A STUDY ON THE FREQUENCY-DOMAIN PRIMARY-AMBIENT EXTRACTION FOR STEREO AUDIO SIGNALS

Jianjun He, Woon-Seng Gan, and *Ee-Leng Tan*

Digital Signal Processing Lab, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore
{jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg, etanel@ntu.edu.sg}

ABSTRACT

Primary-ambient extraction (PAE) has been playing an important role in spatial audio analysis-synthesis. Based on the spatial features, PAE decomposes a signal into primary and ambient components, which are then rendered separately. PAE is performed in subband domain for complex input signals having multiple point-like sound sources. However, the performance of PAE approaches and their key influences for such signals have not been well-studied so far. In this paper, we conducted a study on frequency-domain PAE using principal component analysis (PCA) in the case of multiple sources. We found that the partitioning of the frequency bins is very critical in PAE. Simulation results reveal that the proposed top-down adaptive partitioning method achieves superior performance as compared to the conventional partitioning methods.

Index Terms—Primary-ambient extraction (PAE), spatial audio, principal component analysis (PCA), multiple sources, frequency domain

1. INTRODUCTION

In digital media, physical or virtual sound scenes are typically represented in channel-based representation or object-based representation [1]. The channel-based representation, such as stereo or 5.1 surround sound, is most widely used because of its direct relation to the speaker configuration of the playback system. But the channel-based representation lacks of the flexibility to support different speaker configurations. On the other hand, object-based representation can be applied to any loudspeaker configuration by rendering the sound objects based on their spatial attributes [2]. The difficulty with object-based representation is the requirement of significantly larger storage and higher transmission bandwidth [3]. To avoid these problems, a new representation approach inspired by human auditory system is developed, which exploits the representation of the foreground and background sound. These sound components are usually referred to as the primary and ambient components, respectively [4]. The primary components usually consist of multiple point-like sound sources, whereas the ambient components are made up of environmental sound, such as the reverberation, applause, or nature sound like waterfall. Such primary and ambient based representation facilitates flexible rendering of the sound scene based on the loudspeaker configuration without degrading the efficiency in the reproduction.

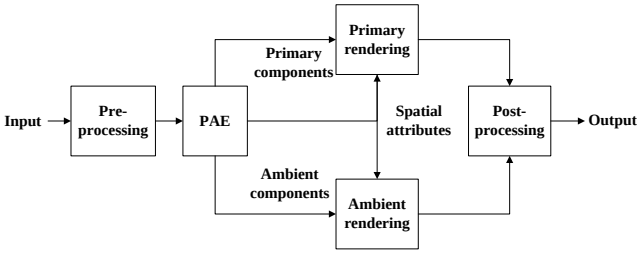

Fig. 1 An illustration of PAE based spatial audio systems.

However, the primary and ambient components are usually mixed in the audio signal for existing channel-based audio formats, which necessitate the extraction of primary and ambient components from the audio signal. Fig. 1 illustrates the primary-ambient extraction (PAE) based spatial audio systems. Prior to PAE, preprocessing such as short-time Fourier transform (STFT) can be applied. The output of PAE will be the extracted primary and ambient components, along with their spatial attributes. These spatial attributes can either be incorporated in the extracted components or transmitted to the receiver for flexible rendering. Post-processing techniques, which may include enhancement [5], [6], coding [4], [7], re-mixing [8]-[11], or simply sending to the playback systems [12]-[14], can be employed in the receiver depending on the requirements of the applications.

To date, many approaches have been proposed for PAE from stereo signals, which include time-frequency masking [8], least squares [15], and principal component analysis (PCA) [16]-[21], [22]. Among these approaches, PCA is the most widely used. These PAE approaches are essentially applied by modelling the stereo signal as a linear mixture of one dominant source and an ambient sound in every subband [22]. However, to the best of our knowledge, the

performance of the subband PAE in dealing with multiple sources has not been investigated.

In this paper, we focus on the study of frequency-domain PAE in the case of multiple sources. PCA based approaches are selected for our testing and only the extraction of primary components is discussed. The rest of this paper is organized as follows. In Section 2, we review the stereo signal model for PAE. Subsequently, two PCA based PAE approaches are introduced. Section 3 discusses in detail the most important step of frequency-domain PAE, i.e., partitioning of the frequency bins. Section 4 presents a series of simulations to validate the PAE approaches. Finally, we conclude this work in Section 5.

2. PAE IN FREQUENCY DOMAIN

In this section, we shall discuss the basic stereo signal model and two PCA based approaches in time domain and frequency domain.

2.1. Stereo Signal Model

Given that primary and ambient components are directional and diffuse, respectively, PAE aims to separate the primary components with the ambient components based on their perceptual spatial features. The perceptual spatial features of these components can be characterized by the inter-channel relationships, which include inter-channel time difference (ICTD), inter-channel level difference (ICLD) and inter-channel cross-correlation coefficient (ICC) [23]. As the number of the sources in the primary components is usually unknown, a common practice in spatial audio processing is to transform the signals into time-frequency domain using short-time Fourier transform (STFT) [8], [15], [16], [24] or subband via filter banks like hybrid quadrature mirror filter banks [25]. Each frequency band or subband is generally assumed to contain one dominant source as the primary component and an ambient component [8], [15], [16], [18].

PAE is performed in each subband of each frame independently and the extracted primary and ambient components are combined via inverse transform or synthesis filter banks. Denoting the *m*th subband of the input stereo signals at time index *l* as $\mathbf{x}_0[m,l]=\left[x_0(0),\ldots,x_0(N-1)\right]^T$, and $\mathbf{x}_1[m,l]=\left[x_1(0),\ldots,x_1(N-1)\right]^T$, where *N* is the frame length. Thus, we can express the stereo signal model as:

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

where $\mathbf{p}_0,\mathbf{p}_1$ and $\mathbf{a}_0,\mathbf{a}_1$ are the primary and ambient components in the two channels, respectively. Since the subband analysis is generally used in PAE, the indices $[m,l]$ are omitted for brevity. This stereo signal model assumes that the primary and ambient components in the two channels are correlated and uncorrelated, respectively. In [15], [16], [18], the correlated primary component is assumed to be amplitude panned, i.e., $\mathbf{p}_1=k\mathbf{p}_0$, where *k* is referred to as the primary panning factor (PPF). As ambient component comprises environmental sound, it is usually considered to be uncorrelated with the primary component [26]. Considering the diffuseness of the ambient component, it is uncorrelated between the two channels and the ambient power is relatively balanced in the two channels of the stereo signal. To quantify the power difference between the primary and ambient components, we introduce the primary power ratio (PPR), which is defined as the ratio of total primary power to total signal power in two channels. Summarizing the assumptions for the stereo signal model, we have

$$\mathbf{p}_1=k\mathbf{p}_0,\ \mathbf{a}_0\perp\mathbf{a}_1,\ \mathbf{p}_i\perp\mathbf{a}_j,\forall i,j\in\{0,1\}, \tag{2}$$

where $\perp$ represents that two signals are uncorrelated.

2.2. PAE using PCA and Shifted PCA

PCA is applied in PAE to decompose the covariance matrix of the input signal [16], [18], [19] into its eigenvectors and eigenvalues. As discussed in [27], the extracted primary components are estimated using

$$\hat{\mathbf{p}}_{\text{PCA},0}=\frac{1}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right),\ \hat{\mathbf{p}}_{\text{PCA},1}=\frac{k}{1+k^2}\left(\mathbf{x}_0+k\mathbf{x}_1\right). \tag{3}$$

The assumption in the stereo signal model confines the primary component to be correlated at zero lag and does not consider the ICTD of the primary component. Studies in [27] have shown that conventional PCA degrades the PAE performance when primary components are partially correlated at zero lag. To solve this problem, shifted PCA (SPCA) is introduced by compensating the ICTD via time shifting [27]. Suppose we find the ICTD to be *d* samples, the output of the each sample in the extracted primary components can be expressed as

$$\begin{aligned}\hat{\mathbf{p}}_{\text{SPCA},0}(\mathrm{j})&=\frac{1}{1+k^2}\left[\mathbf{x}_0(j)+k\mathbf{x}_1(j-d)\right],\\ \hat{\mathbf{p}}_{\text{SPCA},1}(\mathrm{j})&=\frac{k}{1+k^2}\left[\mathbf{x}_0(j+d)+k\mathbf{x}_1(j)\right],\end{aligned} \tag{4}$$

PCA is considered as a special case of SPCA by setting *d*=0.

2.3. PAE using SPCA in Frequency Domain

Next, we consider PAE in the frequency domain by converting the previous time-domain analysis into frequency domain. From (3)-(4), only parameters *k* and *d* are relevant to the extracted primary components in PCA and SPCA, and both parameters are computed using the correlations [27]. Therefore, we shall see how correlations are computed in frequency domain. As discussed in [28], the correlation between two signals $\mathbf{x}_i$ and $\mathbf{x}_j$ can be computed by

$$r_{ij}(\tau)=\begin{cases} IDFT\left(\mathbf{X}_i^*(f)\mathbf{X}_j(f)\right), \tau \ge 0 \\ IDFT\left(\mathbf{X}_i(f)\mathbf{X}_j^*(f)\right), \tau < 0 \end{cases} \quad (5)$$

where $\mathbf{X}_i(f)$ is the DFT of $\mathbf{x}_i$ and * denotes complex conjugate. The ICTD is determined based on the maximum of the cross-correlation

$$d = \arg\max_{\tau}\left\{r_{01}(\tau)\right\}. \quad (6)$$

Time-shifting in time domain is equivalent to phase-shifting in frequency domain [29], that is,

$$\mathbf{x}_i\left[(n-d)_N\right] \overset{DFT}{\leftrightarrow} \mathbf{X}_i(f)e^{-j2\pi fd/N}. \quad (7)$$

Thus, we can rewrite (4) in the frequency domain as

$$\begin{aligned} \hat{\mathbf{P}}_{\mathrm{SPCA},0}(f) &= \frac{1}{1+k^2}\left[\mathbf{X}_0(f)+k\mathbf{X}_1(f)e^{-j2\pi fd/N}\right], \\ \hat{\mathbf{P}}_{\mathrm{SPCA},1}(f) &= \frac{k}{1+k^2}\left[\mathbf{X}_0(f)e^{j2\pi fd/N}+k\mathbf{X}_1(f)\right]. \end{aligned} \quad (8)$$

3. FREQUENCY BIN PARTITIONING IN PAE

To effectively handle multiple sources in the primary components, frequency bins of the input signal are grouped into several partitions. In each partition, there is only one dominant source and hence one corresponding value of *k* and *d* is computed. Ideally, the number of partitions should be the same as the number of sources, and the frequency bins should be grouped in a way such that the magnitude of one source in each partition is significantly higher than the magnitude of other sources. However, the number and spectra of the sources in any given input signals are usually unknown. Hence, the ideal partitioning is difficult or impossible to achieve.

Alternatively, we consider two types of feasible partitioning methods, namely, fixed partitioning and adaptive partitioning. Regardless of the input signal, the fixed partitioning classifies the frequency bins into a certain number of partitions uniformly [8], [15] or non-uniformly, such as equivalent rectangular bandwidth (ERB) [24]. By contrast, adaptive partitioning takes into account of the input signal via the top-down (TD) or bottom-up (BU) method. BU method starts with every bin as one partition and then gradually reduces the number of partitions by combining the bins. Conversely, TD starts from one partition containing all frequency bins and iteratively divides each partition into two sub-partitions, according to certain conditions. As the number of partitions is usually limited, TD is more efficient than BU, and hence preferred.

To determine whether one partition requires further division, ICC-based criteria are proposed in TD partitioning. First, if the ICC of the current partition is already high enough, we consider only one source is dominant in the current partition and cease further division of the partition. Otherwise, the ICCs of the two divided sub-partitions are examined. The partitioning is continued only when at least one of two ICCs of the sub-partitions becomes higher, and neither ICC of the sub-partitions becomes too small which indicates no source is dominant. Suppose the ICCs of the current partition, and two uniformly divided sub-partitions are ϕ_0, ϕ_1, ϕ_2, as shown in Fig. 2. For generality, a higher threshold of ICC ϕ_H and a lower threshold ϕ_L are introduced. Thus, we propose the following three criteria for the continuation of partitioning in TD:

a) $\phi_0 < \phi_H$, and
b) $\mathrm{Max}(\phi_1, \phi_2) > \phi_0$, and
c) $\mathrm{Min}(\phi_1, \phi_2) > \phi_L$.

The partitioning is stopped when any criterion is unsatisfied.

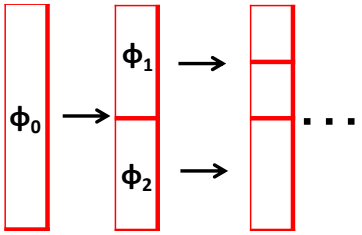

Fig. 2. Top-down partitioning.

4. SIMULATION TESTING AND DISCUSSIONS

To evaluate the performance of frequency-domain PAE approaches, a number of simulations are conducted. In these simulations, speech and music signals are selected as two sources in the primary components, which are amplitude panned and time-shifted separately to simulate different directions. To fulfill the assumptions of the stereo signal model, uncorrelated white Gaussian noise is used as the ambient component. Subsequently, the primary and ambient components are linearly mixed by letting PPR=0.9. DFT of size *N*=4096, and Hanning window with 50% overlapping is applied. Both PCA and SPCA are employed in the testing, and their settings are listed as follows:

a) Full-band, without partitioning (denoted by F);
b-e) Fixed partitioning, with 2, 8, 32 uniform (U) partitions or 20 non-uniform (N) partitions based on ERB [22], (denoted by 2U, 8U, 32U, and 20N, respectively);
f) TD adaptive partitioning, with ϕ_H =0.7, ϕ_L=0.05.

The performance of PAE is determined by the error-to-signal ratio (ESR) [22], which can be computed as

$$\mathrm{ESR(dB)} = 10\log_{10}\left[0.5\left(\frac{\left\|\hat{\mathbf{p}}_0-\mathbf{p}_0\right\|_2^2}{\left\|\mathbf{p}_0\right\|_2^2}+\frac{\left\|\hat{\mathbf{p}}_1-\mathbf{p}_1\right\|_2^2}{\left\|\mathbf{p}_1\right\|_2^2}\right)\right]. \quad (9)$$

A better performance is achieved when ESR is smaller.

First, we test these PAE approaches with a signal containing one source (a speech) in the primary components and the ESR results are presented in Table I. SPCA is better than PCA since it takes the time difference of the primary component into consideration. Comparing the results of SPCA in fixed partitioning with that in the full-band, we

Table I ESR of PAE for one source

	F	2U	8U	32U	20N	TD
PCA	-3.69	-3.38	-3.34	-3.16	-3.33	-3.72
SPCA	-14.78	-12.34	-11.52	-10.63	-10.34	-14.13

Table II ESR of PAE for two sources

Approach	Setting	DS	C	SS
PCA	F	-4.74	-8.06	-4.18
	2U	-5.04	-8.19	-3.95
	8U	-5.22	-8.34	-3.91
	32U	-5.48	-8.44	-3.89
	20N	-6.85	-9.55	-3.98
	TD	-5.03	-8.44	-4.19
SPCA	F	-6.45	-8.07	-10.16
	2U	-6.85	-8.38	-9.89
	8U	-7.11	-8.57	-9.8
	32U	-7.25	-8.44	-9.07
	20N	-7.73	-9.07	-7.29
	TD	-7.93	-8.58	-10.41

observed that the PAE performance degrades as the number of partitions increases. This observation indicates that the partitioning is not required and should be avoided for the single source case. Nevertheless, the performance of TD is quite close to the full-band approach.

Next, we test the performance of PAE when there are two sources in the primary components. Basically, three cases for the directions of two sources are specified as,

a) DS: in different sides, i.e., one in the left, the other in the right;
b) C: one in the center, the other in the left or right;
c) SS: in the same side, i.e., both are in the left or right.

The ESR results are shown in Table II. First, we found that the performance of PCA is worse than that of SPCA, especially when no sources are in the center. Second, not all SPCA approaches with partitioning can yield a better performance than SPCA in full-band, especially when the directions of the two sources are closer (e.g., SS), as shown in Fig. 3. Generally, TD performs better than the fixed partitioning approaches, as well as the full-band approach. As the directions of the two sources get closer (i.e., from DS to SS), better performance with TD is usually achieved.

It is worthwhile to mention that the partitioning of the frequency bins is very critical as there exists overlapping in the spectra of the sources. As the content of the sources varies, the partitioning method should be designed to adapt to these variations. Even though a few more considerations need to be addressed in the future work, TD partitioning is a promising way to divide the frequency bins. Specifically, the selection of the two thresholds, which essentially determine the partitioning, should be investigated further. Also, it is interesting to consider other partitioning methods other than uniform division in TD.

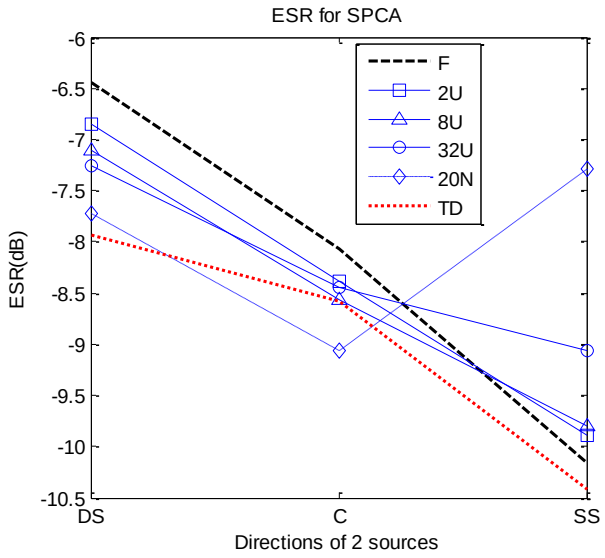

Fig. 3. Comparison of ESR for SPCA with different partitioning settings.

5. CONCLUSIONS

In this paper, we investigated the frequency-domain PAE when there are multiple sources in the primary components of the stereo signals. PCA and SPCA based PAE approaches are employed in this study. We find that frequency bin partitioning is unnecessary for one source, but this partitioning plays an essential role for multiple sources. Conventional fixed partitioning and proposed top-down adaptive partitioning methods were compared for both PCA and SPCA in our simulation. Generally, SPCA outperforms PCA regardless of the partitioning methods. As for the influence of different partitioning methods in SPCA, we found that not all partitioning methods yield better performance than the full-band approach, while the best performance is obtained with the proposed ICC-based TD partitioning method. Future works include the study on the selection of the thresholds and other division methods in TD.

ACKNOWLEDGMENT

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

REFERENCES

[1] 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.

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

[3] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," in *AES 124th Conv.*, Amsterdam, The Netherlands, 2008.

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

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

[6] 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, Nov. 2008.

[7] 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.

[8] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 505-517, Feb. 2014.

[23] F. Baumgarte and C. Faller, "Binaural cue coding—part I: psychoacoustic fundamental and design principals," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

[24] 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.

[25] 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.

[26] 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.

[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] S. Wang, D. Sen, and W. Lu, "Subband Analysis of Time Delay Estimation in STFT domain," in *Proc. 11th Australian International Conference on Speech Science & Technology*, New Zealand, 2006.

[29] S. K. Mitra, *Digital signal processing: a computer-based approach, 3rd ed*. New York: McGraw-Hill, 2006.