Primary-Ambient Extraction Using Ambient Phase Estimation with a Sparsity Constraint

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

***Abstract*—Spatial audio reproduction addresses the growing commercial need to recreate an immersive listening experience of digital media content, such as movies and games. Primary-ambient extraction (PAE) is one of the key approaches to facilitate flexible and optimal rendering in spatial audio reproduction. Existing approaches, such as principal component analysis and time-frequency masking, often suffer from severe extraction error. This problem is more evident when the sound scene contains a relatively strong ambient component, which is frequently encountered in digital media. In this Letter, we propose a novel PAE approach by estimating the ambient phase with a sparsity constraint (APES). This approach exploits the equal magnitude of the uncorrelated ambient components in the two channels of a stereo signal and reformulates the PAE problem as an ambient phase estimation problem, which is then solved using the criterion that the primary component is sparse. Our experimental results demonstrate that the proposed approach significantly outperforms existing approaches, especially when the ambient component is relatively strong.**

***Index Terms*—Ambient phase, primary-ambient extraction (PAE), sparsity, spatial audio.**

I. Introduction

SPATIAL audio reproduction of digital media content (e.g., movies, games, etc.) has gained popularity in recent years. Reproduction of sound scenes essentially involves the reproduction of point-like directional sound sources and the diffuse sound environment, which are often referred to as primary and ambient components, respectively [1], [2]. 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 [2], [3]. However, existing mainstream channel-based audio formats (such as stereo and multichannel signals) provide only the mixed signals [4], which necessitate the extraction of the primary and ambient components from the mixed signals. This extraction process is usually known as primary-ambient extraction (PAE). To date, PAE has been applied in spatial audio processing [3], [5]–[9], spatial audio coding [8], [10], [11], audio re-mixing [1], [9], [12], [13], and hybrid loudspeaker systems [14]–[16] as well as natural sound rendering headphone systems [17].

Numerous PAE approaches are applied to stereo and multichannel signals. For the basic signal model for stereo signals, 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 [2]. Based on this model, several time-frequency masking approaches were introduced, where the time-frequency masks are obtained as a nonlinear function of the inter-channel coherence of the input signal [1] or derived based on the criterion of equal level of ambient components between the two channels [18], [19]. Further investigation of the differences between two channels of the stereo signals has led to several types of linear estimation based approaches [20], including principal component analysis (PCA) based approaches [2], [16], [19], [21]–[27] and least-squares based approaches [20], [28]. These linear estimation based approaches extract the primary and ambient components using different performance-related criteria [20]. To deal with digital media signals that cannot fit into the basic signal model, there are other PAE approaches that consider signal model classification [29], time/phase differences in primary components [27], [30], [31], non-negative matrix factorization [32], independent component analysis [33], etc.

The above-mentioned PAE approaches often suffer from severe extraction error that takes the form of residual uncorrelated ambient component in the extracted primary and ambient components, especially for digital media content having relatively strong ambient power [20]. In this Letter, we aim to improve the performance of PAE by exploiting the characteristics of uncorrelated ambient components of digital media content and the sparsity of the primary components [34]. These considerations have led to the novel approach to solve the PAE problem using ambient phase estimation with a sparsity constraint (APES).

The rest of this Letter is structured as follows. The stereo signal model is reviewed in Section II. Section III discusses the proposed APES approach, followed by the experimental results in Section IV. Finally, our conclusions are drawn in Section V.

II. Stereo Signal Model

In spatial audio, PAE is often considered in time-frequency domain [1], [2], [8], [10], [19], [28], [35]. It is generally assumed that within a time frame (consisting of N short frames), each subband of the input signal contains only one dominant source, which is considered as the primary component, and PAE

Manuscript received September 14, 2014; revised November 04, 2014; accepted December 23, 2014. Date of publication December 31, 2014; date of current version January 15, 2015. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under Grant MOE2010-T2-2-040. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Emauel A. P. Habets.

The authors are with Digital Signal Processing Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jhe007@e.ntu.edu.sg; ewsgan@ntu.edu.sg; etanel@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2387021

is independently carried out on each subband of each frame of the input signal [1], [2], [19], [28]. Denoting the stereo signal in time-frequency domain at time index n and frequency bin index l as $X_c(n,l)$, where the (stereo) channel index $c \in \{0,1\}$. The stereo signal at subband b is denoted as $\mathbf{X}_c[n,b] = [X_c(n,l_{b-1}+1), X_c(n,l_{b-1}+2), \ldots, X_c(n,l_b)]^T$, where l_b is the upper boundary of bin index at subband b [34]. 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 cth channel of the stereo signal, respectively. Since the subband 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 components in stereo signals are correlated, while the ambient components in the two channels are uncorrelated. Correlated primary component could involve inter-channel time and amplitude differences [36]. For this Letter, we shall only consider the primary component to be amplitude panned, i.e., $\mathbf{P}_1 = k\mathbf{P}_0$, where $k \geq 1$ is referred to as the primary panning factor [2], [19], [28]. Amplitude-panned primary components are commonly found in stereo recordings using coincident microphone techniques as well as sound mixes using pan-pot stereo techniques [4]. Considering a channel-based signal, where only the mixed signal is given as input, it is necessary to estimate k. In [20], k is estimated as $k = \frac{r_{11}-r_{00}}{2r_{01}} + \sqrt{\left(\frac{r_{11}-r_{00}}{2r_{01}}\right)^2 + 1}$, where r_{00}, r_{11}, and r_{01} are the autocorrelations and cross-correlation of the input signal in the two channels. Other approaches such as amplitude histograms [37] can also be used to estimate k.

For an ambient component that comprises environmental sound, it is usually considered to be uncorrelated with the primary component [1], [2], [20], [30], [38], [39], as well as having equal power between the two channels. To quantify the power difference between the primary and ambient components, we define the primary power ratio γ as the ratio of total primary power to total signal power in two channels, and $\gamma \in [0,1]$. Previous study revealed that the performance of PAE is highly dependent on γ, where lower γ generally indicates inferior performance [20]. Using the method described in [20], we computed the (estimated) γ for many movie and gaming tracks (e.g., Avatar, Brave, Battlefield 3, BioShock Infinite, etc.), and found that the percentages for the time frames having relative strong ambient power (i.e., $\gamma \leq 0.75$) are often over 50% in these digital media content. These occurrences of strong ambient power case degrade the overall performance of PAE, and therefore a PAE approach that is able to perform well even in the presence of strong ambient power is desired.

III. Ambient Phase Estimation with A Sparsity Constraint

The diffuseness of ambient components usually leads to low correlation between the two channels. To produce diffuse ambient components from raw recordings, decorrelation techniques are commonly used, which mainly include artificial diffuse reverberation [40], [41] that are widely used in studio, as well as other decorrelation techniques, such as introducing delay [42], all-pass filtering [43]–[45], and binaural reverberation [46]. These decorrelation techniques typically produce equal magnitude of ambient components in the two channels of the stereo signal. As such, we can express the spectrum of ambient components as

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

where $\odot$ denotes element-wise Hadamard product, $|\mathbf{A}_c| = |\mathbf{A}|$ represents the equal magnitude of the ambient components, and the element in the bin (n,l) of $\mathbf{W}_c$ is expressed as $W_c(n,l) = e^{j\theta_c(n,l)}$, where $\theta_c(n,l)$ is the bin (n,l) of $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_c = \angle\mathbf{A}_c$ is the phase (in radians) of the ambient components. Considering the panning of the primary component $\mathbf{P}_1 = k\mathbf{P}_0$, the primary component in (1) can be eliminated and (1) can be reduced to

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

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

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

where $./$ represents the element-wise division. Because $|\mathbf{A}|$ is real and non-negative, we derive the relation between the phases of the two ambient components as

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

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

$$\begin{aligned} \mathbf{A}_c &= (\mathbf{X}_1 - k\mathbf{X}_0)\,./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c, \\ \mathbf{P}_c &= \mathbf{X}_c - (\mathbf{X}_1 - k\mathbf{X}_0)\,./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c. \end{aligned} \tag{6}$$

Since $\mathbf{X}_c$ and k can be computed from the input [20], $\mathbf{W}_c$ is the only unknown variable in the right hand sides of (6). It becomes clear that the primary and ambient components are determined by $\mathbf{W}_c$, which is solely related to the phase 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 $\boldsymbol{\theta}_1$ in (5), only $\boldsymbol{\theta}_1$ needs to be estimated. A critical relation in the APE framework is that good extraction performance can be obtained via accurate estimation of ambient phase. Such a relation is a key advantage of APE formulation as similar relations are not found in existing PAE frameworks (e.g., time-frequency masking [1] or linear estimation based PAE [20]).

In general, estimation of ambient phase requires additional criteria that are based on the characteristics of the primary and ambient components. One of the most important characteristics of sound source signals is sparsity, which has been widely used as the critical criterion in finding optimal solutions in many audio and music signal processing applications [34]. In PAE, since the primary components are essentially sound sources, they can be considered to be sparse in the time-frequency domain [34]. Therefore, we estimate $\boldsymbol{\theta}_1$ by restricting the extracted primary component to be sparse, i.e., minimizing the sum of the magnitudes of the primary components for all time-frequency bins:

$$\hat{\boldsymbol{\theta}}_1^* = \arg\min_{\hat{\boldsymbol{\theta}}_1} \left\| \hat{\mathbf{P}}_1 \right\|_1. \tag{7}$$

TABLE I
STEPS IN APES

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

We refer to this approach as the ambient phase estimation with a sparsity constraint.

However, the objective function in (7) is not convex. Therefore, convex optimization techniques are inapplicable, and heuristic methods, such as simulated annealing (SA) [47], are more suitable to solve APES. But SA might not be efficient since optimization is required for all the phase variables. Based on the following two observations, we propose to use a simple but more efficient method to estimate the ambient phase. First, the magnitude of the primary component is independently determined by the phase of the ambient component at the same time-frequency bin and hence, the estimation in (7) can be independently performed for each time-frequency bin. Note that with this approximation, a sufficient condition of the sparsity constraint is applied in practice. Second, the phase variable is bounded to $(-\pi, \pi]$ and high precision of the estimated phase may not be necessary. Thus, we select the optimal phase estimates from an array of discrete phase values $\hat{\boldsymbol{\theta}}_1(d) = (2\pi d/D - \pi)$, where $d \in \{1, 2, \ldots, D\}$ with D being the total number of discrete phase values to be considered. We refer to this method as discrete searching (DS). Following (5) and (6), 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., $\hat{\boldsymbol{\theta}}_1^* = \hat{\boldsymbol{\theta}}_1(d^*)$, where $d^* = \arg\min_{d\in\{1,2,\ldots,D\}} |\hat{P}_1(d)|$. Clearly, the value of D affects the extraction and the computational performance of APES using DS. The detailed steps of APES are listed in Table I.

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

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiments using synthesized mixed signals were carried out to evaluate the proposed approach. One frame (consists of 4096 samples) of speech signal is selected as the primary component, which is amplitude panned to channel 1 with a panning factor $k = 4, 2, 1$. A wave lapping sound recorded at the beach is selected as the ambient component, which is decorrelated using all-pass filters with random phase [45]. The stereo input signal is obtained by mixing the primary and ambient components using different values of primary power ratio ranging from 0 to 1 with an interval of 0.1.

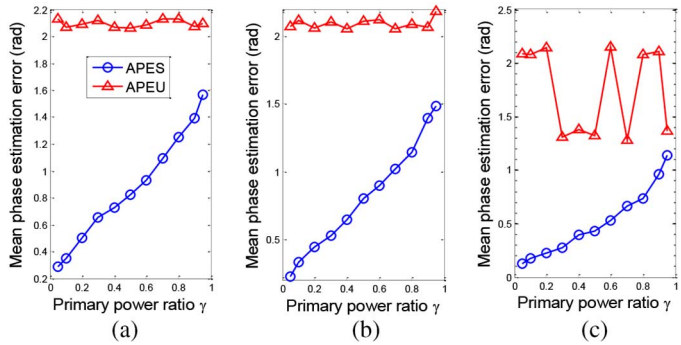

Fig. 1. Comparison of ambient phase estimation error between APES and APEU with (a) $k = 4$; (b) $k = 2$; and (c) $k = 1$. Legend in (a) applies to all the plots.

Our experiments compare the extraction performance of APES, APEU, PCA [2], and two time-frequency masking approaches: Masking 1 [19] and Masking 2 [1]. In the first three experiments, DS with $D = 100$ is used as the searching method of APES. Extraction performance is quantified by the error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components, where lower ESR indicates a better extraction. The ESR for the primary and ambient components are computed as

$$\mathrm{ESR}_{\mathbf{y}} = 10\log_{10}\left\{\sum_{c=0}^{1}\frac{\|\hat{\mathbf{y}}_c - \mathbf{y}_c\|_2^2}{2\|\mathbf{y}_c\|_2^2}\right\}, \forall \mathbf{y} = \mathbf{p}, \quad \text{or} \quad \mathbf{a}. \tag{8}$$

First, we examine the significance of ambient phase estimation by comparing the performance of APES with APEU. In Fig. 1, we show the mean phase estimation error and it is observed that compared to a random phase in APEU, the phase estimation error in APES is much lower. As a consequence, ESRs in APES are significantly lower than those in APEU, as shown in Fig. 2. This result indicates that obviously, close ambient phase estimation is necessary.

Second, we compare the APES with some other PAE approaches in the literature. From Fig. 2, it is clear that APES significantly outperforms other approaches in terms of ESR for $\gamma \leq 0.8$ and $k \neq 1$, suggesting that a better extraction of primary and ambient components is found with APES when primary components is panned and ambient power is strong. When $k = 1$, APES has comparable performance to the masking approaches, and performs slightly better than PCA for $\gamma \leq 0.5$. Referring to Fig. 1 that the ambient phase estimation error is similar for different k values, we can infer that the relatively poorer performance of APES for $k = 1$ is an inherent limitation of APES. Moreover, we compute the mean ESR across all tested

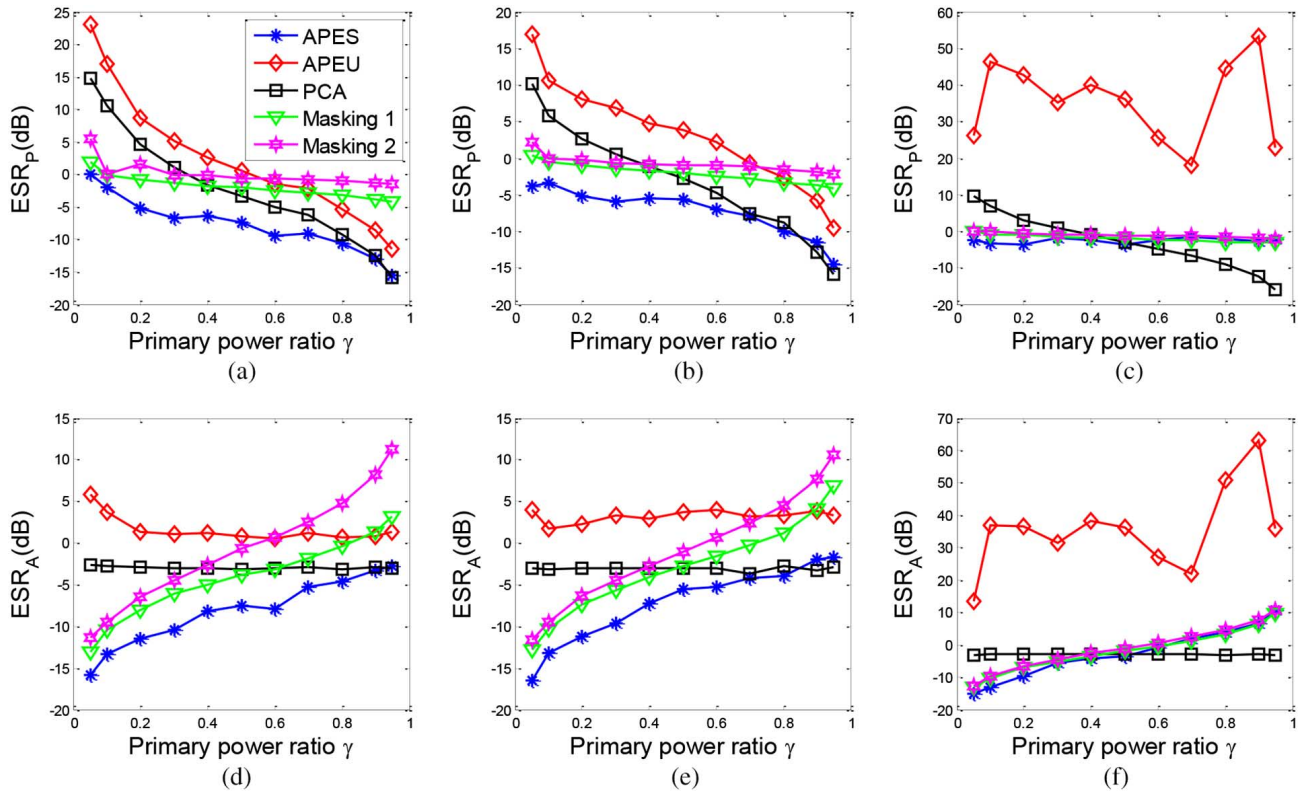

Fig. 2. ESR of (a)–(c) extracted primary component and (d)–(f) extracted ambient component using APES, APEU, PCA [2], Masking 1 [19], and Masking 2 [1]. Three different values of primary panning factor are used: (a), (d) $k = 4$; (b) $k = 2$; (c), (f) $k = 1$. Legend in (a) applies to all the plots.

TABLE II
COMPARISON OF APES WITH DIFFERENT SEARCHING METHODS

Method	Computation time (s)	ESR_P (dB)	ESR_A (dB)
DS (D=10)	0.18	-7.28	-7.23
DS (D=100)	1.62	-7.58	-7.50
SA	426	-7.59	-7.51

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

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

For the purpose of reproducible research, the source code and demo tracks can be found in [48]. However, it shall be noted that the influence of time-frequency transform, though not studied in this paper, is very critical and requires further investigation. Meanwhile, the performance of these PAE approaches shall also be evaluated using more practical signals. Moreover, ambient components in the complex signals are more prone to inter-channel magnitude variations, and therefore probabilistic models based on the statistics of these variations shall be studied to improve the robustness of PAE approaches.

V. CONCLUSIONS

In this Letter, we presented a novel approach to solve the PAE problem using APES. Considering that the diffuse ambient components in two channels of a stereo signal exhibit equal magnitude, the PAE problem is reformulated as an ambient phase estimation problem. Our novel APE formulation provides a promising way to solve PAE as the extraction performance is solely determined by ambient phase estimation accuracy. In this Letter, APE is solved based on the sparsity of the primary components. Based on our experiments using synthesized signals, we found that though under imperfect ambient phase estimation, the proposed approach still showed significant improvement (3-6 dB average reduction in ESR) over existing approaches, especially in the presence of strong ambient components and panned primary components. Moreover, the efficiency of APES can be improved by lowering the precision of the phase estimation, without introducing significant degradation on the extraction performance. Future work includes the study on the influence of time-frequency transform, handling more complex stereo and multichannel signals using probabilistic models, and other optimization criteria in APE.

This paper previously published in IEEE Signal Processing Letters

REFERENCES

[1] 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.

[2] 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.

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

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

[6] J. Breebaart and C. Faller, *Spatial audio processing: MPEG surround and other applications*. Chichester, U.K.: Wiley, 2007.

[7] 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.

[8] 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, NY, USA, 2007.

[9] 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.

[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, "Spatial audio scene coding," in *Proc. 125th Audio Eng. Soc. Conv.*, San Francisco, 2008.

[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 *Proc. 128th Audio Eng. Soc. Conv.*, London, UK, 2010.

[14] 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.

[15] 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.

[16] 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.

[17] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," *IEEE Signal Process. Mag.*, Mar. 2015, in press.

[18] 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, CA, USA, 2012.

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

[20] 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.

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

[22] 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.

[23] 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, CA, 2012.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.

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

[38] 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.

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

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

[41] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1421–1448, Jul. 2012.

[42] 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.

[43] 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.

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

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

[46] 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.

[47] P. J. V. Laarhoven and E. H. Aarts, *Simulated Annealing*. Amsterdam, The Netherlands: Springer, 1987.

[48] J. He, Ambient phase estimation APE [Online]. Available: http://jhe007.wix.com/main#!ambient-phase-estimation/cied Feb. 24, 2014

This paper previously published in IEEE Signal Processing Letters