DTS: X) [3], most existing digital media content is still in channel-based formats (such as stereo and multichannel signals). The channel-based audio is usually specific in its playback configuration, and it does not support flexible playback configurations in domestic or personal listening circumstances [1]. Considering the wide diversity of today's playback systems [4], it becomes necessary to process audio signals such that the reproduction of the audio content is not only compatible with various playback systems 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 [7]. The challenge in loudspeaker playback deals with the mismatch of loudspeaker playback systems in home theater applications, where the number of loudspeakers [8] or even the type of loudspeakers [9]-[11] between the intended loudspeaker system (based on the audio content) and the actual loudspeaker system is different. Conventional techniques to solve this challenge are often referred to as audio remixing (i.e., downmix and upmix), for example, "Left only, Right only (LoRo)", "Left total, Right total (LtRt)", matrix-based mixing surround sound systems, etc [8], [12]-[14]. 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 [15], where virtualization of loudspeakers is achieved by binaural rendering, i.e., convolving the channel-based signals with head-related impulse responses (HRIRs) of the corresponding loudspeaker positions. These conventional techniques in spatial audio reproduction are capable of solving the compatibility issue, but the spatial quality of the reproduced sound scene is usually limited [12], [16]-[18]. To improve the spatial quality of the sound reproduction, the MPEG audio standardization group proposed MPEG Surround and related techniques, which typically address the multichannel and binaural audio reproduction problem based on human perception [19]-[21]. In the synthesis, these techniques usually employ the one-channel downmixed signal and the subband spatial cues, which better suit the reproduction of the distinct directional source signals as compared to the diffuse signals [19], [22].

To further improve the quality of the reproduced sound scene, the perception of the sound scenes is considered as a combination of the foreground sound and background sound [23], which are often referred to as primary (or direct) and ambient (or diffuse) components, respectively [24]-[27]. 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,

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2010-T2-2-040.

Jianjun He, and Woon-Seng Gan are with Digital Signal Processing Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798. Ee-Leng Tan is currently with Beijing Sesame World Co. Ltd, China. (Email: jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg, joseph@sesame-world.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2434272, IEEE/ACM Transactions on Audio, Speech, and Language Processing

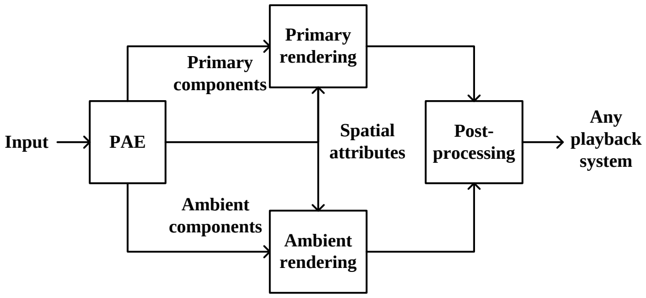


Fig. 1. Block diagram of PAE based spatial audio reproduction [25].

applause, or nature sound like waterfall [22], [28]. 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 [22], [29]. However, the existing channel-based audio formats provide only the mixed signals [30], which necessitate the process of extracting primary and ambient components from the mixed signals. This extraction process is usually known as the primary-ambient extraction (PAE).

As a spatial audio processing tool [8], [16], [17], [19], [26], [29], PAE can also be incorporated into spatial audio coding systems, such as spatial audio scene coding [24], [31], and directional audio coding [32]. Essentially, PAE serves as a front-end to facilitate flexible, efficient, and immersive spatial audio reproduction. First, by decomposing the primary and ambient components of the sound scene, PAE enables the sound reproduction format to be independent of the input format, hence increasing the flexibility of spatial audio reproduction [31], [33]. 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 [25]. 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 [24], [34]-[36].

Figure 1 illustrates the PAE based spatial audio reproduction system, where the primary and ambient components undergo different rendering schemes [25]. The rendering schemes differ for loudspeaker or headphone playback [28], [34], [37]. For loudspeaker playback, the primary components are reproduced using vector base amplitude panning (VBAP) [32] or vector base intensity panning [38], [39] to reproduce the accurate direction of the sound sources. The ambient components, on the other hand, are further decorrelated and distributed to all the loudspeaker channels to create an envelopment effect of the sound environment [24], [40]. For headphone playback, the conventional virtualization that simply applies binaural rendering to the mixed channel-based signals is problematic [16], [17]. PAE based virtualization resolves this problem by applying binaural rendering to the extracted primary components, creating accurate virtual sound sources in the desired directions [17] for headphone playback [26], [41]. Similar to the loudspeaker playback case, the ambient components are decorrelated using artificial reverberation [19], [24], [28], [29] to create a more natural sound environment.

Numerous approaches are applied to solve PAE with stereo and multichannel input signals [22], [42]. In this paper, we focus on stereo input signals since they are still one of the most widely used formats and the PAE approaches for stereo signals can be extended to deal with multichannel signals [22], [42]-[44]. For the basic signal model, the (stereo) mixed signal is considered as the sum of the primary and ambient components. The primary and ambient components are mainly discriminated by their inter-channel cross-correlations, i.e., the primary and ambient components are considered to be correlated and uncorrelated, respectively [22]. Based on this model, time-frequency masking approaches were introduced, where the mask is obtained as a nonlinear function of the inter-channel coherence of the input signal [28] or derived based on the characteristic that ambient components have equal level in the two channels of the stereo signal [42], [45]. Further investigation of the differences between two channels of the stereo signals has led to several types of linear estimation based approaches [46], including principal component analysis (PCA) based approaches [11], [22], [45]-[53] and least-squares based approaches [40], [46], [54]. These linear estimation based approaches extract the primary and ambient components using different performance-related criteria [46]. To deal with complex input signals that do not satisfy the basic stereo signal model, other PAE approaches consider signal model classification [55], time/phase differences in primary components [25], [35], [53], [56] non-negative matrix factorization [57], independent component analysis [58], etc.

Due to the nature of summing input signals directly [46], the aforementioned PAE approaches often have difficulty in removing uncorrelated ambient component in the extracted primary and ambient components. The extraction error in these PAE approaches is more severe when the ambient component is relatively strong compared to the primary component [46], as often encountered in digital media content, including busy sound scenes with many discrete sound sources that contribute to the environment as well as strong reverberation indoor environment. In [59], we proposed an ambient phase estimation (APE) framework to improve the performance of PAE. The APE framework exploits the equal-magnitude characteristic of uncorrelated ambient components in the mixed signals of digital media content and was solved by pursuing the sparsity of the primary components [60] (this approach is known as APES [59]). However, due to the trigonometric operations required in the estimation of the ambient phase, the computational cost of APES is still too high. In this paper, we re-consider the PAE problem from a higher level, i.e., from the perspective of ambient spectrum estimation (ASE). Besides APE, a new formulation referred to as ambient magnitude estimation (AME) is derived and solved using the same criterion as in APES. Furthermore, an approximate solution to the ASE problem shall be discussed, so as to further reduce the computational cost.

A comparative analysis on the objective performance of the proposed ASE approaches and existing PAE approaches in terms of exaction error and computational efficiency is conducted with our simulations. To perform an in-depth evaluation of these PAE approaches, the performance measures proposed in [46] are adopted. However, the calculation of these performance measures in [46] is only applicable for PAE approaches with analytic solutions. Therefore, we propose a novel technique to compute these measures for PAE approaches without analytic solutions, as is the case with the proposed ASE approaches. Moreover, statistical variations are introduced to the ambient magnitudes to examine the robustness of the proposed ASE approaches. Furthermore, subjective listening tests are conducted to complement the objective evaluation.

The remainder of this paper is structured as follows. In Section II, we review the basic stereo signal model. The ambient spectrum estimation framework for PAE, including APE and AME, is formulated in Section III. This is followed by the proposed solutions for ASE in Section IV. Section V explains the calculation of the performance measures, which are used to evaluate the PAE approaches in the experiments in Section VI. Finally, Section VII concludes this paper.

II. Stereo signal model

In spatial audio, PAE is often considered in time-frequency domain [17], [20], [22], [28], [32], [40], [45]. It is generally assumed that there is only one dominant directional source (a.k.a., primary component) in each frequency band of the input signal. PAE is independently carried out on each frequency band of each frame (consisting of *N* short frames) of the input signal [22], [28], [33], [40], [45]. We denote the stereo signal in time-frequency domain at time index *n* and frequency bin index *l* as $X_c(n,l)$, where the channel index $c \in \{0,1\}$. Hence, the stereo signal at subband *b* that consists of bins from $l_{b-1}+1$ to l_b is expressed as $\mathbf{X}_c[n,b] = \left[X_c(n,l_{b-1}+1), X_c(n,l_{b-1}+2), \ldots, X_c(n,l_b)\right]^T$ [38]. The stereo signal model is expressed as:

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

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

The stereo signal model assumes that the primary and ambient components in the two channels of the stereo signals are correlated and uncorrelated, respectively. Correlated primary component can be characterized by time and amplitude differences between the two channels [61]. For this paper, we shall only consider the primary component to be amplitude panned, that is, $\mathbf{P}_1 = k\mathbf{P}_0$, where *k* is referred to as the primary panning factor [22], [40], [45]. For this paper, we assume $k \geq 1$ such that the channel containing the stronger directional primary component is channel 1. This amplitude panned primary component is commonly found in stereo recordings using coincident techniques and sound mixes using conventional amplitude panning techniques [30]. Considering that only the mixed signal is given as input and no prior information is available, it is necessary to estimate *k* using correlations [46] or histograms of amplitude differences [62]. For example, based on the autocorrelations of the two channels r_{00}, r_{11}, and cross-correlations between the two channels r_{01}, we can estimate *k* as $k = \frac{r_{11}-r_{00}}{2r_{01}} + \sqrt{\left(\frac{r_{11}-r_{00}}{2r_{01}}\right)^2 + 1}$.

For an ambient component that is made up of environmental sound, it is usually considered to be uncorrelated with the primary component [35], [63], [64], as well as being balanced in two channels in terms of signal power. To quantify the power difference between the primary and ambient components, the primary power ratio γ is defined as the ratio of total primary power to total signal power in two channels: $\gamma = \sum_{c=0}^{1}\left\|\mathbf{P}_c\right\|_2^2 \Big/ \sum_{c=0}^{1}\left\|\mathbf{X}_c\right\|_2^2$, $\gamma \in [0,1]$. Previous study revealed that the performance of PAE is highly dependent on γ, where lower γ generally indicates inferior overall extraction performance [46]. Using the method described in [46], we computed γ for many movie and gaming tracks (e.g., Avatar, Brave, Battlefield 3, BioShock Infinite), and found that the percentage for the cases with over half of the time frames having relative strong ambient power (i.e., $\gamma \leq 0.75$) is around 70% in these digital media content examples. Since high occurrence of strong ambient power case degrades the overall performance of PAE, a PAE approach that also performs well in the presence of strong ambient power is desired.

III. Ambient Spectrum estimation

The diffuseness of ambient components usually leads to low cross-correlation between the two channels of the ambient components in the stereo signal. During the mixing process, the sound engineers synthesize the ambient component using various decorrelation techniques, such as introducing delay [65], all-pass filtering [66]-[68], artificial reverberation [15], and binaural artificial reverberation [69]. These decorrelation techniques often maintain the magnitude of ambient components in the two channels of the stereo signal. As such, we can express the spectrum of ambient components as

$$\mathbf{A}_c = \left|\mathbf{A}_c\right| \odot \mathbf{W}_c, \ \forall c \in \{0,1\}, \tag{2}$$

where $\odot$ denotes element-wise Hadamard product, $\left|\mathbf{A}_0\right| = \left|\mathbf{A}_1\right| = \left|\mathbf{A}\right|$ is the equal magnitude of the ambient components, and the element in the bin (*n*, *l*) of $\mathbf{W}_c$ is $W_c(n,l) = e^{j\theta_c(n,l)}$, where $\theta_c(n,l)$ is the bin (*n*, *l*) of $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_c = \angle\mathbf{A}_c$ is the vector of phase samples (in radians) of the ambient components. Following these discussions, we shall derive the ASE framework for PAE in two ways: ambient phase estimation [55] and ambient magnitude estimation.

A. Ambient Phase Estimation

Considering the panning of the primary component $\mathbf{P}_1 = k\mathbf{P}_0$, the primary component in (1) can be cancelled out and we arrive at

$$\mathbf{X}_1 - k\mathbf{X}_0 = \mathbf{A}_1 - k\mathbf{A}_0. \tag{3}$$

By substituting (2) into (3), we have

$$|\mathbf{A}| = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0), \qquad (4)$$

where $./$ represents the element-wise division. Because ambient magnitude $|\mathbf{A}|$ is real and non-negative, we derive the relation between the phases of the two ambient components as (refer to Appendix A for detailed derivation)

$$\boldsymbol{\theta}_0 = \boldsymbol{\theta} + \arcsin\left[k^{-1}\sin(\boldsymbol{\theta} - \boldsymbol{\theta}_1)\right] + \boldsymbol{\pi}, \qquad (5)$$

where $\boldsymbol{\theta} = \angle(\mathbf{X}_1 - k\mathbf{X}_0)$. Furthermore, by substituting (4) and (2) into (1), we have

$$\begin{aligned} \mathbf{A}_c &= (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c, \\ \mathbf{P}_c &= \mathbf{X}_{\mathrm{c}} - (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c,\ c \in \{0,1\}. \end{aligned} \qquad (6)$$

Since $\mathbf{X}_c$ and k can be directly computed from the input [46], $\mathbf{W}_0, \mathbf{W}_1$ are the only unknown variables on the right hand side of the expressions in (6). In other words, the primary and ambient components are determined by $\mathbf{W}_0, \mathbf{W}_1$, which are solely related to the phases of the ambient components. Therefore, we reformulate the PAE problem into an ambient phase estimation (APE) problem. Based on the relation between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ stated in (5), only one ambient phase $\boldsymbol{\theta}_1$ needs to be estimated.

B. *Ambient Magnitude Estimation*

To reformulate the PAE problem as an ambient magnitude estimation problem, we rewrite (1) for every time-frequency bin as:

$$X_0^{'} = kX_0 = P_1 + kA_0,\ X_1 = P_1 + A_1. \qquad (7)$$

Consider these bin-wise spectra stated in (7) as vectors in complex plane (represented by an arrow on top), we can express their geometric relations in Fig. 2 as

$$\begin{aligned} &\overrightarrow{X_0^{'}} = \overrightarrow{\mathrm{OB}} = (\mathrm{B}_{\mathrm{Re}}, \mathrm{B}_{\mathrm{Im}}),\ \overrightarrow{X_1} = \overrightarrow{\mathrm{OC}} = (\mathrm{C}_{\mathrm{Re}}, \mathrm{C}_{\mathrm{Im}}), \\ &\overrightarrow{P_1} = \overrightarrow{\mathrm{OP}} = (\mathrm{P}_{\mathrm{Re}}, \mathrm{P}_{\mathrm{Im}}), \\ &k\overrightarrow{A_0} = \overrightarrow{\mathrm{PB}},\ \overrightarrow{A_1} = \overrightarrow{\mathrm{PC}}. \end{aligned} \qquad (8)$$

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

$$\begin{aligned} (\mathrm{P}_{\mathrm{Re}} - \mathrm{B}_{\mathrm{Re}})^2 + (\mathrm{P}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}})^2 &= k^2\hat{r}^2, \\ (\mathrm{P}_{\mathrm{Re}} - \mathrm{C}_{\mathrm{Re}})^2 + (\mathrm{P}_{\mathrm{Im}} - \mathrm{C}_{\mathrm{Im}})^2 &= \hat{r}^2. \end{aligned} \qquad (9)$$

The solution of $(\mathrm{P}_{\mathrm{Re}}, \mathrm{P}_{\mathrm{Im}})$ for (9) is given by:

$$\begin{aligned} \hat{\mathrm{P}}_{\mathrm{Re}} &= \frac{\mathrm{B}_{\mathrm{Re}} + \mathrm{C}_{\mathrm{Re}}}{2} + \frac{(\mathrm{C}_{\mathrm{Re}} - \mathrm{B}_{\mathrm{Re}})(k^2 - 1)\hat{r}^2 \pm (\mathrm{B}_{\mathrm{Im}} - \mathrm{C}_{\mathrm{Im}})\beta}{2|\overrightarrow{\mathrm{BC}}|^2}, \\ \hat{\mathrm{P}}_{\mathrm{Im}} &= \frac{\mathrm{B}_{\mathrm{Im}} + \mathrm{C}_{\mathrm{Im}}}{2} + \frac{(\mathrm{C}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}})(k^2 - 1)\hat{r}^2 \mp (\mathrm{B}_{\mathrm{Re}} - \mathrm{C}_{\mathrm{Re}})\beta}{2|\overrightarrow{\mathrm{BC}}|^2}, \end{aligned} \qquad (10)$$

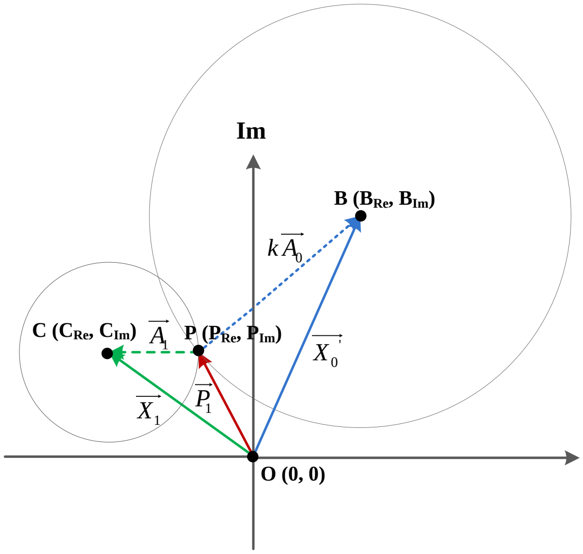

Im
B (B$_{\mathrm{Re}}$, B$_{\mathrm{Im}}$)
$k\overrightarrow{A_0}$
C (C$_{\mathrm{Re}}$, C$_{\mathrm{Im}}$)
$\overrightarrow{A_1}$
P (P$_{\mathrm{Re}}$, P$_{\mathrm{Im}}$)
$\overrightarrow{X_0^{'}}$
$\overrightarrow{X_1}$
$\overrightarrow{P_1}$
Re
O (0, 0)

Fig. 2. Geometric representation of (7) in complex plane in AME.

where the Euclidean distance between the points B and C, $|\overrightarrow{\mathrm{BC}}| = \sqrt{(\mathrm{C}_{\mathrm{Re}} - \mathrm{B}_{\mathrm{Re}})^2 + (\mathrm{C}_{\mathrm{Im}} - \mathrm{B}_{\mathrm{Im}})^2}$ and $\beta = \sqrt{\left[(k+1)^2\hat{r}^2 - |\overrightarrow{\mathrm{BC}}|^2\right]\left[(k-1)^2\hat{r}^2 - |\overrightarrow{\mathrm{BC}}|^2\right]}$. Based on (8), the spectra of the primary and ambient components can then be derived as:

$$\begin{aligned} &\hat{P}_1 = \hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}},\ \hat{P}_0 = k^{-1}(\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}), \\ &\hat{A}_1 = X_1 - (\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}),\ \hat{A}_0 = X_0 - k^{-1}(\hat{\mathrm{P}}_{\mathrm{Re}} + j\hat{\mathrm{P}}_{\mathrm{Im}}). \end{aligned} \qquad (11)$$

Therefore, the PAE problem becomes the problem of determining r, i.e., ambient magnitude estimation. The approach to determine r and select one of the two solutions in (10) will be discussed in Section IV. It can be inferred from Fig. 2 that determining the ambient magnitude is equivalent to determine the ambient phase as either of them will lead to the other. Therefore, we conclude that APE and AME are equivalent and they are collectively termed as ambient spectrum estimation. The block diagram of the ASE based PAE is illustrated in Fig. 3. We argue that in theory, by accurately obtaining the spectra of ambient components, it is possible to achieve perfect extraction (i.e., error-free) of the primary and ambient components using the formulation of ASE, which is not possible with existing PAE approaches as a consequence of residue error from the uncorrelated ambient component [46].

IV. AMBIENT SPECTRUM ESTIMATION WITH A SPARSITY CONSTRAINT

The proposed ambient spectrum estimation framework can greatly simplify the PAE problem into an estimation problem with only one unknown parameter per time-frequency bin. To estimate these parameters, we shall exploit other characteristics of the primary and ambient components that have not been used in previous derivation. One of the most important characteristics of sound source signals is sparsity, which has

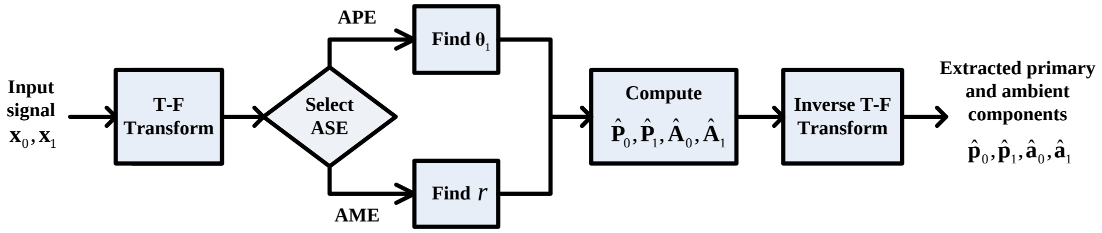

APE
Find θ_1
Input signal $\mathbf{x}_0, \mathbf{x}_1$
T-F Transform
Select ASE
Compute $\hat{\mathbf{P}}_0, \hat{\mathbf{P}}_1, \hat{\mathbf{A}}_0, \hat{\mathbf{A}}_1$
Inverse T-F Transform
Extracted primary and ambient components $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \hat{\mathbf{a}}_0, \hat{\mathbf{a}}_1$
Find r
AME

Fig. 3. Block diagram of ASE based PAE.

been widely used as a critical criterion in finding optimal solutions in many audio and music signal processing applications [60]. In PAE, since the primary components are essentially directional sound sources, they can be considered to be sparse in the time-frequency domain [60]. Therefore, we estimate the ambient phase or magnitude spectrum by restricting that the extracted primary component is sparse. We refer to these approaches as ambient spectrum estimation with a sparsity constraint (ASES). By applying the sparsity constraint in APE and AME, ASES can be divided into two approaches, namely, APES and AMES.

A. Ambient Phase Estimation with a Sparsity Constraint

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

$$\hat{\boldsymbol{\theta}}_1^* = \arg\min_{\hat{\boldsymbol{\theta}}_1} \left\| \hat{\mathbf{P}}_1 \right\|_1, \tag{12}$$

where $\left\| \hat{\mathbf{P}}_1 \right\|_1$ is the 1-norm of the primary component, which is equal to the sum of the magnitudes of the primary component over the time-frequency bins. Since the objective function in (12) is not convex, convex optimization techniques are inapplicable. Heuristic methods, like simulated annealing [70], require optimization to be performed for all the phase variables, and hence are inefficient in solving APES [59]. On this note, a more efficient method referred to as discrete searching (DS) to estimate ambient phase was proposed in [59]. DS is proposed based on the following two observations. First, the magnitude of the primary component at one time-frequency bin is solely determined by the phase of the ambient component at the same time-frequency bin and hence, the estimation in (12) can be independently performed for each time-frequency bin. Second, the phase variable is bounded to $(-\pi, \pi]$ and high precision of the estimated phase may not be necessary. Thus, the optimal phase estimates can be selected from an array of discrete phase values $\hat{\theta}_1(d) = (2\pi d/D - \pi)$, where $d \in \{1, 2, \ldots, D\}$ with D being the total number of phase values to be considered. In general, the value of D affects the extraction and the computational performance of APES using DS [59]. Following (5) and (6), a total number of D estimates of the primary components can be computed. The estimated phase then corresponds to the minimum of magnitudes of the primary component, i.e., $\hat{\theta}_1^* = \hat{\theta}_1(d^*)$, where $d^* = \arg \min_{d \in \{1,2,\ldots,D\}} \left| \hat{P}_1(d) \right|$. Finally, the extracted primary and ambient components are computed using (6). It shall be noted that in DS, a sufficient condition of the sparsity constraint was employed in solving the APES problem in (12).

B. Ambient Magnitude Estimation with a Sparsity Constraint

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

$$\hat{\mathbf{r}}^* = \arg\min_{\hat{\mathbf{r}}} \left\| \hat{\mathbf{P}}_1 \right\|_1, \tag{13}$$

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

To derive the solution for AMES, we follow the geometric relation illustrated in Fig. 2. To ensure the existence of intersection point P, the following constraint

$$\left|\overrightarrow{\mathrm{PC}}\right| - \left|\overrightarrow{\mathrm{PB}}\right| \le \left|\overrightarrow{\mathrm{BC}}\right| \le \left|\overrightarrow{\mathrm{PB}}\right| + \left|\overrightarrow{\mathrm{PC}}\right|, \tag{14}$$

has to be satisfied, which leads to:

$$r \in \left[r_{lb}, r_{ub} \right], \tag{15}$$

where $r_{lb} = \dfrac{\left|\overrightarrow{\mathrm{BC}}\right|}{k+1}, r_{ub} = \dfrac{\left|\overrightarrow{\mathrm{BC}}\right|}{k-1}, \forall k \ne 1.$ When k = 1, there is no physical upper bound from (14). Based on the objective of minimizing the magnitude of primary component, we can actually enforce an approximate upper bound for k = 1, for example, let $r_{ub} = \left|\overrightarrow{\mathrm{OB}}\right| + \left|\overrightarrow{\mathrm{OC}}\right|$, $\forall k = 1$. Thus, the ambient magnitude is bounded, and the same numerical method DS (as used in APES) is employed to estimate r in AMES. Consider an array of discrete ambient magnitude values $\hat{r}(d) = \left(1 - \dfrac{d-1}{D-1}\right) r_{lb} + \dfrac{d-1}{D-1} r_{ub}$, where $d \in \{1, 2, \ldots, D\}$ with D being the total number of ambient magnitude estimates considered. For each magnitude estimate $\hat{r}(d)$, we select the one $\left(\hat{\mathrm{P}}_{\mathrm{Re}}, \hat{\mathrm{P}}_{\mathrm{Im}}\right)$ of two solutions from (10) which gives the smaller primary magnitude. After derivation, we can unify the solution for the selected $\left(\hat{\mathrm{P}}_{\mathrm{Re}}, \hat{\mathrm{P}}_{\mathrm{Im}}\right)$ as

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

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

D: number of phase or magnitude estimates in discrete searching

$$
\begin{aligned}
\hat{\mathrm{P}}_{\mathrm{Re}}\left(d\right) &= \frac{\mathrm{B}_{\mathrm{Re}}+\mathrm{C}_{\mathrm{Re}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Re}}-\mathrm{B}_{\mathrm{Re}}\right)\left(k^2-1\right)\hat{r}^2\left(d\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2} \\
&\quad +\frac{\left(\mathrm{B}_{\mathrm{Im}}-\mathrm{C}_{\mathrm{Im}}\right)\beta\left(d\right)\operatorname{sgn}\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2}, \\
\hat{\mathrm{P}}_{\mathrm{Im}}\left(d\right) &= \frac{\mathrm{B}_{\mathrm{Im}}+\mathrm{C}_{\mathrm{Im}}}{2}+\frac{\left(\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\right)\left(k^2-1\right)\hat{r}^2\left(d\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2} \\
&\quad +\frac{-\left(\mathrm{B}_{\mathrm{Re}}-\mathrm{C}_{\mathrm{Re}}\right)\beta\left(d\right)\operatorname{sgn}\left(\mathrm{B}_{\mathrm{Re}}\mathrm{C}_{\mathrm{Im}}-\mathrm{B}_{\mathrm{Im}}\mathrm{C}_{\mathrm{Re}}\right)}{2\left|\overrightarrow{\mathrm{BC}}\right|^2},
\end{aligned}
\tag{16}
$$

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

$$\left|\hat{P}_1\left(d\right)\right|=\sqrt{\hat{\mathrm{P}}_{\mathrm{Re}}^2\left(d\right)+\hat{\mathrm{P}}_{\mathrm{Im}}^2\left(d\right)}, \tag{17}$$

The estimated ambient magnitude then corresponds to the minimum of the primary component magnitude, i.e., $\hat{r}^*=\hat{r}\left(d^*\right)$, where $d^*=\arg\min_{d\in\{1,2,\ldots,D\}}\left|\hat{P}_1\left(d\right)\right|$. Finally, the extracted primary and ambient components are computed using (11).

C. Computational Cost of APES and AMES

In this subsection, we compare the computational cost of APES and AMES, as shown in Table I. In general, both AMES and APES are quite computational extensive. AMES requires more operations which include square root, addition, multiplication, and division, but requires no trigonometric operations. By contrast, APES requires 7*D*+6 times of trigonometric operations for every time-frequency bin. The computational efficiency of these two approaches is affected by the implementation of these operations.

D. An Approximate solution: APEX

To obtain a more efficient approach for ambient spectrum estimation, we derive an approximate solution in this subsection. For every time-frequency bin, we can rewrite (1) for the two channels as:

$$
\begin{aligned}
\left|X_0\right|^2 &= \left|P_0\right|^2+\left|A_0\right|^2+2\left|P_0\right|\left|A_0\right|\cos\theta_{PA0} \\
&= k^{-2}\left|P_1\right|^2+\left|A\right|^2+2k^{-1}\left|P_1\right|\left|A\right|\cos\theta_{PA0}, \\
\left|X_1\right|^2 &= \left|P_1\right|^2+\left|A_1\right|^2+2\left|P_1\right|\left|A_1\right|\cos\theta_{PA1} \\
&= \left|P_1\right|^2+\left|A\right|^2+2\left|P_1\right|\left|A\right|\cos\theta_{PA1},
\end{aligned}
\tag{18}
$$

where θ_{PA0}, θ_{PA1} are the phase differences between the spectra of the primary and ambient components in channel 0 and 1, respectively. From (18), we can obtain that

$$
\begin{aligned}
&\left(1-k^{-2}\right)\left|P_1\right|^2+2\left|A\right|\left(\cos\theta_{PA1}-k^{-1}\cos\theta_{PA0}\right)\left|P_1\right| \\
&-\left(\left|X_1\right|^2-\left|X_0\right|^2\right)=0.
\end{aligned}
\tag{19}
$$

Solving (19) for $\left|P_1\right|$, we arrive at (20). From (20), when *k* > 1, the minimization of $\left|P_1\right|$ can be approximately achieved by minimizing $k^{-1}\cos\theta_{PA0}-\cos\theta_{PA1}$ (considering that $\left|X_1\right|^2\geq\left|X_0\right|^2$ in most cases since $k\geq1$), which leads to $\theta_{PA0}=\pi,\ \theta_{PA1}=0.$ According to the relation between the two ambient phases in (5), we can infer that it is impossible to always achieve both $\theta_{PA0}=\pi$ and $\theta_{PA1}=0$ at the same time. Clearly, since $k>1$, a better approximate solution would be taking $\theta_{PA1}=0.$ On the other hand, when *k* = 1, one approximate solution to minimize $\left|P_1\right|$ would be letting $\theta_0-\theta_1=\pi.$ These constraints can be applied in either APE or AME framework. Here, applying the constraints in APE is more straightforward and we shall obtain the approximate phase estimation as:

$$\hat{\boldsymbol{\theta}}_1^*=\begin{cases}\angle\mathbf{X}_1, & \forall k>1\\ \angle\left(\mathbf{X}_1-\mathbf{X}_0\right), & \forall k=1\end{cases}. \tag{21}$$

As the phase (or the phase difference) of the input signals is employed in (21), we refer to this approximate solution as APEX. As shown in Table I, APEX requires the lowest computational cost and is significantly more efficient than either APES or AMES. The performance of these approaches will be evaluated in the following sections.

V. Evaluation of PAE

An evaluation framework for PAE was initially proposed in [46]. In general, we are concerned with the extraction accuracy and spatial accuracy in PAE. The overall extraction accuracy of PAE is quantified by error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components, where lower ESR indicates better extraction of these components. The ESR for the primary and ambient components are computed as

$$
\begin{aligned}
\mathrm{ESR}_{\mathrm{P}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{p}}_c-\mathbf{p}_c\right\|_2^2}{\left\|\mathbf{p}_c\right\|_2^2}\right\}, \\
\mathrm{ESR}_{\mathrm{A}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{a}}_c-\mathbf{a}_c\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}.
\end{aligned}
\tag{22}
$$

The extraction error can be further decomposed into three components, namely, the distortion, interference, and leakage (refer to [46] for the explanation of these three error components). Corresponding performance measures of these error components can be computed directly for PAE approaches with analytic solutions. As there is no analytic solution for these ASE approaches, we need to find alternative ways to compute these measures. In this section, we propose a novel optimization technique to estimate these performance measures.

First, we consider the extracted primary component in time domain $\hat{\mathbf{p}}_c$. Since the true primary components in two channels are completely correlated, no interference is incurred [46]. Thus we can express $\hat{\mathbf{p}}_c$ as

$$\hat{\mathbf{p}}_c = \mathbf{p}_c + Leak_{\mathbf{p}_c} + Dist_{\mathbf{p}_c}, \tag{23}$$

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

$$\left(w^*_{Pc,0}, w^*_{Pc,1}\right) = \arg \min_{\left(w_{Pc,0}, w_{Pc,1}\right)} \left\| \hat{\mathbf{p}}_c - \mathbf{p}_c - \left(w_{Pc,0}\mathbf{a}_0 + w_{Pc,1}\mathbf{a}_1\right)\right\|_2^2, \tag{24}$$

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

$$\begin{aligned} \mathrm{LSR}_{\mathrm{P}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|w^*_{Pc,0}\mathbf{a}_0 + w^*_{Pc,1}\mathbf{a}_1\right\|_2^2}{\left\|\mathbf{p}_c\right\|_2^2}\right\}, \\ \mathrm{DSR}_{\mathrm{P}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{p}}_c - \mathbf{p}_c - \left(w^*_{Pc,0}\mathbf{a}_0 + w^*_{Pc,1}\mathbf{a}_1\right)\right\|_2^2}{\left\|\mathbf{p}_c\right\|_2^2}\right\}. \end{aligned} \tag{25}$$

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

$$\hat{\mathbf{a}}_c = \mathbf{a}_c + Leak_{\mathbf{a}_c} + Intf_{\mathbf{a}_c} + Dist_{\mathbf{a}_c}, \tag{26}$$

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

$$\begin{aligned} &\left(w^*_{Ac,c}, w^*_{Ac,1-c}\right) \\ &= \arg \min_{\left(w_{Ac,c}, w_{Ac,1-c}\right)} \left\|\hat{\mathbf{a}}_c - \mathbf{a}_c - \left(w_{Ac,c}\mathbf{p}_c + w_{Ac,1-c}\mathbf{a}_{1-c}\right)\right\|_2^2, \end{aligned} \tag{27}$$

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

$$\begin{aligned} \mathrm{LSR}_{\mathrm{A}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|w^*_{Ac,c}\mathbf{p}_c\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}, \\ \mathrm{ISR}_{\mathrm{A}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|w^*_{Ac,1-c}\mathbf{a}_{1-c}\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}, \\ \mathrm{DSR}_{\mathrm{A}} &= 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{a}}_c - \mathbf{a}_c - \left(w_{Ac,c}\mathbf{p}_c + w_{Ac,1-c}\mathbf{a}_{1-c}\right)\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}. \end{aligned} \tag{28}$$

Previous experience on evaluating linear estimation based PAE approaches such as PCA and least-squares suggests that these parameters $w_{Pc,0}, w_{Pc,1}, w_{Ac,c}, w_{Ac,1-c}$ are bounded to [-1, 1], hence we can employ a simple numerical searching method similar to DS to determine the optimal estimates of these parameters using a certain precision [46]. As audio signals from digital media are quite non-stationary, these measures shall be computed for every frame and can be averaged to obtain the overall performance for the whole track.

On the other hand, spatial accuracy is measured using the inter-channel cues. For primary components, the accuracy of the sound localization is mainly evaluated using inter-channel time and level differences (i.e., ICTD and ICLD). In this paper, there is no ICTD involved in the basic mixing model for stereo input signals, and the ICLD is essentially determined by the estimation of *k*, which is common between the proposed approaches and the existing linear estimation based approaches such as PCA [46]. For these two reasons, spatial accuracy is not evaluated for primary component extraction, but is focused on the extraction of ambient components. The spatial accuracy of the ambient component is evaluated in terms of its diffuseness, as quantified by inter-channel cross-correlation coefficient (ICC, from 0 to 1) and the ICLD (in dB). It is clear that a more diffuse ambient component requires both ICC and ICLD to be closer to 0.

VI. Experimental results and discussions

In this section, we shall present a comprehensive objective and subjective evaluation of the proposed ASE approaches and two existing PAE approaches, namely, PCA [22], and time-frequency masking [45]. In these experiments, the searching method of APES or AMES is DS with *D* = 100. Based on the performance measures introduced in Section V, we shall compare the overall extraction error performance, the specific error performance including leakage, distortion, and interference, as well as the spatial accuracy of the ambient components. Additionally, we will also compare the efficiency of these PAE approaches in terms of the computation time based on our simulation. To examine the robustness of these

$$\left|P_1\right| = \begin{cases} \dfrac{\left|A\right|\left(k^{-1}\cos\theta_{PA0} - \cos\theta_{PA1}\right) + \sqrt{\left|A\right|^2\left(k^{-1}\cos\theta_{PA0} - \cos\theta_{PA1}\right)^2 + \left(1-k^{-2}\right)\left(\left|X_1\right|^2 - \left|X_0\right|^2\right)}}{1-k^{-2}}, & \forall k > 1 \\ \dfrac{\left|X_1\right|^2 - \left|X_0\right|^2}{2\left|A\right|\left(\cos\theta_{PA1} - \cos\theta_{PA0}\right)} = \dfrac{\left|X_1\right|^2 - \left|X_0\right|^2}{-4\left|A\right|\left(\sin\dfrac{\theta_{PA0}+\theta_{PA1}}{2}\sin\dfrac{\theta_0 - \theta_1}{2}\right)} & , \forall k = 1 \end{cases}. \tag{20}$$

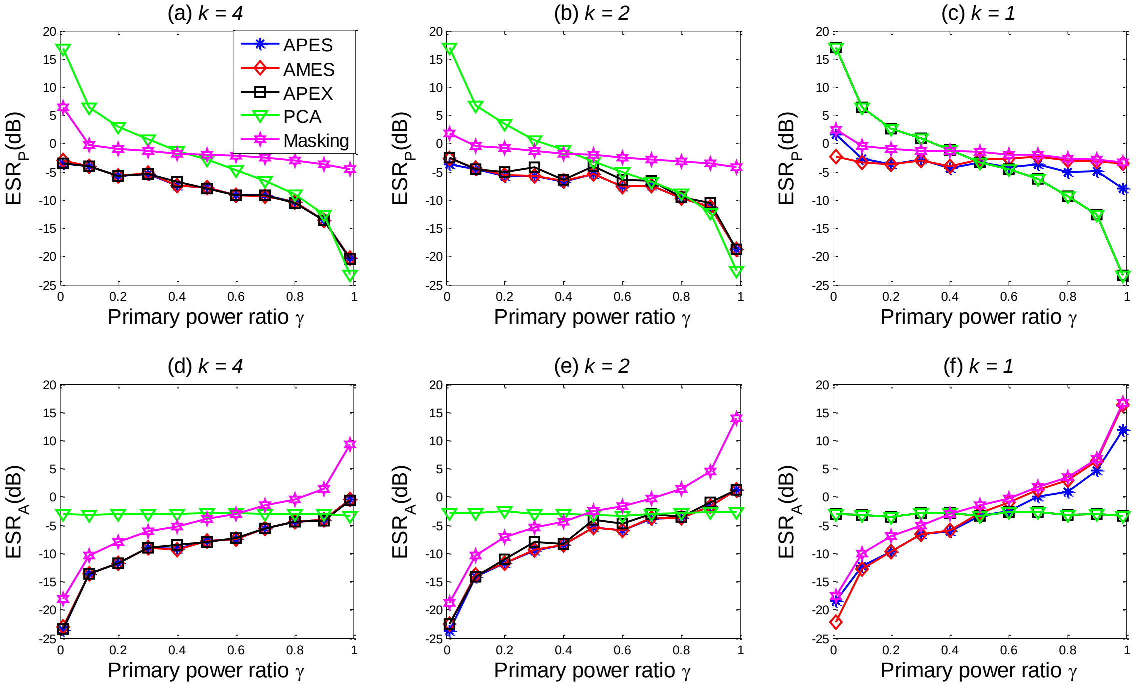


Fig. 4. Comparison of the ESR of (a-c) extracted primary components and (d-f) extracted ambient components, with respect to different *k* values, using APES, AMES, APEX, PCA [22], and Masking [45].

PAE approaches, we evaluate the proposed approaches using synthesized mixed signal with unequal ambient magnitude in two channels. Lastly, subjective listening tests were conducted to examine the perceptual timbre and spatial quality of different PAE approaches. 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 $k \in \{1,\ 2,\ 4\}$. A wave lapping sound recorded at the beach is selected as the ambient component, which is decorrelated using all-pass filters with random phase [68]. 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.

First, we compare the overall performance of the three ASE approaches with two other PAE approaches in the literature, namely, PCA [22] and Masking [45]. For the proposed ASE approaches, FFT size is set as 4096 while for Masking, the best setting for FFT size is found as 64. The ESR of these approaches with respect to different values of γ and *k* is illustrated in Fig. 4. Our observations of the ESR performance are as follows:

1) Generally, the performance of all these PAE approaches varies with γ. As γ increases, ESR$_\text{P}$ decreases while ESR$_\text{A}$ increases (except ESR$_\text{A}$ of PCA). Considering primary components to be more important in most applications, it becomes apparent that the two representative existing approaches cannot perform well when γ is low.
2) Primary panning factor *k* is the other factor that affects the ESR performance of these PAE approaches except PCA. For the Masking approach, the influence of *k* is insignificant for most cases except ESR$_\text{P}$ at very low γ and ESR$_\text{A}$ at very high γ. By contrast, the ASE approaches are more sensitive to *k*. The ESR of APES and AMES are lower at higher *k*, especially when γ is high. For APEX, the performance varies between $k > 1$, and $k = 1$, which was implied in (21).
3) Irrespective of γ and *k*, APES and AMES perform quite similar. Both APES and AMES outperform existing approaches at lower γ, i.e., from $\gamma < 0.8$ when $k = \{2, 4\}$ to $\gamma < 0.5$ when $k = 1$. APEX can be considered as an approximate solution to APES or AMES for $k > 1$, and when $k = 1$, it becomes identical to PCA (this can also be verified theoretically).

Second, we look into the specific error performance of ASE approaches at $k = 2$. Note that there are some slight variations in these error measures for close γ values, which is due to the inaccuracy in the estimation of specific error components. Nevertheless, we can observe the following trends. As shown in Fig. 5(a) and 5(b), we found that the performance improvement of ASE approaches in extracting primary components lies in the reduction of the ambient leakage, though at the cost of introducing more distortion. For ambient component extraction, PCA and Masking yield the least amount of leakage and interference, respectively. Note that the little amount of leakage in PCA and interference in Masking are actually due to the estimation error, since none of them theoretically exist in the extracted ambient components. Nevertheless, the ASE approaches yields moderate amount of these errors, which results in a better overall performance.

Third, we examine the spatial accuracy of PAE in terms of the diffuseness of the extracted ambient components. As shown

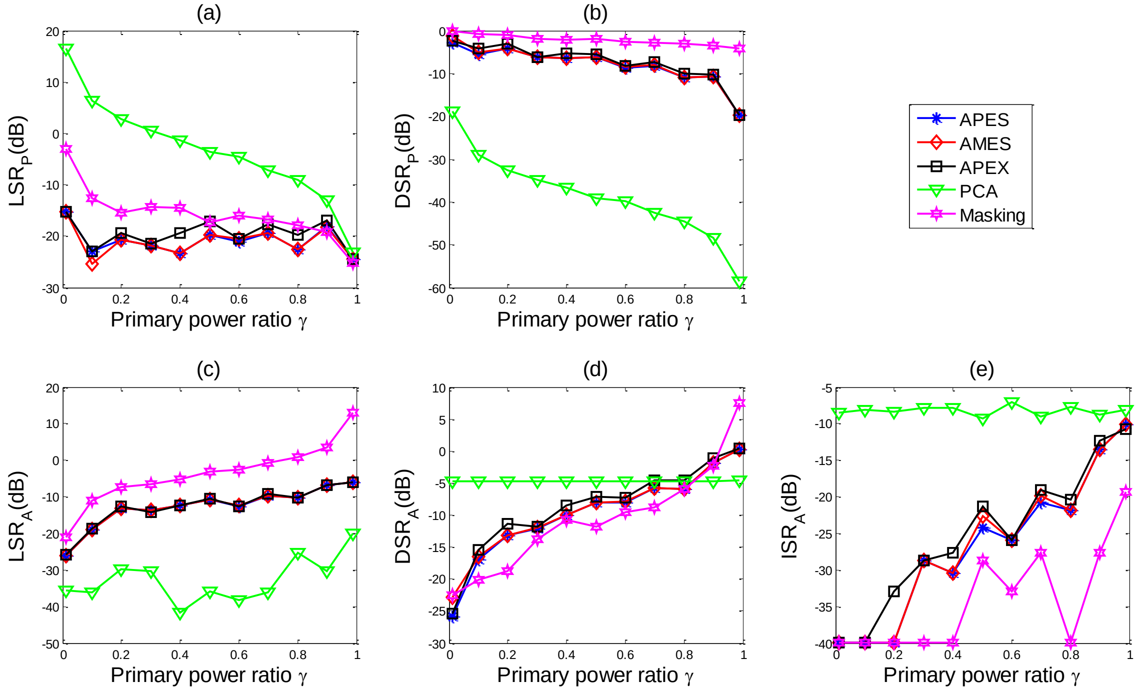


Fig. 5. Comparison of the specific error performance of (a-b) LSR and DSR in the extracted primary components and (c-e) LSR, DSR, and ISR in the extracted ambient components using APES, AMES, APEX, PCA, and Masking.

in Fig. 6(a)-(c), the lowest and highest ICC are achieved with true ambient components and ambient components extracted by PCA, respectively. The ASE approaches outperform the existing approaches, and are more effective in extracting diffuse ambient components at higher k and lower γ. For ICLD of the extracted ambient components as shown in Fig. 6(d)-(f), we observed that all approaches extract ambient components with equal level between the two channels, whereas PCA works only for $k = 1$.

Fourth, we compare the extraction performance as well as the computation time among these PAE approaches. The simulation was carried out on a PC with i5-2400 CPU, 8 GB RAM, 64-bit windows 7 operating system and 64-bit MATLAB 7.11.0. Though MATLAB simulations do not provide precise computation time measurement compared to the actual implementation, we could still obtain the relative computation performance among the PAE approaches. The results of computation time averaged across all the γ and k values are summarized in Table II. It is obvious that the three ASE approaches perform better than PCA and Masking on the average. But when we compare the computation time among APES, AMES, and APEX, we found that AMES is around 20x faster than APES, but is still far away from the computation time of the existing approaches. The APEX, which estimates the ambient phase directly using the phase of the input signals, is over 40x faster as compared to AMES and becomes quite close to the Masking approach, and hence can be considered as a good alternative ASE approach for PAE. Furthermore, in order to achieve real-time performance (in frame-based processing), the processing time must be less than 4096/44.1 = 92.88 (ms). It is clear that APEX, together with PCA and Masking satisfies this real-time constraint.

Fifth, we study the robustness of the proposed ASE approaches using experiments with the input signals containing unequal ambient magnitudes in the two channels. To quantify the violation of the assumption of equal ambient magnitude, we introduce an inter-channel variation factor v that denotes the range of variation of the ambient magnitude in one channel as compared to the other channel. Let us denote the ambient magnitude in the two channels as r_0, r_1. The variation of ambient magnitude is expressed as $v = 10\log_{10}\left(r_1/r_0\right) \quad (\text{dB})$. In the ideal case, we always have $v = 0$. To allow variation, we consider v as a random variable with mean equal to 0, and variance as σ^2. In this experiment, we consider two types of distributions for the variation, namely, normal distribution and uniform distribution, and examine the performance of these PAE approaches with respect to different variance of variations, i.e., $\sigma^2 \in [0,\ 10]$, at $\gamma = 0.5$, and $k = 2$. We run the experiment 10 times and illustrate the averaged performance in terms of ESR and ICC in Figs. 7 and 8. We observed that as the variance of the variation increases, the ESR performance of proposed ASE approaches becomes worse, though ICC was not affected much. The ASE approaches are more robust to ambient magnitude variations under normal distribution compared to uniform distribution. Compared to PCA and Masking, the proposed approaches are still better with the variance of variation up to 10 dB. Therefore, we conclude that the three ASE approaches are in general robust to ambient magnitude variations.

Lastly, subjective tests were carried out to evaluate the perceptual performance of these PAE approaches. A total of 17

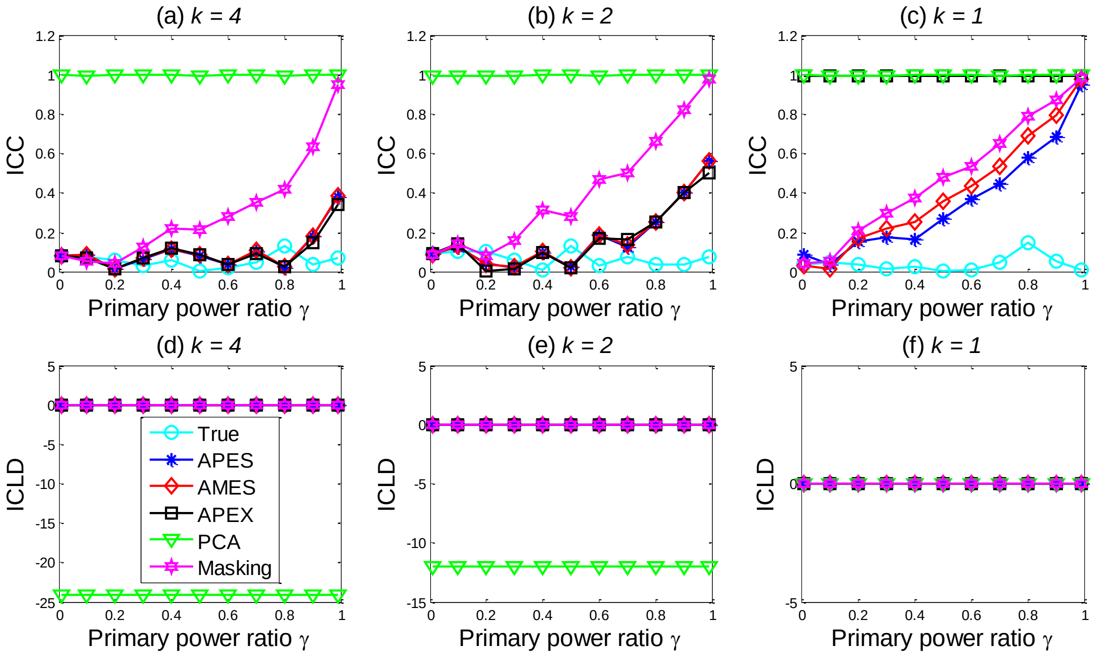


Fig. 6. Comparison of the diffuseness of the extracted ambient components in terms of (a)-(c) ICC and (d)-(f) ICLD using APES, AMES, APEX, PCA, and Masking.

Table II: Average ESR, ICC, and computation time of PAE approaches

Method	APES	AMES	APEX	PCA [22]	Masking [45]
ESR_P (dB)	-6.73	-6.31	-6.25	-3.02	-1.57
ESR_A (dB)	-6.73	-6.31	-6.25	-3.02	-2.77
ICC of ambient components	0.19	0.22	0.42	1	0.40
Computation time (ms)	3921.8	217.1	4.8	0.06	5.0

subjects (15 males and two females), who were all between 20-30 years old, participated in the listening tests. None of the subjects reported any hearing issues. The tests were conducted in a quiet listening room at Nanyang Technological University, Singapore. An Audio Technica MTH-A30 headphone was used. The stimuli used in this test were synthesized using amplitude panned (k = 2) primary components (speech, music, and bee sound) and decorrelated ambient components (forest, canteen, and waterfall sound) based on two values of primary power ratio ($\gamma = 0.3,\ 0.7$) for the duration of 2-4 seconds. Both the extraction accuracy and spatial accuracy were examined. The testing procedure was based on MUSHRA [71], [72], 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 [72]. The MATLAB GUI was modified based on the one used in [73]. Subjects were asked to listen to the clean reference sound and tested sounds obtained from different PAE approaches, and give a score of 0-100 as the response, where 0-20, 21-40, 41-60, 61-80, and 81-100 represent a bad, poor, fair, good, and excellent quality, respectively. Finally, we analyzed the subjects' responses for the hidden reference (clean primary or ambient components), mixture, and three PAE approaches, namely, Masking [45], PCA [22], and APEX. Note that APEX is selected as the representative of ASE approaches because APES and AMES exhibit very similar extraction results. The box plots of the subjective scores of the extraction and spatial accuracy for the tested PAE approaches are illustrated in Figs. 9. 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, we observed that APES outperforms the other PAE approaches in extracting accurate primary components, as shown in Fig. 9(a). In Fig. 9(b), APEX, though slightly worse off than PCA, still produces considerable accuracy in ambient extraction. The good perceptual performance of ambient components extracted from PCA lies in the very low amount of primary leakage, as shown in Fig. 5(c). On the other hand, we found that the spatial performance were also affected by the undesired leakage signals as compared to the clean reference, as found in the mixtures, which preserve the same spatial quality as the reference, but were rated lower than the reference. With respect to the diffuseness of the ambient components, APEX performs the best 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. A further analysis of the ANOVA results shows that the p-values are extremely small, which reveals that the differences among the performance of these PAE approaches are significant. To sum up the subjective evaluation results, the proposed ASE approaches yield the best performance in terms

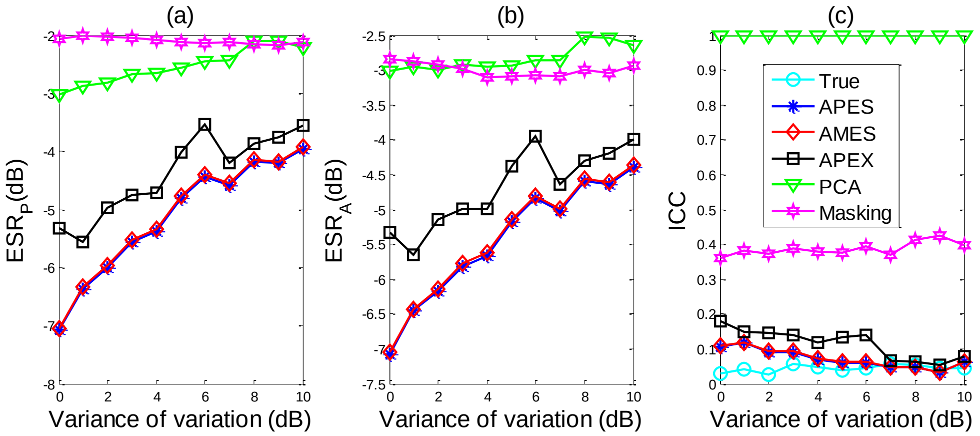


Fig. 7. Comparison of the performance of PAE approaches in the presence of normally distributed variations in the ambient magnitudes in two channels (with $\gamma = 0.5$, $k = 2$): (a) ESR$_P$, (b) ESR$_A$, (c) ICC of ambient components.

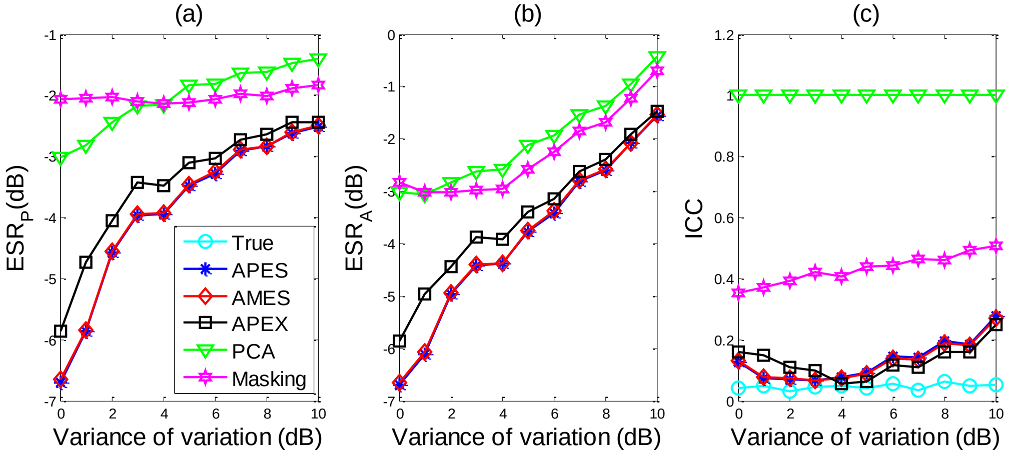


Fig. 8. Comparison of the performance of PAE approaches in the presence of uniformly distributed variations in the ambient magnitudes in two channels (with $\gamma = 0.5$, $k = 2$): (a) ESR$_P$, (b) ESR$_A$, (c) ICC of ambient components.

of extraction and spatial accuracy, which is consistent with our objective evaluation results.

For the purpose of reproducible research, the source code and some of the processed tracks used in our experiments can be found in [74]. Despite the improved performance of the proposed ASE approach as shown in these simulations and experiments, there are a few issues need to be carefully considered to generalize the results to more complex audio signals in digital media. One of them is the time-frequency transform. The proposed PAE approaches as well as the existing PAE approaches were proposed based on a basic stereo signal model. How to obtain the most appropriate time-frequency representation so that all the assumptions in this signal model are satisfied are extremely important to ensure a good PAE performance. Secondly, though only the sparsity constraint is used in this paper, other constraints can also be employed to improve the performance of ambient spectrum estimation, especially for the case with k close to 1. Some of these constraints include the correlation of the ambient components, the independence between primary and ambient components, etc. Thirdly, probabilistic approaches shall be developed to model the ambient magnitude variations better. Last but not least, extending the PAE approaches from stereo to multichannel signals (e.g., 5.1) is also of great practical value. One idea is to apply PAE to the downmixed signals [43], selected pairs [44] or even every pair of the multichannel signals [42]. However, a more comprehensive study on these extensions of PAE approaches needs to be carried out.

VII. Conclusions

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

Based on our experiments, we observed significant performance improvement of the proposed approaches over existing approaches. The improvement on error reduction is

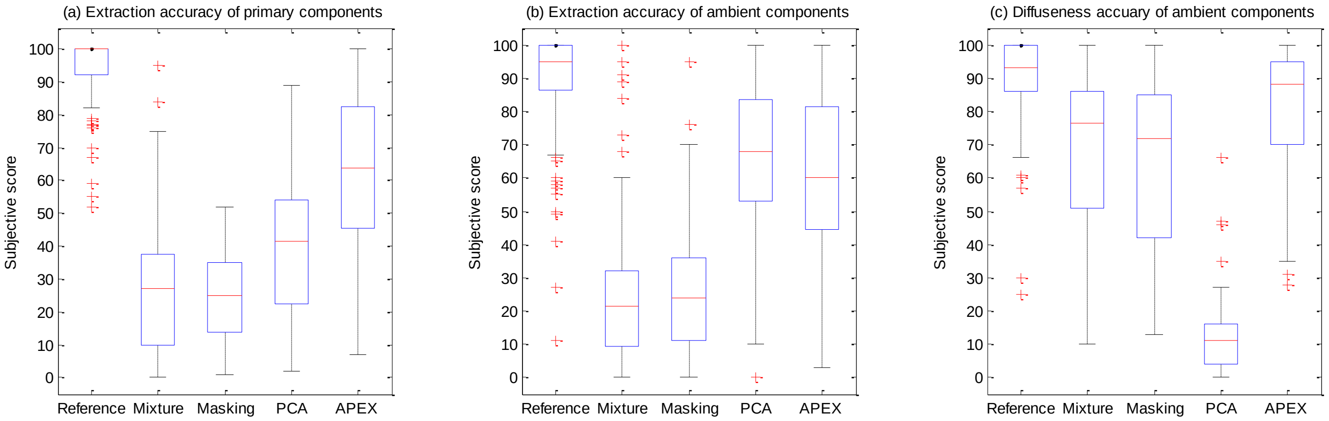


Fig. 9. Subjective performance for (a) the extraction accuracy of primary components, (b) the extraction accuracy of ambient components, and (c) diffuseness accuracy of ambient components.

around 3-6 dB on average and up to 10-20 dB for lower γ, which is mainly due to the lower residual error from the uncorrelated ambient components. Moreover, the ASE approaches perform better for mixed signals having heavily panned primary components than those having slightly panned primary components. In terms of the spatial accuracy, the ASE approaches extract more diffuse ambient components. When it comes to the computational efficiency of these PAE approaches, we found that AMES is an order of magnitude faster than APES under the same setting in MATLAB simulation, but is still not as efficient as existing approaches. For this purpose, we have also derived an approximate solution APEX and verified its effectiveness, as well as the efficiency in our simulation. Besides the ideal situation where the ambient magnitudes are equal in two channels, the robustness of these ASE approaches was also examined by introducing statistical variations to the ambient magnitudes in the two channels of the stereo signal. It was found that the proposed approaches can still yield better results with the variance of variations up to 10 dB. The objective performance of the proposed ASE approaches was also validated in our subjective tests. Future work includes the study on the use of estimation criteria other than the sparsity of the primary component [75], time-frequency transform in PAE, and handling more complex stereo and multichannel signals using ASE.

Appendix A

Derivation of the relation between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ in (5)

We show the derivation for the relation between the ambient phases in two channels. First, we rewrite $\mathbf{W}_1-k\mathbf{W}_0=(\cos\boldsymbol{\theta}_1-k\cos\boldsymbol{\theta}_0)+j(\sin\boldsymbol{\theta}_1-k\sin\boldsymbol{\theta}_0)$. Since $|\mathbf{A}|$ is real, we have the following relation: $\sin\boldsymbol{\theta}./\cos\boldsymbol{\theta}=(\sin\boldsymbol{\theta}_1-k\sin\boldsymbol{\theta}_0)./(\cos\boldsymbol{\theta}_1-k\cos\boldsymbol{\theta}_0)$, which can be further rewritten as

$$\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=k^{-1}\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_1). \qquad (29)$$

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

$$\boldsymbol{\theta}_0^{(1)}=\boldsymbol{\theta}-\boldsymbol{\alpha}\ ,\ \boldsymbol{\theta}_0^{(2)}=\boldsymbol{\theta}+\boldsymbol{\alpha}+\boldsymbol{\pi}, \qquad (30)$$

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

$$\begin{aligned}
&\left.\frac{\operatorname{Im}\{\mathbf{W}_1-k\mathbf{W}_0\}}{\operatorname{Im}\{\mathbf{X}_1-k\mathbf{X}_0\}}\right|_{\boldsymbol{\theta}_0^{(1)}=\boldsymbol{\theta}-\boldsymbol{\alpha}}\\
&=\left.\frac{\sin\boldsymbol{\theta}_1-k\sin\boldsymbol{\theta}_0}{\sin\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0^{(1)}=\boldsymbol{\theta}-\boldsymbol{\alpha}}\\
&=\frac{\sin\boldsymbol{\theta}_1-k\sin(\boldsymbol{\theta}-\boldsymbol{\alpha})}{\sin\boldsymbol{\theta}}\\
&=-\left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1)+k\cos\boldsymbol{\alpha}\right]\\
&=-\left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1)+\sqrt{k^2-1+\cos^2(\boldsymbol{\theta}-\boldsymbol{\theta}_1)}\right]\\
&\le 0.
\end{aligned} \qquad (31)$$

Therefore, the sign of the imagery part of $\mathbf{W}_1-k\mathbf{W}_0$ is different from the sign of imagery part of $\mathbf{X}_1-k\mathbf{X}_0$, resulting in negative values for ambient magnitude $|\mathbf{A}|$. Therefore, the first solution in (30) is inadmissible. Similarly, we take the second solution $\boldsymbol{\theta}_0^{(2)}$ and derive the ratio of imagery part of $\mathbf{W}_1-k\mathbf{W}_0$ to the imagery part of $\mathbf{X}_1-k\mathbf{X}_0$ as

$$\begin{aligned}
&\left.\frac{\operatorname{Im}\{\mathbf{W}_1-k\mathbf{W}_0\}}{\operatorname{Im}\{\mathbf{X}_1-k\mathbf{X}_0\}}\right|_{\boldsymbol{\theta}_0^{(2)}=\boldsymbol{\theta}+\boldsymbol{\alpha}+\boldsymbol{\pi}}\\
&=\left[\cos(\boldsymbol{\theta}-\boldsymbol{\theta}_1)+\sqrt{k^2-1+\cos^2(\boldsymbol{\theta}-\boldsymbol{\theta}_1)}\right]\ge 0.
\end{aligned} \qquad (32)$$

Therefore, the sign of the imagery part of $\mathbf{W}_1-k\mathbf{W}_0$ is the same from the sign of imagery part of $\mathbf{X}_1-k\mathbf{X}_0$, ensuring nonnegative values in ambient magnitude $|\mathbf{A}|$. Hence, we can conclude that based on the second solution, the relation between the ambient phases in two channels is $\boldsymbol{\theta}_0=\boldsymbol{\theta}+\arcsin\left[k^{-1}\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_1)\right]+\boldsymbol{\pi}$.

REFERENCES

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

[2] Dolby (2014 Jul. 11), Dolby Atmos next-generation audio for cinema, issue 3 [Online], Available: http://www.dolby.com/us/en/technologies/dolby-atmos.html

[3] 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.

[4] 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. 2013.

[5] 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.

[6] F. Rumsey, "Spatial audio: eighty years after Blumlein," *J. Audio Eng. Soc.*, vol. 59, no. 1/2, pp. 57–62, Jan./Feb. 2011.

[7] F. Rumsey, "Spatial audio processing: upmix, downmix, shake it all about," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 474–478, Jun. 2013.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* , Cambridge, MA: Academic Press, 1994.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] C. Faller, "Coding of spatial audio compatible with different playback formats," in *117th Audio Eng. Soc. Conv.,* San Francisco, CA, Oct. 2004.

[22] 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.

[23] N. Stefanakis, and A. Mouchtaris, "Foreground suppression for capturing and reproduction of crowded acoustic environments," in *Proc. ICASSP*, Brisbane, Australia, 2015, pp. 51-55.

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

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

[26] 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.

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

[28] 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.

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

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

[31] 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.

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

[33] F. Rumsey, "Time-frequency processing for spatial audio," *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 655–659, Jul./Aug. 2010.

[34] S. Y. 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.

[35] 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.

[36] C. Faller, "Matrix surround revisited," *in Audio Eng. Soc. 30th Int. Conf.*, Saariselka, Finland, Mar. 2007.

[37] 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.

[38] M. M. Goodwin, and J. M. Jot, "Analysis and synthesis for universal spatial audio coding," in *121st Audio Eng. Soc. Conv.,* San Francisco, CA, Oct. 2006.

[39] 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.

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

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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 *Proc. ICASSP*, Dallas, 2010, pp. 385-388.

[52] 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 *Proc. ICME*, Melbourne, Australia, 2012, pp. 628-633.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representation in audio and music: from coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995-1016, Jun. 2010.

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

[62] O. Yilmaz and S. Richard, "Blind separation of speech mixtures via time-frequency masking*," IEEE Trans. Sig. Process*., vol. 52, no. 7, pp.1830-1847, Jul. 2004.

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

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

[65] F. Rumsey, "Controlled subjective assessments of two-to-five channel surround sound processing algorithms," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, pp. 563–582, Jul./Aug. 1999.

[66] M. Schroeder, "An Artificial Stereophonic Effect Obtained from a Single Audio Signal," *J. Audio Eng. Soc.*, vol. 6, no. 2, pp. 74–79, Feb. 1958..

[67] G. Potard, I. Burnett, "Decorrelation techniques for the rendering of apparent sound source width in 3D audio displays," in *Proc. DAFx'04*, Naples, Italy, Oct. 2004.

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

[69] F. Menzer, and C. Faller, "Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching," in *126th Audio Eng. Soc. Conv.*, Munich, Germany, May 2009.

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

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

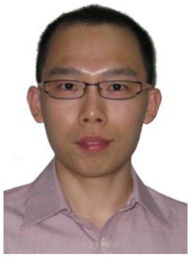
[72] J. Liebetrau, F. Nagel, N. Zacharov, K. Watanabe, C. Colomes, P. Crum, T. Sporer, and A. Mason, "Revision of Rec. ITU-R BS. 1534," in Proc. 137th AES convention, LA, Oct, 2014.

[73] 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.

[74] J. He. (2015 Jan. 14). Ambient Spectrum estimation ASE [online]. Available: http://jhe007.wix.com/main#!ambient-spectrum-estimation/c6bk.

[75] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Process. Magazine*, vol. 31, no. 3, pp. 107-115, 2014.

Jianjun He (S'12) received his B.ENG. degree in automation from Nanjing University of Posts and Telecommunications, China in 2011 and is currently pursuing his Ph.D. degree in electrical and electronic engineering at Nanyang Technological University (NTU), Singapore. In 2011, he was working as a general assistant in Nanjing International Center of Entrepreneurs (NICE), building platforms for start-ups from oversea Chinese scholars in Jiangning District, Nanjing, China. Since 2015, he has been a project officer with School of Electrical and Electronic Engineering in NTU. His Ph.D. work has been published in IEEE Signal Processing Magazine, IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), IEEE Signal Processing Letters, and ICASSP, etc. He has been an active reviewer for various Journals and conferences, including IEEE TASLP, Journal of Audio Engineering Society, etc. Aiming at improving humans' listening, his research interests include audio and acoustic signal processing, 3D audio (spatial audio), psychoacoustics, active noise control, source separation, and emerging audio and speech applications. Currently, He is a student member of the IEEE and Signal Processing Society (SPS), a member of APSIPA, and an affiliate member of IEEE SPS audio and acoustic technical committee.



Ee-Leng Tan received his BEng (1st Class Hons) and PhD degrees in Electrical and Electronic Engineering from Nanyang Technological University in 2003 and 2012, respectively. His research interests include image/audio processing and real-time digital signal processing. To date, his work has been awarded three patents in Japan, Singapore, and US. He currently holds the position of a Chief Science Officer, leading the research and development of the technological company Beijing Sesame World Co. Ltd. Concurrently, Dr Tan consults as the technical advisor for several start-ups.

Woon-Seng Gan (M'93-SM'00) received his BEng (1st Class Hons) and PhD degrees, both in Electrical and Electronic Engineering from the University of Strathclyde, UK in 1989 and 1993 respectively. He is currently an Associate Professor in the School of Electrical and Electronic Engineering in Nanyang Technological University. His research interests span a wide and related areas of adaptive signal processing, active noise control, and spatial audio. He has published more than 250 international refereed journals and conferences, and has granted seven Singapore/US patents. He had co-authored three books on *Digital Signal Processors: Architectures, Implementations, and Applications* (Prentice Hall, 2005), *Embedded Signal Processing with the Micro Signal Architecture*, (Wiley-IEEE, 2007), and *Subband Adaptive Filtering: Theory and Implementation* (John Wiley, 2009). He is currently a Fellow of the Audio Engineering Society(AES), a Fellow of the Institute of Engineering and Technology(IET), a Senior Member of the IEEE, and a Professional Engineer of Singapore. He is also an Associate Technical Editor of the Journal of Audio Engineering Society (JAES); Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing (ASLP); Editorial member of the Asia Pacific Signal and Information Processing Association (APSIPA) Transactions on Signal and Information Processing; and Associate Editor of the EURASIP Journal on Audio, Speech and Music Processing. He is currently a member of the Board of Governor of APSIPA.