Time-shifting based primary-ambient extraction for spatial audio reproduction

He, Jianjun; Gan, Woon-Seng; Tan, Ee-Leng

2015

He, J., Gan, W.-S., & Tan, E.-L. (2015). Time-Shifting Based Primary-Ambient Extraction for Spatial Audio Reproduction. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(10), 1576-1588.

https://hdl.handle.net/10356/81365

https://doi.org/10.1109/TASLP.2015.2439577

Time-Shifting Based Primary-Ambient Extraction for Spatial Audio Reproduction

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

Abstract

One of the key issues in spatial audio analysis and reproduction is to decompose a signal into primary and ambient components based on their directional and diffuse spatial features, respectively. Existing approaches employed in primary-ambient extraction (PAE), such as principal component analysis (PCA), are mainly based on a basic stereo signal model. The performance of these PAE approaches has not been well studied for the input signals that do not satisfy all the assumptions of the stereo signal model. In practice, one such case commonly encountered is that the primary components of the stereo signal are partially correlated at zero lag, referred to as the primary-complex case. In this paper, we take PCA as a representative of existing PAE approaches and investigate the performance degradation of PAE with respect to the correlation of the primary components in the primary-complex case. A time-shifting technique is proposed in PAE to alleviate the performance degradation due to the low correlation of the primary components in such stereo signals. This technique involves time-shifting the input signal according to the estimated inter-channel time difference of the primary component prior to the signal decomposition using conventional PAE approaches. To avoid the switching artifacts caused by the varied time-shifting in successive time frames, overlapped output mapping is suggested. Based on the results from our experiments, PAE approaches with the proposed time-shifting technique are found to be superior to the conventional PAE approaches in terms of extraction accuracy and spatial accuracy.

Index Terms

Primary-ambient extraction (PAE), spatial audio, principal component analysis (PCA), spatial cues.

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, 100010. (Email: jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg, joseph@sesame-world.com).

I. INTRODUCTION

With 3D video technology gaining prevalence, consumers are demanding a more immersive listening experience to better match the 3D visual effects, resulting in a growing need for 3D audio or spatial audio reproduction. Sound scenes in moving pictures and video games can generally be decomposed into directional and diffuse components, which are also referred to as primary (or direct) and ambient (or diffuse) components, respectively [1], [2]. To achieve accurate and flexible rendering of spatial audio, different processing techniques should be employed to reproduce the primary and ambient components of the audio signals [1], [3]-[5]. However, the primary and ambient components are usually mixed in conventional channel-based audio formats, such as stereo and surround sound formats [6]. Such channel-based audio formats make primary-ambient extraction (PAE) an essential step in spatial audio reproduction [4], [5]. In recent years, PAE has been incorporated into a wide range of applications, including spatial audio processing [7]-[9], spatial audio coding [10], [11], audio mixing [12]-[14], and emerging loudspeaker and headphone reproduction systems [15], [4].

There are two emerging frameworks for spatial audio coding: spatial audio scene coding (SASC) [11], [16] and directional audio coding (DirAC) [10]. Both SASC and DirAC extract the primary and ambient components and then synthesize the output based on the playback system configuration. In SASC, the localization analysis and synthesis, based on Gerzon localization vector [17], are independently performed on the primary and ambient components. In DirAC, the primary components are reproduced using vector base amplitude panning [18], while the ambient components are decorrelated and channeled to all loudspeakers to create the surrounding sound environment.

Incorporating PAE into various up-mixing techniques has been discussed in [2], [13], [14]. The PAE based up-mixing is particularly suitable for a hybrid loudspeaker system proposed by Gan *et al.* [19], [15]. This hybrid loudspeaker system uniquely combines parametric and conventional loudspeakers, taking advantage of the high directivity of the parametric loudspeakers to render accurate localization of the primary components and reproduce spaciousness of the ambient components using the conventional

loudspeakers [20]. Furthermore, PAE based spatial audio reproduction for headphone playback has been shown to create a more natural and immersive listening experience than conventional headphone rendering systems [9], [4].

To date, many approaches have been proposed for PAE. For these PAE approaches, the stereo input signal is generally modeled as a directional primary sound source linearly mixed with the ambient component. The assumptions of the stereo signal model are as follows. First, the primary and ambient components are considered to be independent with each other. Second, the primary components in the two channels are assumed to be correlated at zero lag. Third, the ambient components in the two channels are uncorrelated. Assuming that the ambient components in two channels of the stereo signal have equal level, Avendano and Jot [2] used a time-frequency mask to extract the ambient components from the stereo signal. Their time-frequency mask approach can also be extended to multichannel input signals [21]. A least-squares approach, proposed by Faller, estimated the primary and ambient components by minimizing the mean-square error [22]. Control of spatial cues of the ambient components was also combined with least-squares [23]. Recently, He *et al.* proposed a new ambient spectrum estimation framework and derived a sparsity constrained solution for PAE [24], [25]. Principal component analysis (PCA) based approaches remain the most widely studied approaches for PAE [1], [26]-[33]. Considering the independence between the primary and ambient components, the stereo signal is decomposed into two orthogonal components in each channel using the Karhunen-Loève transform [34]. Assuming that the primary component is relatively stronger in power than the ambient component, the component having larger variance is considered to be the primary component and the remaining component is considered as the ambient component. A comprehensive evaluation and comparison on these PAE approaches can be found in [35]. Other techniques such as non-negative matrix factorization [36] and independent component analysis [37] are also applied in PAE.

In practice, PAE is usually applied to the input signals without any reference or prior information. To achieve better extraction of the primary and ambient components, PAE requires the signal model to match

the input signal more closely. To date, little work has been reported to deal with input signals that do not fulfill all the assumptions of the stereo signal model. In [38], a normalized least-mean-square approach was proposed to address the problem in extracting the reverberation from stereo microphone recordings. Härmä [39] tried to improve the performance of PAE by classifying the time-frequency regions of the stereo signal into six classes. Thompson *et al.* [21] introduced a primary extraction approach that estimates the magnitude and phase of the primary component from a multichannel signal by using a linear system of the pairwise correlations. The latter approach requires at least three channels of the input signal and is not applicable to stereo input signals.

This paper focuses on PAE that deals with real-world stereo input signals that may not fit the typical PAE signal model. As seen in stereo microphone recordings, movies, and gaming tracks, the primary components in stereo signals can be amplitude panned and time-shifted. In addition, spectral differences can be found in the primary components that are obtained using binaural recording or binaural synthesis based on head-related transfer functions (HRTFs) [40]. We shall classify this type of stereo signals as the primary-complex signals. The primary components in the primary-complex signals usually exhibit partial correlation at zero lag. Other types of complex stereo signals, such as those involving (partially) correlated ambient components, are less common, and hence are not considered in this paper.

Therefore, we shall focus our study of PAE on two cases, namely, the ideal and primary-complex cases, where the primary components are completely correlated and partially correlated at zero lag, respectively. The performance of PAE is quantified by the measures of extraction accuracy and spatial accuracy. Performance degradation due to the mismatch of the input signal with the stereo signal model, and the proposed solution to deal with this mismatch is extensively studied in this paper. Some preliminary results of this study for primary component extraction have been reported in [41]. In this paper, we extend our study to both primary and ambient extraction. PCA is taken as a representative PAE approach in our study. More in-depth analysis on the performance of PCA based PAE is conducted for the extraction of both the primary and ambient components. In the primary-complex case, the performance degradation of PCA based

PAE with respect to the value of primary correlation is discussed, and we find the main cause of low primary correlation and the consequent performance degradation to be the time difference of the primary component. Hence, we propose a time-shifting technique to deal with PAE in the primary-complex case. The time-shifting technique is incorporated into PCA based PAE, resulting in a new approach referred to as time-shifted PCA (SPCA). A new overlapped output mapping method has also been proposed to avoid the switching artifacts caused by time-shifting. To validate the advantages of the proposed time-shifting technique and verify the improved performance of the proposed approach over conventional approaches more comprehensively, four experiments have been conducted using more realistic test signals, as compared to [41]. It shall be noted that the proposed time-shifting technique, though studied with PCA in this paper, can be incorporated into any other PAE approaches that are derived based on the stereo signal model.

The remainder of this paper is organized as follows. In Section II, we review the stereo signal model. PAE using PCA in the ideal case is discussed in Section III. Section IV presents the performance analysis of PCA based PAE in the primary-complex case. The proposed SPCA based PAE to address the problem in the primary-complex case is discussed in Section V. Section VI presents our comparative evaluation on the performance of PCA and SPCA based PAE using four experiments. Finally, we conclude this work in Section VII.

II. STEREO SIGNAL MODEL

In this section, we introduce the basic stereo signal model and its key assumptions for audio signals in digital media. A summary of key symbols used in this paper is presented in Table I. In general, we consider the stereo signals as mixtures of two constitutes [1]: (i) directional point-like sound sources referred to as the primary component; and (ii) a diffuse sound environment referred to as the ambient component. To deal with multiple concurrent sound sources in the primary component, a common practice in spatial audio processing is to preprocess the stereo signals using subband decomposition (or time-frequency transform)

Table I
Summary of key symbols used in this paper

$c \in \{0,1\}$	Channel index	m	Time frame index
$\mathbf{x}_c$	Mixed signal	b	Subband index
$\mathbf{p}_c$	Primary component	n	Sample index
$\mathbf{a}_c$	Ambient component	N	Frame length
r	Correlation	ϕ_x	Correlation coefficient of the mixed signal
τ	Lag index	τ_o	ICTD
k	Primary panning factor	γ	Primary power ratio
$\hat{k}_{ic}$, $\hat{\gamma}_{ic}$	Estimates of k and γ in ideal case, as in (7) and (8)	$\hat{k}_{pc}$, $\hat{\gamma}_{pc}$	Estimates of k and γ in the primary-complex case, as in (14) and (15)
$\hat{\mathbf{p}}_{\text{PCA},\,c}$	Primary component extracted using PCA	$\hat{\mathbf{a}}_{\text{PCA},\,c}$	Ambient component extracted using PCA
ϕ_{P}	Correlation coefficient of the primary component (at zero lag)	Δk, $\Delta\gamma$	Ratio between the estimated k, γ and their true values in the primary-complex case, as in (16) and (17)
$\hat{p}_{\text{SPCA},\,c}$	Primary component (one sample) extracted using SPCA	$\hat{a}_{\text{SPCA},\,c}$	Ambient component (one sample) extracted using SPCA

[1], [2], [22], [28], [42]. PAE is applied for signals in every subband independently by considering only one dominant sound source in the primary component in one subband [1], [2], [22], [28]. Finally, the extracted primary and ambient components in the subbands are combined. Denoting one subband of the stereo signal as $\mathbf{x}_c[m,b] = [x_c(mN,b), x_c(mN+1,b), \ldots, x_c(mN+N-1,b)]^T$, where $c \in \{0,1\}$, m, b, N and T are the channel index, time frame index, subband index, frame length, and the transpose operator, respectively. The basic signal model is formulated as:

$$\begin{aligned} \mathbf{x}_0[m,b] &= \mathbf{p}_0[m,b] + \mathbf{a}_0[m,b], \text{ and} \\ \mathbf{x}_1[m,b] &= \mathbf{p}_1[m,b] + \mathbf{a}_1[m,b], \end{aligned} \tag{1}$$

where $\mathbf{p}_0$, $\mathbf{p}_1$ and $\mathbf{a}_0$, $\mathbf{a}_1$ are the primary and ambient components in two channels of the stereo signal, respectively. Since the frame-based subband analysis is generally used in the discussions of PAE approaches in this paper, the indices $[m,b]$ are omitted for brevity.

In the stereo signal model, the primary and ambient components are differentiated by their correlations. The correlation coefficient between signals $\mathbf{x}_0$ and $\mathbf{x}_1$ is defined as follows:

$$\phi_x(\tau) = \frac{r_{01}(\tau)}{r_{00} r_{11}} = \frac{\sum_{n=0}^{N-1}\left[x_0(n)\, x_1(n+\tau)\right]}{\sqrt{\sum_{n=0}^{N-1}\left[x_0^2(n)\right]\sum_{n=0}^{N-1}\left[x_1^2(n+\tau)\right]}}, \tag{2}$$

where $r_{01}(\tau)$ is the cross-correlation of $\mathbf{x}_0$ and $\mathbf{x}_1$ at lag τ, and r_{00}, r_{11} are the auto-correlations for the two channels. Two signals having the highest absolute value of correlation coefficient $\max\left|\phi_x(\tau)\right|$ as one and zero are considered as correlated and uncorrelated signals, respectively. A correlation coefficient between zero and one indicates that the two signals are partially correlated. In the stereo signal model, it is assumed that the primary and ambient components in the two channels are correlated and uncorrelated, respectively [1].

As pointed out by Blauert [43], the correlated primary component in the stereo signal satisfies either one or both of the following conditions: i) amplitude panned, i.e., $\mathbf{p}_1 = k\mathbf{p}_0$, where k is the primary panning factor; ii) time-shifted, i.e., $p_1(n) = p_0(n+\tau_o)$, where $p_1(n)$ is the nth sample in $\mathbf{p}_1$ and τ_o is the inter-channel time difference (ICTD) between the two channels. In this stereo signal model, the correlated primary component is assumed to be only amplitude panned between the two channels of the stereo signal, and the primary component is uncorrelated with the ambient component [1]. Considering the diffuseness of the ambient component, the ambient power in the two channels of the stereo signal is relatively balanced.

To determine the power difference between the primary and ambient components, we introduce the primary power ratio, which is defined as the ratio of primary power to the sum of the primary and ambient power:

$$\gamma = \frac{P_{\mathbf{p}_0} + P_{\mathbf{p}_1}}{P_{\mathbf{p}_0} + P_{\mathbf{p}_1} + P_{\mathbf{a}_0} + P_{\mathbf{a}_1}},\ \gamma \in [0,1], \tag{3}$$

where $P_{(.)}$ denotes the mean square power of the signal in the subscript. The assumptions for the stereo signal model are summarized as:

$$\mathbf{p}_1 = k\mathbf{p}_0,\ \mathbf{a}_0 \perp \mathbf{a}_1,\ \mathbf{p}_i \perp \mathbf{a}_j,\ \forall i, j \in \{0,1\}, \tag{4}$$

$$P_{\mathbf{p}_1} = k^2 P_{\mathbf{p}_0},\ P_{\mathbf{a}_1} = P_{\mathbf{a}_0}, \tag{5}$$

where $\perp$ represents uncorrelated signals. In the ideal case, assumptions in (4) and (5) are completely satisfied. Thus, we can express the auto- and cross-correlations of the stereo input signal at zero lag as (the lag index 0 is omitted):

$$r_{00} = N\left(P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{11} = N\left(k^2 P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{01} = NkP_{\mathbf{p}_0}. \tag{6}$$

From (6), we obtain the estimates of k and γ of the stereo signal model as

$$\hat{k}_{ic} = \frac{r_{11} - r_{00}}{2r_{01}} + \sqrt{\left(\frac{r_{11} - r_{00}}{2r_{01}}\right)^2 + 1}, \tag{7}$$

$$\hat{\gamma}_{ic} = \frac{2r_{01} + \left(r_{11} - r_{00}\right)\hat{k}_{ic}}{\left(r_{11} + r_{00}\right)\hat{k}_{ic}}, \tag{8}$$

where the subscript "*ic*" stands for "ideal case". In the ideal case, the estimates of k and γ are equal to their true values, i.e., $\hat{k}_{ic} = k,\ \hat{\gamma}_{ic} = \gamma$.

Based on the stereo signal model, we can characterize the stereo signal using k and γ. Primary panning factor k can be considered as the square root of the inter-channel level difference (ICLD) of the primary components, which is one of the most important localization cues in spatial audio [43]. The primary component is panned to channels 0 and 1 for $k < 1$, and $k > 1$, respectively. As γ increases, the primary component becomes more prominent in the input signal. In the following sections, we shall use k and γ to determine the performance of PAE.

III. PCA BASED PAE IN IDEAL CASE

PCA is a widely used method in multivariate analysis [34], and it was first introduced to solve the PAE problem in [26]. The central idea of PCA is to linearly transform the input signal into orthogonal principal components with descending variances. In this section, we assume that the input signal satisfies all the

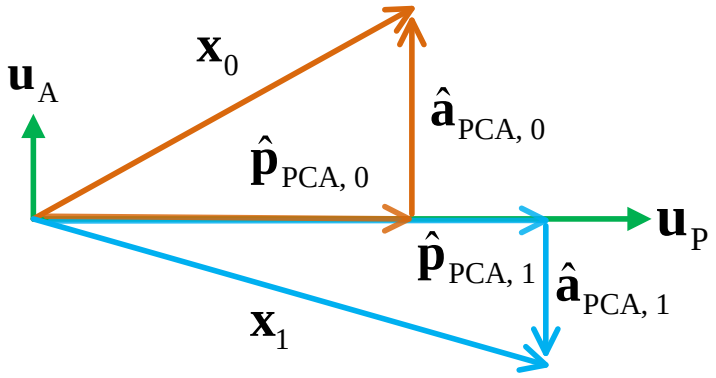


Fig. 1. A geometric representation of PCA based PAE.

assumptions discussed in Section II, which is referred to as the ideal case for PAE. Based on the stereo signal model, PAE using PCA decomposition can be mathematically described as [1]:

$$\begin{aligned}
&\mathbf{u}_P = \arg\max_{\mathbf{u}_P} \left(\left\| \mathbf{u}_P{}^T \mathbf{x}_0 \right\|^2 + \left\| \mathbf{u}_P{}^T \mathbf{x}_1 \right\|^2 \right), \\
&\mathbf{u}_A = \arg\min_{\mathbf{u}_A} \left(\left\| \mathbf{u}_A{}^T \mathbf{x}_0 \right\|^2 + \left\| \mathbf{u}_A{}^T \mathbf{x}_1 \right\|^2 \right), \\
&\text{s.t. } \mathbf{u}_P \perp \mathbf{u}_A, \ \left\| \mathbf{u}_P \right\| = \left\| \mathbf{u}_A \right\| = 1,
\end{aligned} \tag{9}$$

where $\mathbf{u}_P$ and $\mathbf{u}_A$ are the primary and ambient basis vectors, respectively. As depicted in Fig. 1, $\mathbf{u}_P$ and $\mathbf{u}_A$ maximizes and minimizes the total projection energy of the input signal vectors, respectively. The solution to (9) can be obtained by eigenvalue decomposition of the input covariance matrix [28].

In general, the primary component possesses more power than the ambient component. Hence, the larger eigenvalue and the corresponding basis vector are related to the primary component and the smaller eigenvalue related to the ambient component. The simplified solutions for the extracted primary and ambient components (denoted by a hat symbol on the top) extracted using PCA [35], [41] are given by

$$\hat{\mathbf{p}}_{PCA,\,0} = \frac{1}{1+\hat{k}_{ic}^2} \left(\mathbf{x}_0 + \hat{k}_{ic} \mathbf{x}_1 \right), \ \hat{\mathbf{p}}_{PCA,\,1} = \hat{}_{ic} \hat{\mathbf{p}}_{PCA,\,0}. \tag{10}$$

$$\hat{\mathbf{a}}_{PCA,\,0} = \frac{\hat{k}_{ic}}{1+\hat{}_{ic}^2} \left(\hat{k}_{ic} \mathbf{x}_0 - \mathbf{x}_1 \right), \ \hat{\mathbf{a}}_{PCA,\,1} = -\frac{1}{\hat{k}_{ic}} \hat{\mathbf{a}}_{PCA,\,0}. \tag{11}$$

From (10)-(11), we observe that the extracted primary and ambient components are weighted sums of the

input signals. Note that the weighted sum form of solution in PCA can be generalized into the linear estimation based PAE, as studied in [35]. In the ideal case, the performance of PCA and other linear estimation based PAE approaches are affected by k and γ [35]. Detailed analysis on the performance of these PAE approaches is studied in depth in [35].

IV. PCA BASED PAE IN THE PRIMARY-COMPLEX CASE

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

Considering a stereo signal having a partially correlated primary component at zero lag, the first assumption of the stereo signal model as stated in (4) becomes

$$0 < \left[\phi_{\mathrm{P}} = \frac{\mathbf{p}_0{}^T \mathbf{p}_1}{\sqrt{\left(\mathbf{p}_0{}^T \mathbf{p}_0\right)\left(\mathbf{p}_1{}^T \mathbf{p}_1\right)}} \right] < 1, \tag{12}$$

where ϕ_{P} is the correlation coefficient of the primary component at zero lag (primary correlation for short), and the rest of the assumptions in (4) and (5) remain unchanged. Here, only the positive primary correlation is considered, since the negatively correlated primary component can be converted into positive by simply multiplying the primary component in either channel by -1. In primary-complex case, the correlations of the input signals at zero lag are computed as:

$$r_{00} = N\left(P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{11} = N\left(k^2 P_{\mathbf{p}_0} + P_{\mathbf{a}_0}\right),\ r_{01} = N\phi_{\mathrm{P}} k P_{\mathbf{p}_0}. \tag{13}$$

Hence, the estimated k and γ are:

$$\hat{k}_{pc} = \phi_{\mathrm{P}} \frac{r_{11} - r_{00}}{_{01}} + \sqrt{\left(\phi_p \frac{r_{11} - r_{00}}{2r_{01}} \right)^2 + 1}, \tag{14}$$

$$\hat{\gamma}_{pc} = \frac{2r_{01} + \phi_{\mathrm{P}} \left(r_{11} - r_{00} \right) \hat{k}_{pc}}{\phi_{\mathrm{P}} \left(r_{11} + r_{01} \right) \hat{k}_{pc}}, \tag{15}$$

where the subscript "*pc*" stands for "the primary-complex case". Clearly, accurate estimation of *k* and γ in the primary-complex case requires the additional knowledge about the primary correlation ϕ_{P}. However, this primary correlation is usually unavailable as only the mixed signal is given as input. In PCA based PAE, the estimates of *k* and γ for the ideal case, given in (7)-(8), are usually employed. To see how accurate these ideal case estimates are, we substitute (13) into (7) and (8), and compute the ratio between the estimated *k* and true *k*, and the ratio between estimated γ and true γ as

$$\Delta k = \frac{\hat{k}_{ic}}{k} = \frac{k^2 - 1}{2\phi_{\mathrm{P}} k^2} + \sqrt{\left(\frac{k^2 - 1}{2\phi_{\mathrm{P}} k^2} \right)^2 + \frac{1}{k^2}}, \tag{16}$$

$$\Delta\gamma = \frac{\hat{\gamma}_{ic}}{\gamma} = \frac{k^2 - 1 + 2\phi_{\mathrm{P}}}{k^2 + 1}. \tag{17}$$

Using (16) and (17), the ratios of the *k* and γ in the primary-complex case with respect to the primary correlation are plotted in Fig. 2. It is clear that *k* is only correctly estimated (i.e., Δk = 0 dB) when it equals one; and the estimation of γ is more accurate (i.e., $\Delta\gamma$ closer to 1) as *k* increases. The estimations of *k* and γ become less accurate as the primary correlation decreases from one to zero. The inaccuracy in the estimates of *k* and γ results in an incorrect ICLD of the extracted primary components and hence degrades the extraction performance.

Next, we analyze the extraction performance of PCA based PAE in the primary-complex case. First, we rewrite (10)-(11) using the true primary and ambient components:

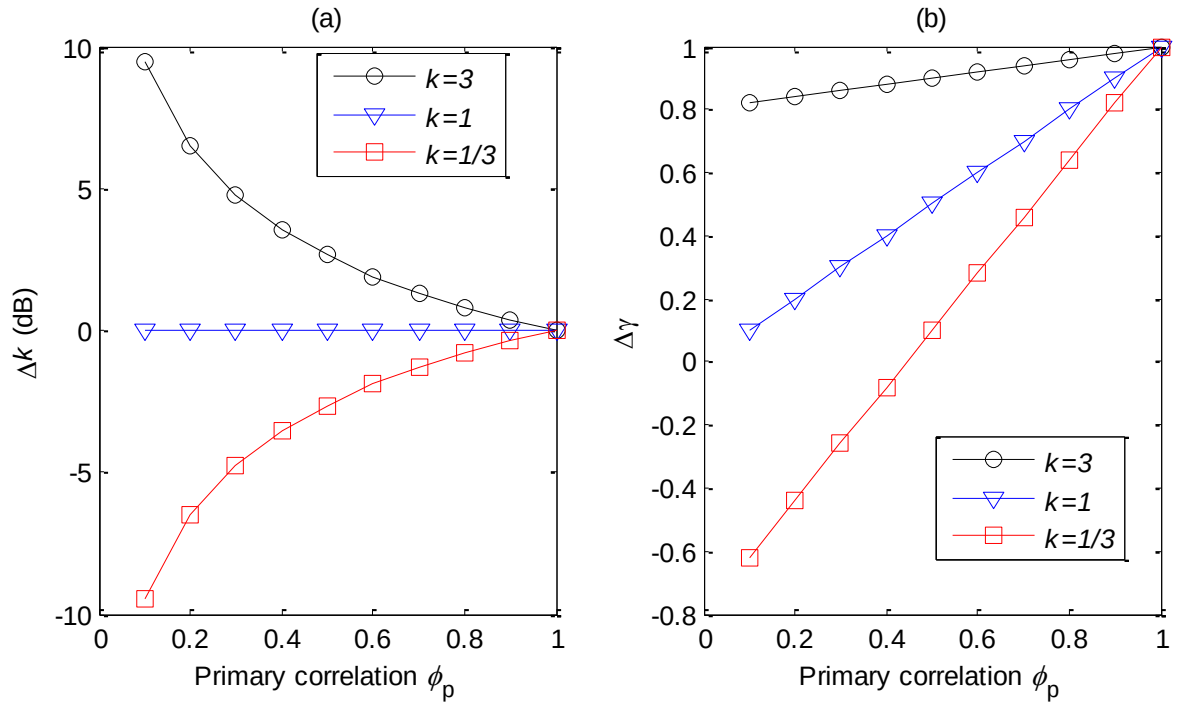


Fig. 2. Estimation of (a) primary panning factor k, and (b) primary power ratio γ in the primary-complex case with varying ϕ_p. The estimations are more accurate when Δk and $\Delta\gamma$ are closer to 0 dB and 1, respectively.

$$\begin{aligned}\hat{\mathbf{p}}_{\mathrm{PCA},0} &= \mathbf{p}_0 - \mathbf{v} + \frac{1}{1+\hat{k}_{ic}^{\,2}}\left(\mathbf{a}_0 + \hat{k}_{ic}\mathbf{a}_1\right),\\ \hat{\mathbf{p}}_{\mathrm{PCA},1} &= \mathbf{p}_1 + \frac{1}{\hat{k}_{ic}}\mathbf{v} + \frac{\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\left(\mathbf{a}_0 + \hat{k}_{ic}\mathbf{a}_1\right),\end{aligned} \tag{18}$$

$$\begin{aligned}\hat{\mathbf{a}}_{\mathrm{PCA},0} &= \frac{\hat{k}_{ic}^{\,2}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_0 + \mathbf{v} + \frac{-\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_1,\\ \hat{\mathbf{a}}_{\mathrm{PCA},1} &= \frac{1}{+\hat{k}_{ic}^{\,2}}\mathbf{a}_1 - \frac{1}{\hat{k}_{ic}}\mathbf{v} + \frac{-\hat{k}_{ic}}{1+\hat{k}_{ic}^{\,2}}\mathbf{a}_0,\end{aligned} \tag{19}$$

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

To evaluate the PAE performance, two groups of performance measures quantifying the extraction accuracy and spatial accuracy are introduced [35]. The extraction accuracy is usually quantified by the

extraction error, which is given by the error-to-signal ratio (ESR). This measure ESR is defined as the average of the ratios of the extraction error power to the power of the true component in the two channels of the stereo signal, and the ESR of the extracted primary and ambient components are computed as:

$$\begin{aligned} \mathrm{ESR}_{\mathrm{P}} &= 0.5\left(\frac{P_{\mathbf{p}_0-\hat{\mathbf{p}}_0}}{P_{\mathbf{p}_0}} + \frac{P_{\mathbf{p}_1-\hat{\mathbf{p}}_1}}{P_{\mathbf{p}_1}} \right), \\ \mathrm{ESR}_{\mathrm{A}} &= 0.5\left(\frac{P_{\mathbf{a}_0-\hat{\mathbf{a}}_0}}{P_{\mathbf{a}_0}} + \frac{P_{\mathbf{a}_1-\hat{\mathbf{a}}_1}}{P_{\mathbf{a}_1}} \right). \end{aligned} \tag{20}$$

Smaller value of ESR indicates a better extraction.

In the second group of measures, we consider the spatial accuracy by comparing the inter-channel relations of the extracted primary and ambient components with those of the true components. Due to the differences in the spatial characteristics of the primary and ambient components, we shall evaluate these components separately. For the primary components, there are three widely used spatial cues, namely, inter-channel cross-correlation coefficient (ICC), ICTD, and ICLD. The accuracy of these cues can be used to evaluate the sound localization accuracy of the extracted primary components [7], [44]. There has been extensive research in ICTD estimation after the coincidence model proposed by Jeffress (see [45]-[48] and references therein). Based on the Jeffress model [45], the ICC of different time lags is calculated and the lag number that corresponds to the maximum ICC is determined as the estimated ICTD. ICLD is obtained by taking the ratio of the signal power between the channels 1 and 0. For the extracted ambient components, we evaluate the diffuseness of these components using ICC and ICLD [49]. Since the ambient component is uncorrelated and relatively balanced in the two channels of the stereo signal, a better extraction of the ambient component is achieved when ICC and ICLD of the ambient component is closer to zero and one, respectively.

In Table II, we summarize the results of the performance measures for the extracted primary and ambient components when PCA based PAE is applied in the primary-complex (i.e., $\phi_{\mathrm{P}} \neq 1$) and ideal cases (i.e., $\phi_{\mathrm{P}} = 1$). To illustrate how the extraction accuracy is influenced by ϕ_{P}, the results of ESR using

TABLE II
Performance of PCA based PAE in the primary-complex case.

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

$\gamma \in \{0.2,\ 0.5,\ 0.8\}$ and k = 3, are plotted in Fig. 3. It is clear that ESR is affected by the primary correlation ϕ_{P}. As shown in Fig. 3(a), the error of the extracted primary component decreases as ϕ_{P} approaches one, except for γ = 0.2. This exceptional case arises when γ is low, and the ambient leakage in the extracted primary component becomes the main contributor for the extraction error. From Fig. 2(a), we notice that as ϕ_{P} increases, Δk decreases, which leads to the decrease of $\hat{k}_{ic} = \Delta k \bullet k$; and hence the contributor from the ambient leakage in $\mathrm{ESR_P}$ (i.e., $\left(1+k^{-2}+\frac{k^{2}-k^{-2}}{\hat{k}_{ic}^{\,2}+1}\right)\frac{1-\gamma}{4\gamma}$) increases, which finally leads to the increase of $\mathrm{ESR_P}$ for γ = 0.2. For the ESR of the extracted ambient component ($\mathrm{ESR_A}$) as illustrated in Fig. 3(b), we observed that $\mathrm{ESR_A}$ decreases gradually as ϕ_{P} increases, which leads to an extracted ambient component having less error. Based on these observations, we find that

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

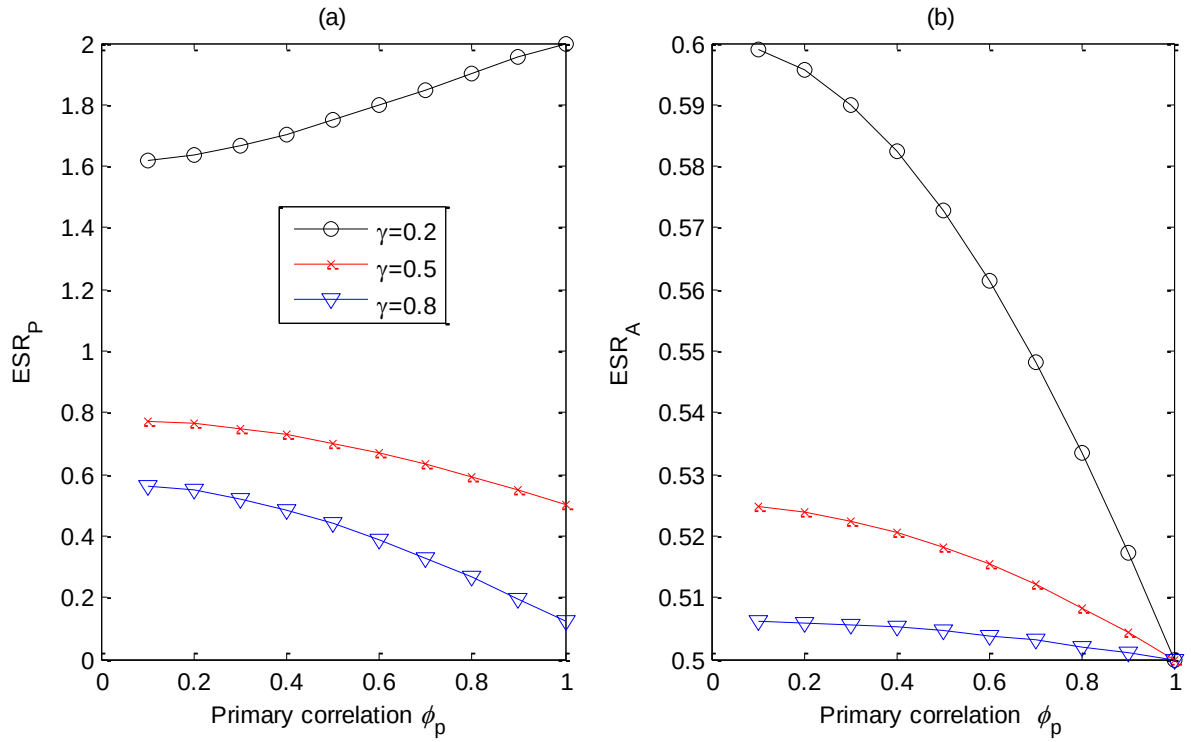


Fig. 3. ESR of (a) primary extraction and (b) ambient extraction using PCA based PAE in the primary-complex case with varying ϕ_p according to the results in Table II. Legend in (a) applies to both plots.

From the above observations, it is concluded that the performance of PCA based PAE is degraded by the partially correlated primary components of the stereo signal in the primary-complex case. The degraded performance, as observed in PCA, actually originates from the inaccurate estimations of k and γ. As found in [31], the linear estimation based PAE approaches are determined by these two parameters. Hence, it can be inferred that these linear estimation based PAE approaches as well as other PAE approaches that are derived based on the basic stereo signal model will encounter a similar performance degradation when dealing with stereo signals having partially correlated primary components.

V. TIME-SHIFTED PCA BASED PAE IN THE PRIMARY-COMPLEX CASE

In the audio of moving pictures and video games, it is commonly observed that the primary components are amplitude panned and/or time-shifted [50], [51], where the latter leads to low correlation of the primary components at zero lag. As mentioned in the previous section, PCA based PAE dealing with such primary-

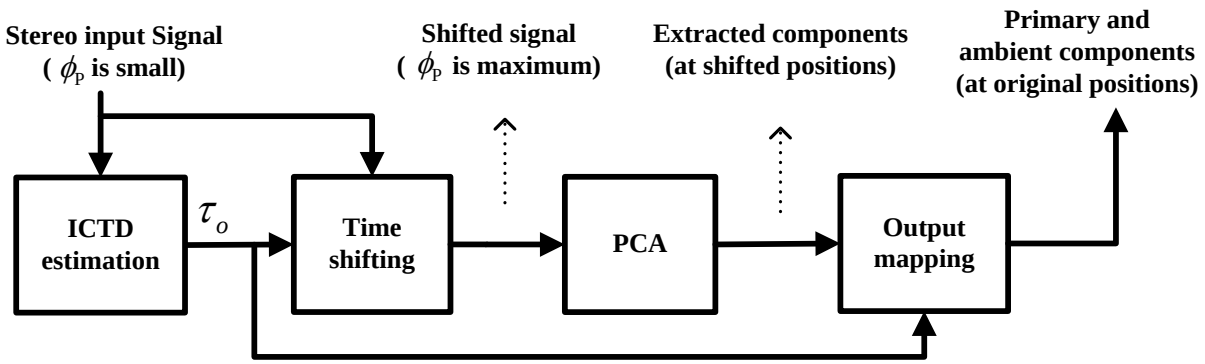


Fig. 4. Block diagram of SPCA based PAE.

complex signals leads to significant extraction error. Furthermore, the ICTD of the primary component is completely lost after the extraction. To overcome these issues, we propose a time-shifting technique to be incorporated into PCA based PAE, which results in the proposed approach, namely, the time-shifted PCA (SPCA) based PAE. The proposed approach aims to retain the ICTD in the extracted primary component and time-shifts the primary components to increase the primary correlation, thereby enhancing the performance of PAE. Some preliminary results have been reported in [41].

The block diagram of the proposed SPCA based PAE is shown in Fig. 4. In SPCA based PAE, the stereo input signal is first time-shifted according to the estimated ICTD of the primary component. Subsequently, PCA is applied to the shifted signal and extracts primary and ambient components at shifted positions. Finally, the time indices of extracted primary and ambient components are mapped to their original positions using the same ICTD. Let τ_o denotes the estimated ICTD, the final output for the *n*th sample in the extracted components can be expressed as

$$\hat{p}_{\mathrm{SPCA},\,0}(n)=\frac{1}{\hat{k}_{ic}^{\,2}}\left[x_0(n)+\hat{k}_{ic}x_1(n-\tau_o)\right],\ \hat{p}_{\mathrm{SPCA},\,1}(n)=\frac{\hat{k}_{ic}}{1+\hat{k}_{ic}}\left[x_0(n+\tau_o)+\hat{k}_{ic}x_1(n)\right], \tag{21}$$

$$\hat{a}_{\mathrm{SPCA},\,0}(n)=\frac{\hat{k}_{ic}}{\hat{k}_{ic}}\left[\hat{k}_{ic}x_0(n)-x_1(n-\tau_o)\right],\ \hat{a}_{\mathrm{SPCA},\,1}(n)=-\frac{1}{1+\hat{k}_{ic}^{\,2}}\left[\hat{k}_{ic}x_0(n+\ _o)-x_1(n)\right]. \tag{22}$$

It can be seen that the proposed approach is related to delayed-and-sum beamformer [52] in the sense that each extracted component is a weighted sum of the input signals but with a delay or advance being applied

in either channel. When ICTD $\tau_o = 0$, the proposed SPCA based PAE reduces to the conventional PCA based PAE.

As mentioned in previous section, estimation of ICTD can be obtained using various approaches. In this paper, we apply the Jeffress model [45], which estimates the ICTD of the primary component using the maximum ICC of the primary component at various lags $\phi_{\mathrm{P}}(\tau)$. When only the stereo signal is available, we cannot compute the ICC of the primary component directly. Instead, the ICC of the stereo input signal $\phi_x(\tau)$ is used to estimate the ICTD of the primary component. Due to the uncorrelated ambient component of the stereo signal, which remains uncorrelated after the stereo signal is time-shifted, we find that for each lag τ,

$$\phi_x(\tau) = g\phi_{\mathrm{P}}(\tau), \tag{23}$$

where $g = \sqrt{\frac{P_{\mathbf{p}_0} P_{\mathbf{p}_1}}{P_{\mathbf{x}_0} P_{\mathbf{x}_1}}}$ is lag-invariant. Therefore, the ICTD $\tau_o = \arg\max_{\tau} \phi_{\mathrm{P}}(\tau) = \arg\max_{\tau} \phi(\tau)$. A detailed study on the estimation of ICTD based on ICC in complex situations is discussed in [50]. Due to the effect of summing localization, the maximum number of lags considered for ICC and ICTD in spatial audio is usually limited to ±1 ms [43]. The positive and negative values of ICTD account for the primary components that are panned to the directions of channel 0 and channel 1 in the auditory scene, respectively. As compared to the conventional PCA based PAE, the estimation of ICTD is one critical additional step, which inevitably incurs more calculations. More specifically, in the conventional PCA, the cross-correlations (i.e., $\phi_x(0)$) is only computed once. By contrast, the proposed SPCA requires a total of 89 times of cross-correlations (i.e., $\phi_x(\tau)$, $\forall \tau \in [-44, 44]$, at a sampling rate fs = 44.1 kHz). One way to reduce the additional computation load is to increase the sample step size in ICTD estimation. For instance, computing only the cross-correlations with odd (or even) indices can reduce the additional computation load by half, at the cost of reducing the resolution of ICTD estimation.

The time-shifting operation is achieved by keeping the signal in channel 0 unchanged but delaying (or advancing) the signal in channel 1 by a duration equal to ICTD when ICTD ≤ 0 (or ICTD > 0). When the amounts of shifts in two successive frames are not the same, a proper mapping strategy is required to shift back the primary and ambient components that are extracted from the shifted signal to the original positions. To show how the change of ICTD affects the final output mapping, we consider two extreme cases, as illustrated in Fig. 5. The table in the top middle of Fig. 5 shows the ICTDs of three successive frames considered for these two cases. In the first case, we consider maximum ICTD decrease, i.e., the ICTD of frame *i-1* is 1 ms, which is decreased to -1 ms in frame *i*. In the second case, we consider maximum ICTD increase, that is, as compared to the frame *i*, the ICTD of frame *i+1* is increased to 1 ms. Consequently, the decrease and increase of ICTDs in these two cases lead to a 2 ms overlap and gap in channel 1 between these frames, respectively, as shown in Fig. 5(a). To generalize these two extreme cases, let us consider the change of ICTD in two successive frames as $\Delta\tau_o(i) = \tau_o(i) - \tau_o(i-1)$. Hence, we have

$$\text{Samples between the two frames of the extracted components in channel 1} = \begin{cases} \text{overlap of } \left|\Delta\tau_o(i)\right|, & \Delta\tau_o(i) < 0 \\ \text{no overlap or gap}, & \Delta\tau_o(i) = 0. \\ \text{gap of } \Delta_{\ o}(i), & \Delta\tau_o(i) > 0 \end{cases} \tag{24}$$

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

To avoid the switching artifacts, all successive frames should be overlapped such that no gap between the frames can be found even when the ICTD increase reaches its maximum. The proposed overlapped output mapping strategy is depicted in Fig. 5(b). Let the duration of the overlapping samples in the stereo signals be *Q* ms. As compared to the conventional output mapping in Fig. 5(a), different amount of

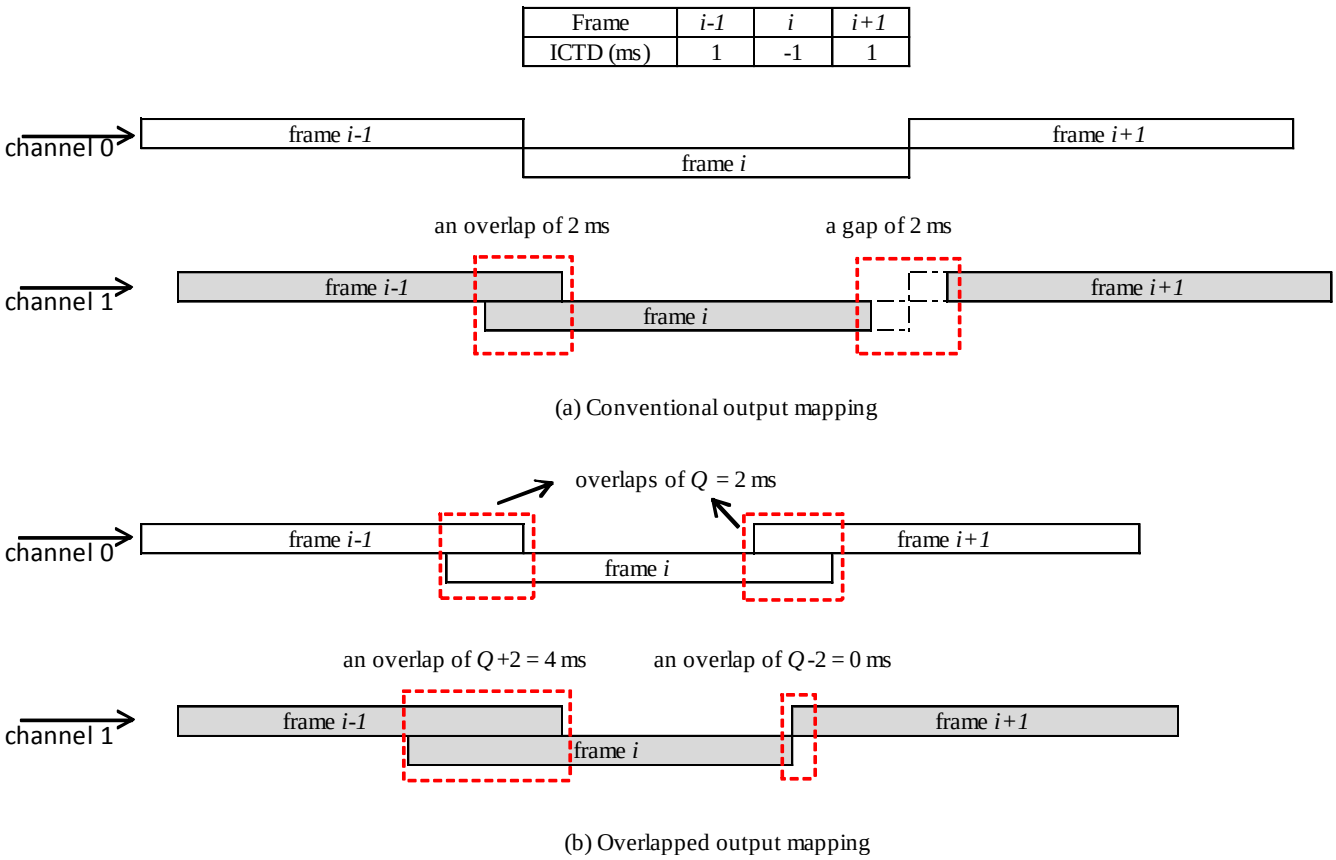


Fig. 5. An illustration of two output mapping strategies in the extreme cases: (a) conventional; (b) overlapped. The two channels 0 and 1 are depicted in white and grey, respectively. The table in the top middle shows the ICTDs for three successive frames. The value of Q in this example is selected as 2 ms.

overlapping samples are found in both channels in Fig. 5(b). In channel 0, exact Q ms between each two frames is overlapped, while in channel 1, the duration of overlapping samples varies from frame to frame according to the change in the ICTDs. That is,

$$\begin{aligned}&\text{Samples between the two frames of the extracted components in channel 1}\\&= \text{overlap of } \left\lfloor \text{Q*}10^{-3}\text{*fs-}\Delta\tau_o(i) \right\rceil.\end{aligned} \tag{25}$$

To correspond to the two extreme cases, the duration of overlapping samples in channel 1 would be from Q-2 ms to Q+2 ms. In order to ensure no gap is found between any two successive frames, the duration of overlapping samples must be equal to or greater than 2 ms, i.e., $Q \geq 2$ ms. As shown in Fig. 5(b), where Q is chosen as the lowest value, i.e., $Q = 2$ ms, we find that even in the extreme case of maximum ICTD increase from frame i to frame $i+1$, there is no gap in channel 1. Therefore, no matter how much the ICTD

Table III
Specifications of the four experiments

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

changes, all frames can be handled appropriately without gap artifacts. Increasing Q would also smoothen the extracted components, especially when the direction of the primary components changes rapidly. It is noted from (25) that the actual overlapping samples in different frames and channels can be varying. Thus, the cross-fading technique is required to adapt to these variations of the overlapping samples.

Based on the above discussions, we shall see that the proposed time-shifting and overlapped output mapping techniques work independently from PCA. Therefore, the same time-shifting and output mapping technique in the proposed SPCA can be applied seamlessly to improve the performance of many other existing PAE approaches, including time-frequency masking [2], PCA based approaches [29]-[31], and other linear estimation based PAE approaches as discussed in [35]. However, it shall be noted that the ICTD estimation and time-shifting operations would incur additional computation and memory cost.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

To validate the performance of the proposed SPCA based PAE, a number of experiments were conducted. As the focus of this paper is to examine PAE with partially correlated primary components rather than the subband decomposition of the stereo signal, we shall consider only one dominant source in the primary component of the stereo signal and perform PAE without subband decomposition in the experiments. Experimental results for PAE with time-shifting on multiple dominant sources and subband decomposition can be found in [53]. Subjective listening test that compares PCA with SPCA in localization of extracted primary components is presented in [54]. In this section, we present the results from four different experiments. To perform an accurate comparative analysis between PCA and SPCA, we manually

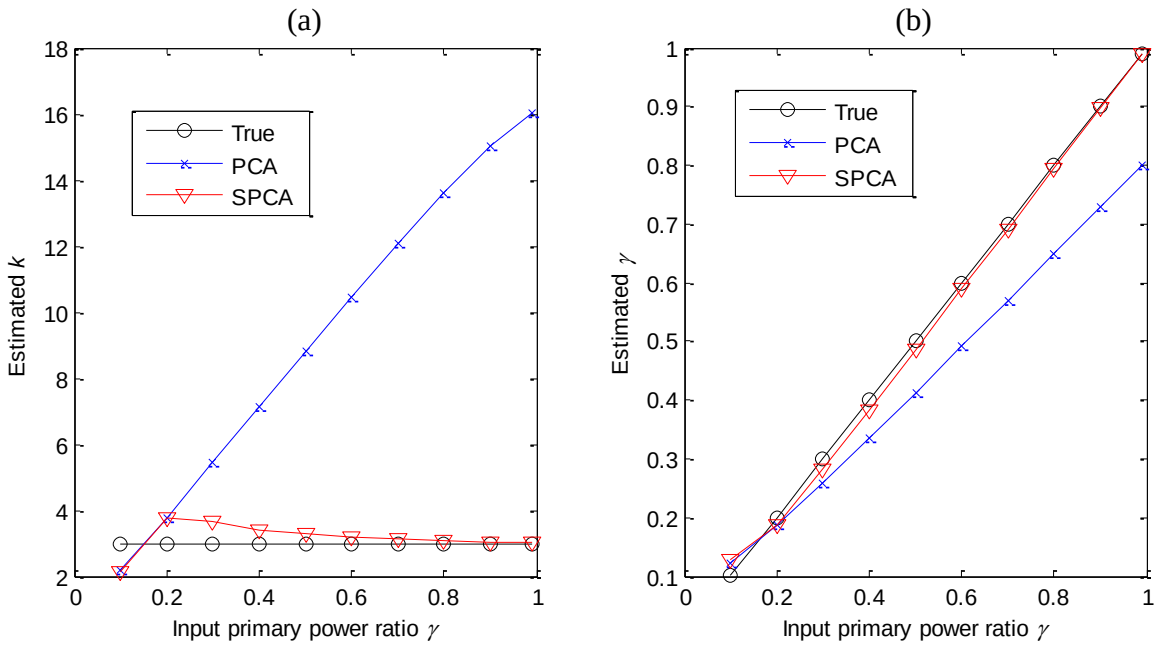


Fig. 6. Comparison of the estimation of (a) k and (b) γ between PCA and SPCA based PAE in the primary-complex case.

synthesized directional signals and mixed them with ambient signals in the first two experiments. The first and second experiments considered static and moving primary component, respectively. In the first experiment, we compared the extraction performance of PCA and SPCA with respect to γ. While the direction of the primary component was fixed in the first experiment, the second experiment examined the estimation of the panning directions of the primary components using PCA and SPCA with γ being close across the frames. The third experiment evaluated how PCA and SPCA perform when dealing with reverberation type of ambient components. To evaluate these two PAE approaches in a more realistic scenario, the fourth experiment was conducted using recorded signals of primary and ambient sound tracks that were played back over loudspeakers around a dummy head. Detailed specifications of the four experiments are given in Table III. Some of the test tracks used in these experiments and MATLAB codes can be found in [55].

In the first experiment, a speech clip was selected as the primary component, which is amplitude panned by k = 3 and time-shifted by $\tau_o = 40$ samples at a sampling rate of 44.1 kHz, both correspond to the direction of channel 1. The ambient component was taken from a stereo recording of lapping wave with

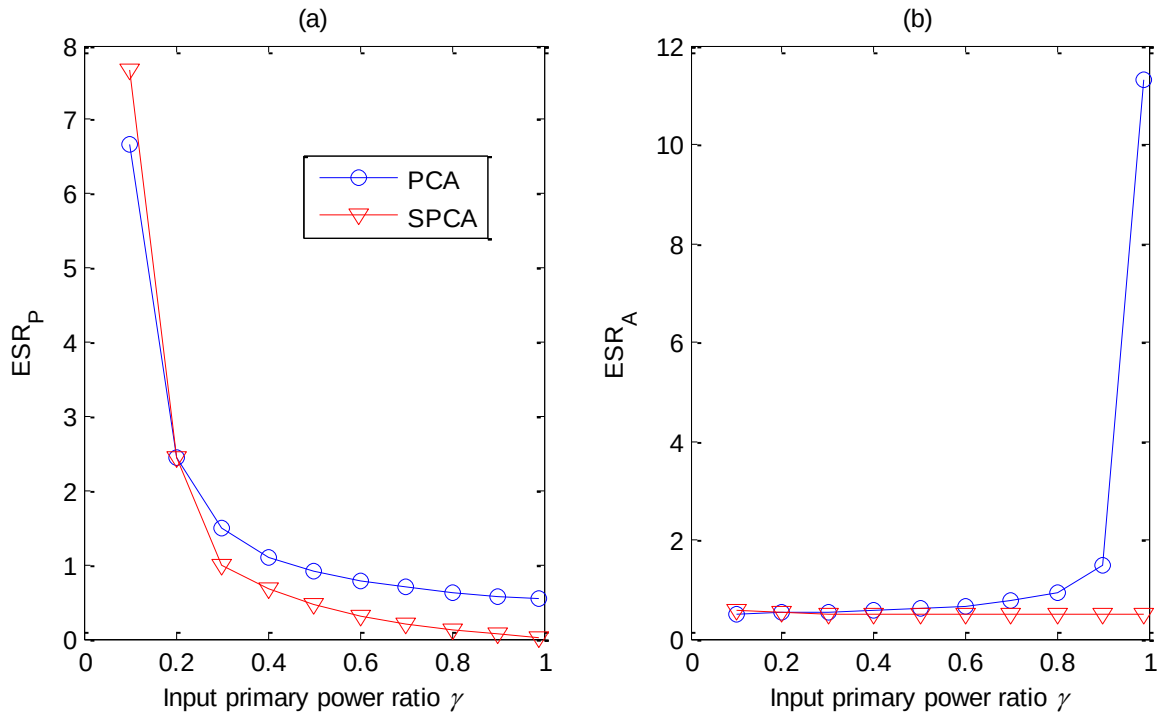


Fig. 7. ESR of (a) primary extraction and (b) ambient extraction using PCA and SPCA in the primary-complex case. Legend in (a) applies to both plots.

low correlation (less than 0.1) and close to unity power ratio between the two channels. Subsequently, the primary and ambient components were linearly mixed based on the values of γ ranging from 0 to 1. Finally, the extraction performance of PCA and SPCA was evaluated using the performance measures introduced in Section IV. Note that the correlation coefficient of the tested primary component at zero lag is 0.17, which is increased to one after time-shifting the synthesized signal by 40 samples according to the estimated ICTD. The unity correlation implies that the primary component is completely correlated in SPCA.

The results of the performance measures of PCA and SPCA are shown in Figs. 6-8. In Fig. 6, there are significant errors in the estimations of k and γ in PCA, which are estimated more accurately in SPCA. Fig. 7 summarizes the ESR of PAE using PCA and SPCA. For primary extraction as shown in Fig. 7(a), significant reduction (more than 50%) of ESR is obtained using SPCA when $\gamma \geq 0.5$. Based on Fig. 7(b), SPCA extracts the ambient components with smaller ESR than PCA, especially when γ is high (more than

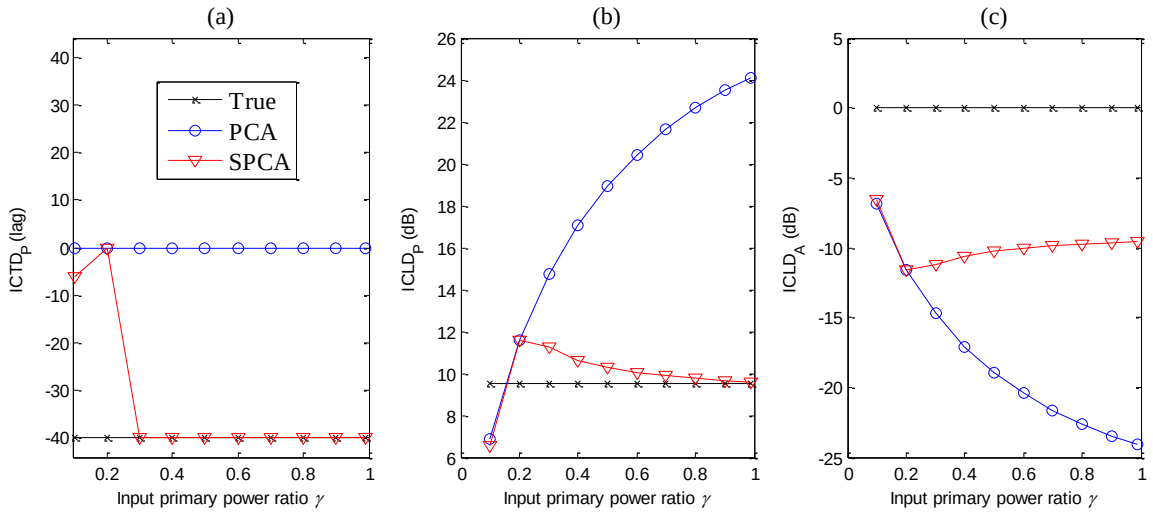


Fig. 8. Comparison of spatial accuracy in PAE using PCA and SPCA in the primary-complex case. (a) ICTD in the extracted primary components; (b) ICLD in the extracted primary components; (c) ICLD in the extracted ambient components. Legend in (a) applies to all plots.

50% reduction for $\gamma \geq 0.8$). The significant improvement lies in the reduction of the leakage from the primary components in the extracted ambient component.

SPCA also outperforms PCA in terms of spatial accuracy of the extracted primary and ambient component. As shown in Fig. 8(a), the ICTD of the primary component extracted by SPCA is closer to the ICTD of the true primary component for $\gamma \geq 0.3$. When the primary components become too weak in the stereo signals, the estimation of ICTD in SPCA is less accurate. For the ICLD whose just-noticeable difference (JND) is generally below 3 dB [56], we found that the ICLD of the primary component extracted by SPCA is significantly closer to the ICLD of the true primary component, as shown in Fig. 8(b). Therefore, the directions of the primary components extracted by SPCA would be more accurately reproduced and localized. For ambient extraction, we observed that the ICLD of the extracted ambient component for SPCA is closer to 0 dB as compared to PCA, as shown in Fig. 8(c). Even though neither approach can extract an uncorrelated and balanced ambient component, a relatively better ambient extraction is obtained with SPCA. Similar to the ideal case, this drawback of ambient extraction is an inherent limitation of PCA [35]. Post-processing techniques like decorrelation [57] and post-scaling [22], [31] can be applied to further enhance ambient extraction. To sum up the first experiment, we can verify

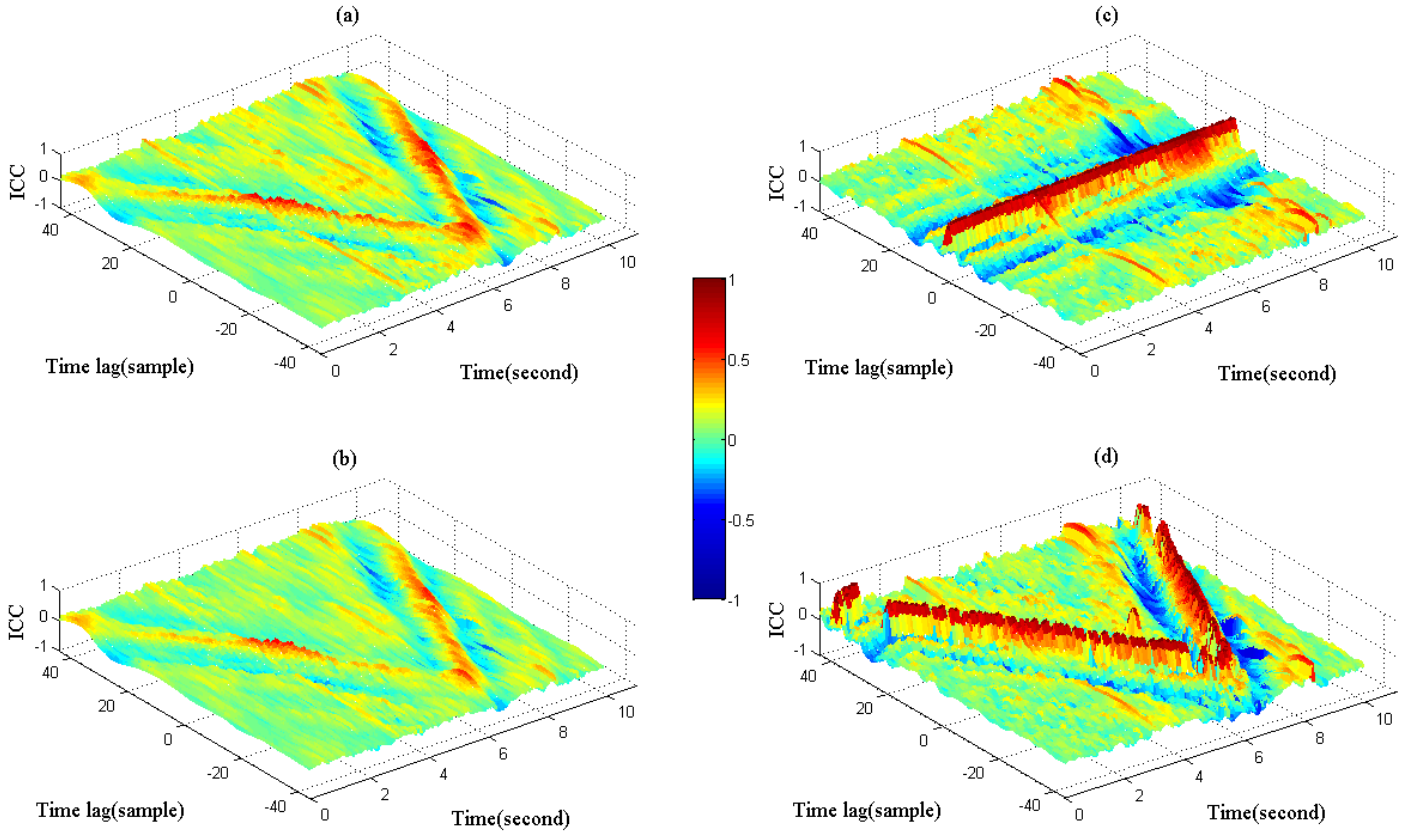
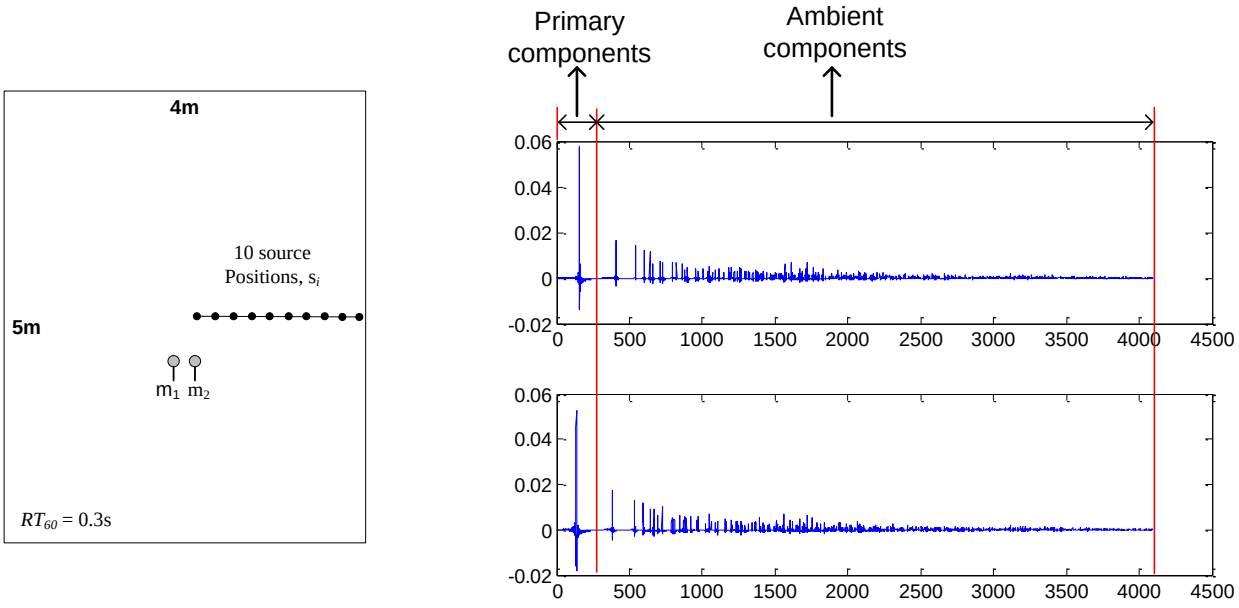


Fig. 9. Short-time cross-correlation function of (a) true primary component; (b) stereo signal with mixed primary and ambient components; (c) primary component extracted using PCA; (d) primary component extracted using SPCA. Frame size is 4096 samples with 50% overlap.

that when dealing with PAE having a directional primary component with time and level differences, SPCA extracts the primary and ambient components more accurately than PCA.

In the second experiment, a binaural recording of a matchbox sound shaking around the dummy head in the anti-clockwise direction was taken as the primary component, and a wave lapping sound was used as the ambient component. The four plots in Fig. 9 illustrate the short-time cross-correlation of the true primary component, mixed signal, primary component extracted by PCA, and primary component extracted by SPCA. The positions of the peaks on the mesh of these plots represent the direction of the primary components, where the time lag at 40 represents extreme left and –40 represents extreme right. The anti-clockwise panning of the primary component around the head, as shown in Fig. 9(a), becomes less obvious after mixing with the ambient component, as shown in Fig. 9(b). Comparing the correlation of the primary component extracted using PCA and SPCA, as shown in Fig. 9(c) and 9(d), respectively, we can easily



(a) (b)

Fig. 10. (a) Specifications of the room, microphone positions and source positions in the reverberation experiment. (b) An example of the generated RIR and the division of the response for primary and ambient components.

verify that only SPCA based PAE preserves the spatial cues of the primary component from the mixed stereo signal. This experiment confirms that SPCA can correctly track the moving directions of the primary components and thus leads to an improved extraction performance with more accurate spatial cues, as compared to PCA.

In the third experiment, we considered the extraction of a direct signal and its reverberation from a stereo recording in a reverberant room. For the purpose of a more accurate evaluation, simulated room impulse responses (RIRs) were used. The RIR was generated using the software from [58], which is created using the image method [59]. As specified in Fig. 10(a), the size of the room is 5×4×6 m^3 with reverberation time RT_{60} set as 0.3s. For the RIR generation, positions for two microphones were set as m_1(2, 1.9, 2) and m_2(2, 2.1, 2). The positions of a speech source varied in 10 locations (one at a time) in a straight line, as (2.5, s_i, 2) with s_i = 1.9+0.2**i*, *i* = 1, 2, …, 10. The length of the RIR is 4096 samples with sampling frequency at 44.1 kHz. In either channel, the mixed signal was obtained by convolving the source with the generated RIR. The true primary components were synthesized by convolving only the direct paths with the

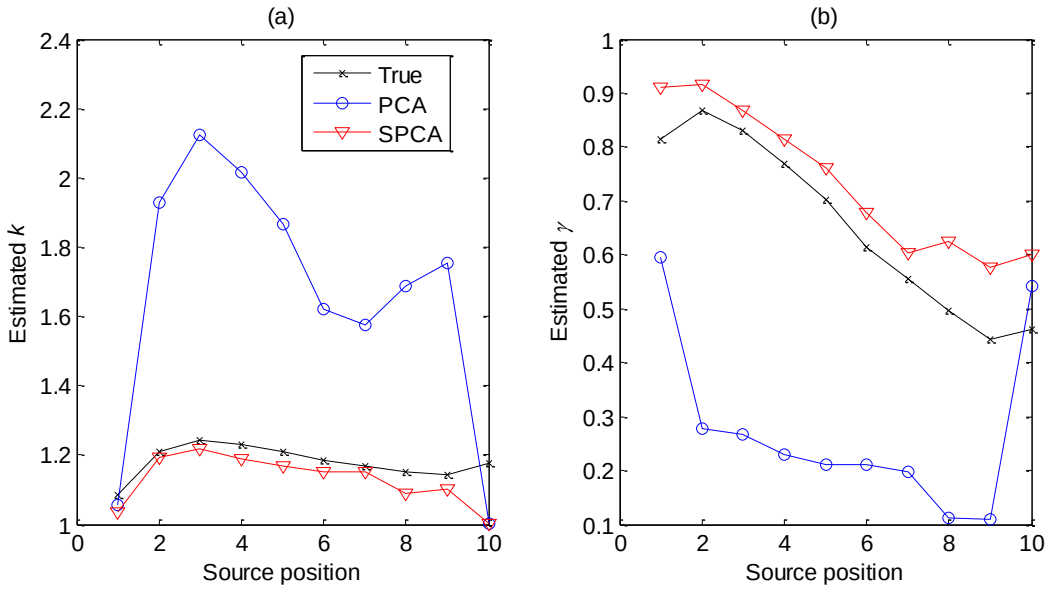


Fig. 11. Comparison of the estimation of (a) k and (b) γ between PCA and SPCA based PAE in the reverberation experiment. Legend in (a) applies to both plots.

source, while the remainder paths are used as the responses for the synthesis of the true ambient components, as shown in Fig. 10(b). Performance of PAE using PCA and SPCA is compared in Figs. 11-13. It can be observed clearly in these figures that as compared to PCA based PAE, SPCA based PAE can estimate k and γ much closer to their true values, thereby yielding a smaller ESR in both primary and ambient extraction, as well as having spatial cues (i.e., ICTD, ICLD) closer to the true values. In particular, we have also applied the normalized least-mean-square (NLMS) approach proposed by Usher [38] in the ambient extraction. As shown in Fig. 12 (b), the proposed SPCA approach also outperforms NLMS significantly.

In the fourth experiment, we tested and compared these PAE approaches using recorded signals. The measurements were conducted in a semi-anechoic recording room (5.4×3.18×2.36 m^3, RT_{60} = 0.2s) at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. The layout of the experiment setup is illustrated in Fig. 14. Four loudspeakers A_1 to A_4 were used to reproduce the ambient sound of a canteen. The primary component, a speech signal, was played back over loudspeaker P, which was placed at each of the three positions with 0°, 45°, and 90° azimuth in the horizontal plane. At the center of the room, a dummy head, which was fitted with a pair of microphones mounted on the two ears,

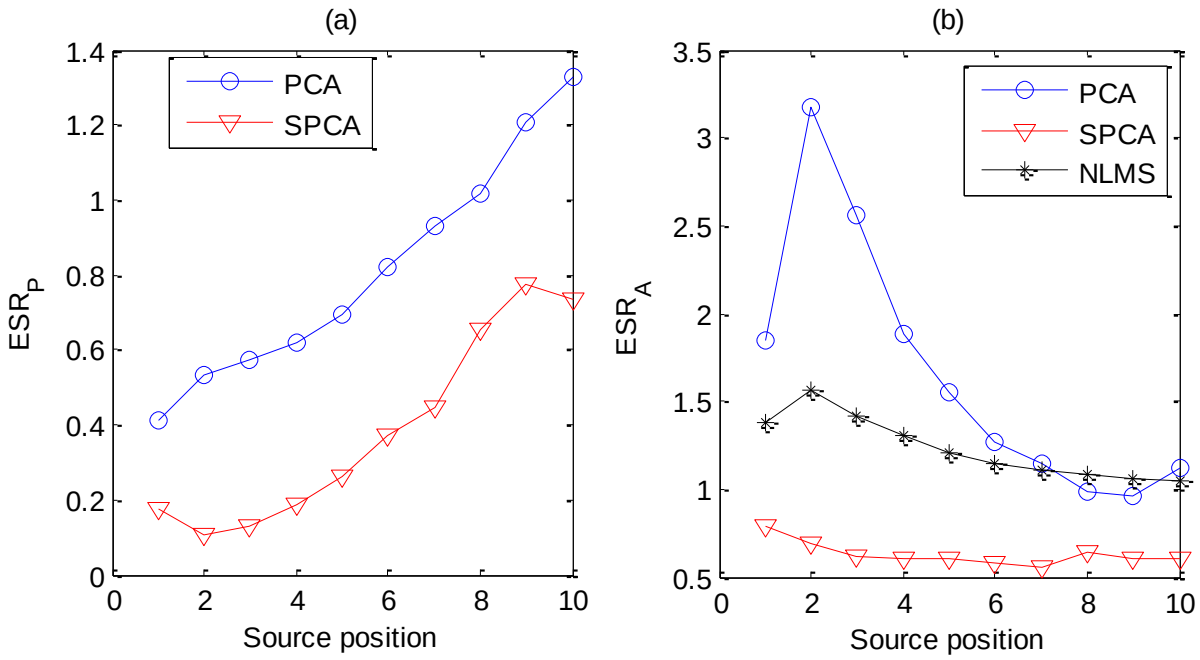


Fig. 12. ESR of (a) primary extraction and (b) ambient extraction using PCA and SPCA in the reverberation experiment. The NLMS approach [35] is included in (b) for comparison of ambient extraction performance.

was used to record the simulated sound scene. To evaluate the performance of the PAE approaches, the "ground truth" reference signals of this experiment (i.e., the true primary and ambient components) were recorded by muting either the one-channel primary loudspeaker or the four-channel ambient loudspeakers. The performance of PCA and SPCA based PAE are summarized in Tables IV and V. In Table IV, the performance of the two PAE approaches is examined by comparing γ, k, and the spatial cues with their true values, respectively. We observed that SPCA based PAE yields much closer results to the true values as compared to PCA based PAE for all directions of the primary component. From Table V, we observed that the values of ESR in SPCA based PAE are lower (up to 50%) than those in PCA based PAE. These observations from the fourth experiment indicate clearly that SPCA based PAE outperforms PCA based PAE in more practical situations.

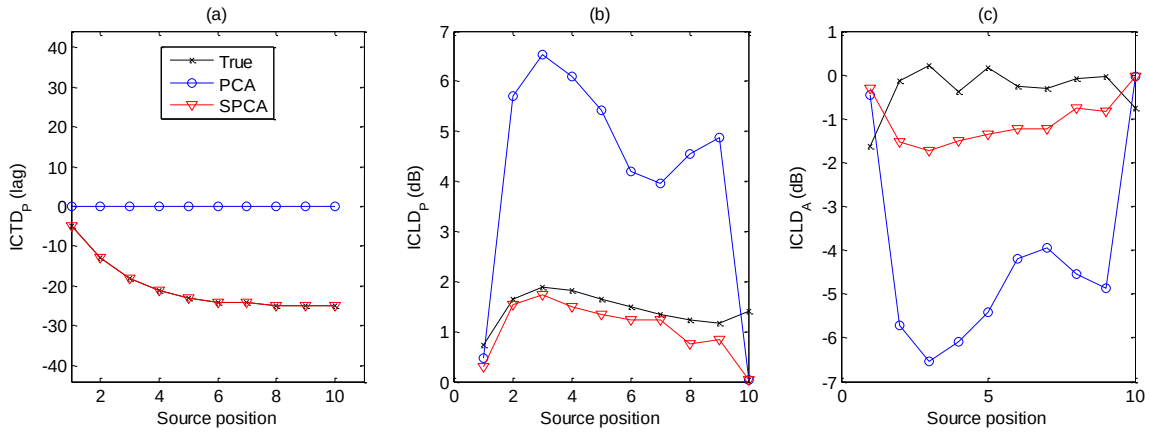


Fig. 13. Comparison of spatial accuracy in PAE using PCA and SPCA in the reverberation experiment. (a) ICTD in the extracted primary components; (b) ICLD in the extracted primary components; (c) ICLD in the extracted ambient components. Legend in (a) applies to all plots.

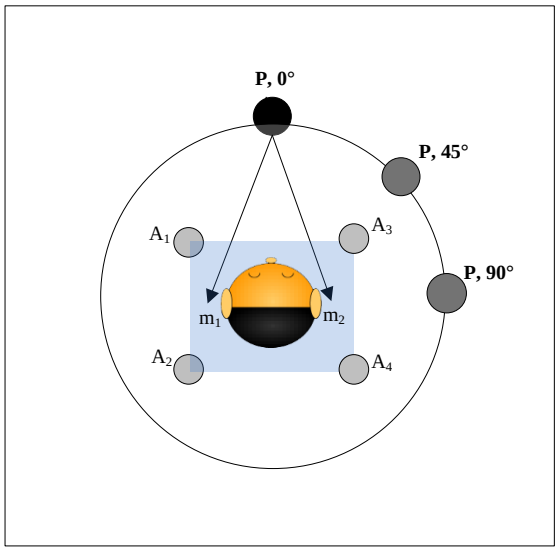


Fig. 14. Layout of the fourth experiment setup. Four ambient loudspeakers are located at A_1-A_4. The primary loudspeaker P is positioned at one of the three directions 0°, 45°, 90° in the horizontal plane with a radius of 1.5 meter. Two microphones m_1 and m_2 are mounted onto the two ears of the dummy head.

VII. CONCLUSIONS

In this paper, we investigated the performance of PCA based PAE in the ideal and primary-complex cases. The performance of PAE was evaluated on extraction accuracy and spatial accuracy. Relatively accurate extraction of primary and ambient components using PCA was found in the ideal case. In practice, the conventional PCA based PAE exhibits severe performance degradation when dealing with the input signals under the primary-complex case, where the primary component is partially correlated at zero lag.

Table IV
Comparison of γ, k, and spatial cues between PCA and SPCA based PAE in the fourth experiment.

	γ			k			$ICTD_P$			$ICLD_P$ (dB)			$ICLD_A$(dB)		
θ	0°	45°	90°	0°	45°	90°	0°	45°	90°	0°	45°	90°	0°	45°	90°
True	0.81	0.79	0.86	0.95	1.47	1.81	1	-17	-31	-1.02	7.74	11.90	1.03	1.18	1.03
PCA	0.66	0.31	0.57	0.93	6.06	3.18	0	0	0	-1.46	36.03	23.11	1.46	-36.03	-23.11
SPCA	**0.76**	**0.73**	**0.72**	**0.94**	**1.54**	**2.18**	**1**	**-17**	**-31**	**-1.26**	**8.65**	**15.60**	**1.26**	**-8.65**	**-15.60**

Table V
Comparison of ESR between PCA and SPCA based PAE in the fourth experiment.

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

Without the knowledge of the correlation of the primary component, the two important parameters primary panning factor and primary power ratio of the stereo signal cannot be estimated accurately. Furthermore, it was found that as the primary correlation decreases, the error in the primary and ambient components extracted by PCA based PAE generally increases. Based on this finding, the proposed SPCA based PAE approach maximizes the primary correlation by appropriately time-shifting the input signals prior to the extraction process. Overlapped output mapping method with a minimum duration of 2 ms overlapping is required to avoid the switching artifacts introduced by time-shifting. As compared to the conventional PCA based PAE, the proposed approach retains the ICTD and corrects the ICLD of the extracted primary component, as well as reduces the extraction error by as much as 50%. With the improved performance of the proposed approach validated using synthesized signals and real-world recordings in our experiments, we conclude that the proposed time-shifting technique can be employed in PAE to handle more generic cases of stereo signals that contains partially correlated primary components. Future work shall investigate the use of subband decomposition in PAE in non-ideal cases.

ACKNOWNLEDGEMENTS

The authors wish to thank the reviewers for their constructive comments and suggestions.

REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

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

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

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

[9] 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.

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

[11] 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.

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

[14] 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.

[15] 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.

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

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

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

[19] E. L. Tan, and W. S. Gan, "Reproduction of immersive sound using directional and conventional loudspeakers," *J. Acoust. Soc. Amer.*, vol. 131, no. 4, pp. 3215-3215, Apr. 2012.

[20] E. L. Tan, W. S. Gan, and C. H. Chen, "Spatial sound reproduction using conventional and parametric loudspeakers,"in *Proc. APSIPA ASC*, Hollywood, CA, Dec. 2012.

[21] 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.

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

[23] 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.

[24] 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.

[25] J. He, W. S. Gan, and E. L. Tan, "Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction," to appear, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. x, pp. xxx-xxx, XXX. 2015.

[26] 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.

[27] 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, France, 2006.

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

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

[30] 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.

[31] 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.

[32] 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.

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

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

[35] 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.

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

[37] 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.

[38] J. Usher and J. Benesty, “Enhancement of spatial sound quality: A new reverberation-extraction audio mixer,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2141-2150, Sep. 2007.

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

[40] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, CA: NASA, 2000.

[41] 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.

[42] 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.

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

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

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

[46] W. A. Yost, “Perceptual models for auditory localization,” in *Proc. 12th Audio Eng. Soc. Int. Conf.*, Copenhagen, Denmark, 1993.

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

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

[49] R. M. Stern, D. Wang, and G. J. Brown, *Computational auditory scene analysis*. Piscataway, NJ: Wiley/IEEE Press, 2006.

[50] WIKIPEDIA. (2013, August 23). Stereophonic sound [Online]. Available: http://en.wikipedia.org/wiki/Stereophonic_sound.

[51] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: a review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp.1920-1938, Sep. 2013.

[52] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[53] 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.

[54] 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.

[55] J. He. (2015 Mar. 16). SPCA [Online]. Available: http://jhe007.wix.com/main#!research/c24xx.

[56] T. Francart and J. Wouters, "Perception of across-frequency interaural level differences," *J. Acoust. Soc. Amer.*, vol. 122, no. 5, pp. 2826-2831, Nov. 2007.

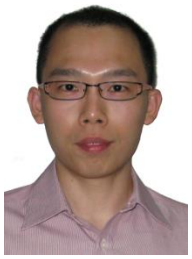
[57] C. Faller, "Parametric multichannel audio coding: synthesis of coherence cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no.1, pp. 299-310, Jan. 2006.

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

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

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.