TIME-SHIFTED PRINCIPAL COMPONENT ANALYSIS BASED CUE EXTRACTION FOR STEREO AUDIO SIGNALS

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

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

ABSTRACT

In spatial audio analysis-synthesis, one of the key issues is to decompose a signal into cue and ambient components based on their spatial features. Principal component analysis (PCA) has been widely employed in cue extraction. However, the performance of PCA based cue extraction is highly dependent on the assumptions of the input signal model. One of these assumptions is the input signal contains highly correlated cue at zero lag. However, this assumption is often unmet. To overcome this problem, time shifted PCA is proposed in this paper, which involves time-shifting the input signal according to the estimated inter-channel time difference (ITD) of the input signal before cue extraction. From our simulation and listening tests results, the proposed method is found to be superior to the conventional PCA based cue extraction method.

Index Terms— Cue extraction, spatial audio, principal component analysis (PCA), stereo audio signal, correlation

1. INTRODUCTION

With the increasing prevalence of 3D video technology, consumers are demanding a more immersive listening experience to better match 3D visual effects, resulting in a growing need for 3D audio or spatial audio reproduction. In spatial audio rendering, different processing schemes should be independently applied to the cue and ambient components to enhance the spatial perception of audio [1]. However, the cue and ambient components are not stored separately in conventional audio formats including stereo and 5.1, which necessitates ambient-cue extraction. In recent years, ambient-cue extraction has widely used in spatial audio processing [2], [3], audio mixing [4]-[6], spatial audio coding [7], [8], and immersive 3D sound system [9], [10].

To date, many approaches have been proposed for ambient-cue extraction. In [11], a time-frequency mask was created to extract ambience from a stereo input signal. Faller introduced a least-square approach to estimate the cue and ambience for surround sound up-mixing [12]. Other techniques like factor analysis [13] and independent component analysis [14] are also applied in the ambient-cue extraction.

Principal component analysis (PCA) remains one of the most widely studied methods applied in ambient-cue extraction [1], [15]-[17]. A stereo signal is generally modeled as a directional sound source mixed with uncorrelated ambience in these works. Taking consideration of the independence between cue and ambience, the stereo signal is decomposed into two orthogonal components using the Karhunen-Loève transform [18]. Based on the assumption that cue is relatively stronger than ambience, the component with larger variance is assumed as cue and the remaining component as ambience.

The performance of ambient-cue extraction is severely degraded when the cue is not completely correlated at zero lag, leading to significant error in the extracted cue and poor estimation of inter-channel time difference (ITD) and inter-channel level difference (ILD) of the cue. These differences between the extracted cue and the true cue can lead to erroneous sound localization. A normalized least-mean-square approach was proposed in [19] to solve this problem in ambience extraction. A complicated approach discussed in [20], [21] involves classification of the time-frequency regions of the stereo signal into six classes before extraction. Recently, Thompson *et al.* [22] introduced a cue extraction method that directly estimates the magnitude and phase of cue from a multichannel audio signal.

In this paper, we focus on the improvement of cue extraction for stereo signals using PCA based methods. This paper is organized as follows. In Section 2, we review the stereo signal model, and the key assumptions of the signal model for PCA based cue extraction [1]. Subsequently, PCA based cue extraction is derived, and two groups of performance measures [15] are presented. Performance degradation due to the mismatch between the input signal and the assumptions of the signal model is also discussed. Section 3 discusses the avoidance of performance degradation of PCA based cue extraction using shifted PCA based cue extraction. Section 4 presents a series of performance comparisons, including simulation results and subjective listening tests. Finally, we conclude this work in Section 5.

2. PCA BASED CUE EXTRACTION

In the stereo signal model, the two-channel input signal is modeled as a directional cue mixed with uncorrelated ambience. In this section, the closed-form expressions of the PCA based cue extraction are derived. The performance of conventional PCA based cue extraction is then evaluated in both ideal and practical cases.

2.1. Basic signal model

In general, a stereo signal model [1] consists of two parts: (i) a directional component referred as the cue; and (ii) a diffused component referred as the ambience. Denoting the time-domain stereo input signals as $\vec{x}_L, \vec{x}_R$, we formulate the basic signal model as:

$$\vec{x}_L = \vec{c}_L + \vec{a}_L,\ \vec{x}_R = \vec{c}_R + \vec{a}_R, \tag{1}$$

where $\vec{c}_L, \vec{c}_R$ and $\vec{a}_L, \vec{a}_R$ are the cue and ambience in the left and right channels, respectively.

For this model, cue and ambience are assumed to be correlated and uncorrelated, respectively. Correlated cue is considered to satisfy one of the following conditions [22]: (i) amplitude panned, i.e., $\vec{c}_R = k\vec{c}_L$, where k is the cue panning factor (CPF); (ii) time shifted, i.e., $c_R(n) = c_L(n+m)$, where $c_R(n)$ is the nth sample in $\vec{c}_R$ and m is the ITD between two channels; (iii) amplitude panned and time shifted, i.e., $c_R(n) = kc_L(n+m)$. The correlated cue is assumed to be amplitude panned in this basic signal model [1]. Furthermore, the cue is assumed to be uncorrelated with the ambience. Considering the diffuseness of ambience, it is relatively balanced in a stereo signal. Generally, cue is found to possess higher energy than ambience. To determine the energy difference between cue and ambience, we introduce the cue energy ratio (CER) γ, which is defined as the ratio of the total cue energy to the total signal energy. Summarizing these assumptions for the stereo signal model, we have:

$$\vec{c}_R = k\vec{c}_L, \vec{a}_L \perp \vec{a}_R, \vec{c}_L(\vec{c}_R) \perp \vec{a}_L(\vec{a}_R), \tag{2}$$

$$E_{\vec{c}_R} = k^2 E_{\vec{c}_L}, E_{\vec{a}_L} = E_{\vec{a}_R}, \gamma \in (0.5,1), \tag{3}$$

where E denotes the signal energy. Given any stereo input signal that fulfills the above conditions, we can relate the auto-correlations and cross-correlation as:

$$r_{LL} = \vec{x}_L^{\,H}\vec{x}_L = E_{\vec{x}_L} = E_{\vec{c}_L} + E_{\vec{a}_L}, \tag{4}$$

$$r_{RR} = \vec{x}_R^{\,H}\vec{x}_R = E_{\vec{x}_R} = k^2 E_{\vec{c}_L} + E_{\vec{a}_L}, \tag{5}$$

$$r_{LR} = \vec{x}_L^{\,H}\vec{x}_R = \vec{c}_L^{\,H}\vec{c}_R = kE_{\vec{c}_L}, \tag{6}$$

where H is the Hermitian operator. From (4)-(6), the CPF and CER are obtained respectively as:

$$k = \frac{r_{RR} - r_{LL}}{2r_{LR}} + \sqrt{\left(\frac{r_{RR} - r_{LL}}{2r_{LR}}\right)^2 + 1}, \tag{7}$$

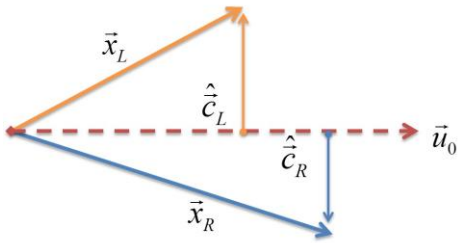

Fig. 1. A geometric representation of PCA-based extraction.

$$\gamma = \frac{2r_{LR} + (r_{RR} - r_{LL})k}{(r_{RR} + r_{LL})k}. \tag{8}$$

Higher values of k and γ indicates that cue is panned more to the right channel, and the cue is more prominent in the input, respectively.

2.2. Cue extraction using PCA decomposition

Based on the signal model, cue extraction using PCA decomposition can be mathematically described as [15]:

$$\vec{u}_0 = \arg\max_{\vec{u}_0} \left(\left\| \vec{u}_0^{\,H}\vec{x}_L \right\|^2 + \left\| \vec{u}_0^{\,H}\vec{x}_R \right\|^2 \right), \tag{9}$$

where $\vec{u}_0$ is the cue basis vector that maximizes the total projection energy of the input signal vectors, as depicted in Fig. 1. A closed-form solution of (9) can be obtained by eigenvalue decomposition of the input covariance matrix. First, we compute the larger eigenvalue and its corresponding cue basis vector using

$$\lambda_0 = 0.5(r_{LL} + r_{RR} + \sqrt{(r_{LL} - r_{RR})^2 + 4r_{LR}^{\,2}}), \tag{10}$$

$$\vec{u}_0 = r_{LR}\vec{x}_L + (\lambda_0 - r_{LL})\vec{x}_R. \tag{11}$$

Next, we compute the extracted cues as

$$\hat{\vec{c}}_L = \frac{\vec{u}_0^{\,H}\vec{x}_L}{\vec{u}_0^{\,H}\vec{u}_0}\vec{u}_0,\ \hat{\vec{c}}_R = \frac{\vec{u}_0^{\,H}\vec{x}_R}{\vec{u}_0^{\,H}\vec{u}_0}\vec{u}_0. \tag{12}$$

Substituting (4)-(8) into (12), the extracted cues are simplified to

$$\hat{\vec{c}}_L = (1+k^2)^{-1}(\vec{x}_L + k\vec{x}_R),\ \hat{\vec{c}}_R = k\hat{\vec{c}}_L. \tag{13}$$

From (13), we observe that the extracted cues are the weighted sum of the stereo input signals, and the cues are scaled by k between the right and left channels. As (13) has only one parameter k, PCA based cue extraction can be efficiently implemented using (13).

2.3. Performance evaluation in ideal and general cases

Generally, it is unlikely for any stereo input signal to satisfy all the assumptions of the signal model reviewed in Section 2.1. In this section, we shall therefore consider the PCA based cue extraction in a more general case, where cue is not only amplitude panned, but partially correlated at zero time lag. In this case, we rewrite (13) using the true cue and ambience:

$$\hat{\vec{c}}_L = \vec{c}_L + \frac{k}{1+k^2}(\vec{n}_R - k\vec{n}_L) + \frac{1}{1+k^2}(\vec{a}_L + k\vec{a}_R),$$
$$\hat{\vec{c}}_R = \vec{c}_R + \frac{1}{1+k^2}(k\vec{n}_L - \vec{n}_R) + \frac{k}{1+k^2}(\vec{a}_L + k\vec{a}_R), \quad (14)$$

where $\vec{n}_L \perp \vec{n}_R$ are the uncorrelated components decomposed from the partially-correlated true cues $\vec{c}_L, \vec{c}_R$ with correlation at zero lag ϕ_c using the signal model discussed in Section 2.1; and k becomes the amplitude difference between the correlated components decomposed from $\vec{c}_L, \vec{c}_R$. Based on (14), the cues are completely extracted but contain two types of errors, which are the ambience leakage (also found in the extracted cues when $\phi_c = 1$) and the error contributed by $\vec{n}_L, \vec{n}_R$ in the extracted cues, for $0 \le \phi_c < 1$.

Based on (14), we introduce two groups of measures to evaluate the performance of cue extraction. The first group measures the extraction accuracy using two metrics. First, we consider error to cue energy ratio (ECR) in the extracted cues, which is defined as the ratio between the total energy of error signals and total energy of the true cues. Second, normalized correlations between the extracted cues and true cues in the left and right channels ϕ_{cL}, ϕ_{cR} are computed to measure their similarity [15]. In the second group, three spatial attributes: inter-channel cross-correlation coefficient (ICC), ITD, and ILD are adopted to evaluate the localization of the extracted cues [23].

Table I summarizes these measures to evaluate the performance of PCA based cue extraction. It is clear from Table I that the accuracy of the cue extraction is dependent on the correlation of the cues ϕ_c, CER, and CPF. To illustrate how the extraction accuracy is influenced by ϕ_c, the results of Table I with CPF = 3 and CER at [0.5, 0.7, 0.9] are plotted in Fig. 2 (a), (b) and (c). Several observations from these plots on the extraction accuracy are as follows. In the ideal case when $\phi_c = 1$, the cues are extracted with relative little error and high similarity to original cues. As ϕ_c decreases, increasing error and decreasing similarity are found in the extracted cues for all values of CER. For the localization of cues extracted using PCA based cue extraction, ICC and ITD are always one and zero, respectively. These values imply that the ITD of the cues is completely lost after PCA extraction. With ILD estimation error (k = 3) plotted in Fig. 2(d), we found that the estimated ILD becomes increasingly unreliable as ϕ_c decreases. Similar observations are also found with other values of CPF and CER. Therefore, it is concluded that the performance of PCA based cue extraction is degraded when cues become partially correlated at zero lag.

3. SHIFTED PCA BASED CUE EXTRACTION

TABLE I: Evaluation results for PCA based cue extraction.

Group 1 Extraction accuracy	**Group 2 Localization parameters**
$\mathrm{ECR} = \frac{1+\beta-\gamma}{2\gamma(1+\beta)}$	ICC = 1, ITD = 0
$\phi_{cL} = \sqrt{\frac{4\gamma}{\left(2+\left(1+k^2\right)\beta\right)(1+\beta+\gamma)}}$	$\mathrm{ILD} = k^2$
$\phi_{cR} = \sqrt{\frac{4k^2\gamma}{\left(2k^2+\left(1+k^2\right)\beta\right)(1+\beta+\gamma)}}$	$\Delta\mathrm{ILD} = \frac{2k^2+k^2\left(1+k^2\right)\beta}{2k^2+\left(1+k^2\right)\beta}$

where $\beta = \sqrt{1+\left(\frac{2k}{1+k^2}\right)^2\left(\frac{1}{\phi_c^2}-1\right)} - 1$, and ΔILD is the ratio between estimated ILD and the true ILD.

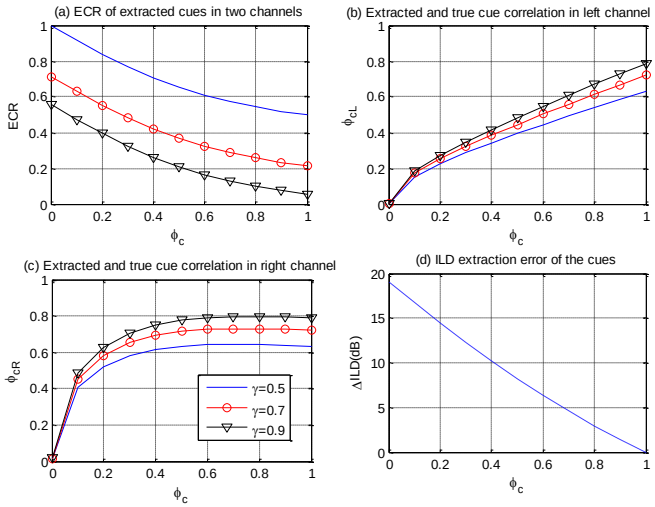

Fig. 2. Performance of PCA based cue extraction in general cases with varying ϕ_c according to the results in Table I (k = 3). (a) ECR of the extracted cues; (b) and (c) normalized correlations between the extracted cues and true cues in the left and right channels; (d) ILD extraction error. Legend in (c) applies to (a), (b) and (c).

In the previous section, we have determined that the performance of PCA based cue extraction is considerably degraded by the low correlation of cue at zero lag. The major cause for lower correlation of cue in most stereo audio signals is that cue is often time shifted and amplitude panned.

The degraded performance includes higher error and loss of ITD in the extracted cue. To overcome these issues, a novel shifted PCA (SPCA) method is proposed to improve the PCA based cue extraction. In shifted PCA, the stereo input signal is first time-shifted according to the estimated ITD of the cue before PCA decomposition. Subsequently, the extracted cue samples are then shifted back using the same ITD. This approach retains the ITD in the extracted cue, and enhances the cue extraction due to the higher correlation of the time shifted cues. The block diagram of the proposed SPCA based cue extraction is shown in Fig. 3.

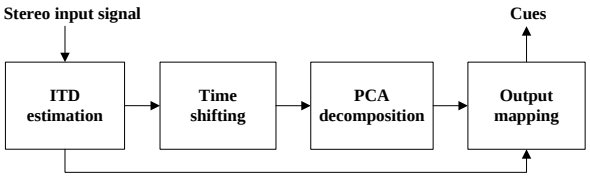

Fig. 3. Block diagram of shifted PCA based cue extraction.

After the coincidence model proposed by Jeffress, there has been extensive research to estimate ITD (see [24]-[29] and references therein). Based on the classical Jeffress's model [24], the ICC of different time lags is first calculated and the lag number corresponds to the maximum ICC would be the estimated ITD of the stereo signal. Note that the conventional PCA is a special case of shifted PCA when ITD is zero. An alternative way to time shifting of cues is to compensate for the phase difference of the cues [30].

In the case of complex input signal, which contains several cues from different directions, the input signal can be decomposed into critical bands and then applying SPCA in each subband, assuming that only one cue is dominant in each critical band. Finally, cues extracted in different critical bands are combined at the output.

4. PERFORMANCE COMPARISON

To evaluate the performance of the proposed SPCA based cue extraction, a number of simulations and subjective listening tests are conducted. In this paper, we shall present one of our test results. Other test results can be found in [31]. In our test, a speech signal is selected as the cue, which is amplitude panned by a factor of 3 and time shifted by 40 time units, both to the right channel; and uncorrelated white Gaussian noise is used as ambience. Subsequently, the cue and ambient components are linearly mixed based on different CERs, which vary from 0.5 to 1 to simulate different scenarios. Next, PCA and SPCA are employed to extract cues from the synthesized stereo signal, respectively. Finally, their performance of cue extraction is determined by computing the two groups of measures discussed in Section 2.3. Note that the normalized correlation of the tested cue at zero lag is 0.1676, which is increased to 1 after shifting the mixed signal according to the estimated ITD. The unity correlation implies that the cues are completely correlated in SPCA.

The simulation results of the performance measures are shown in Fig. 4. Figures 4(a) and 4(b) present the ECR for the left and right channels, respectively. Although the ECR of the right channel is highly similar between PCA and SPCA, we observe a significant drop of ECR in the left channel for SPCA. Similar results are found in the correlations between the extracted and true cues, which are shown in Figs. 4(c) and 4(d). SPCA produces better estimates of the cues in the left channel as compared to PCA. When considering the localization of the extracted

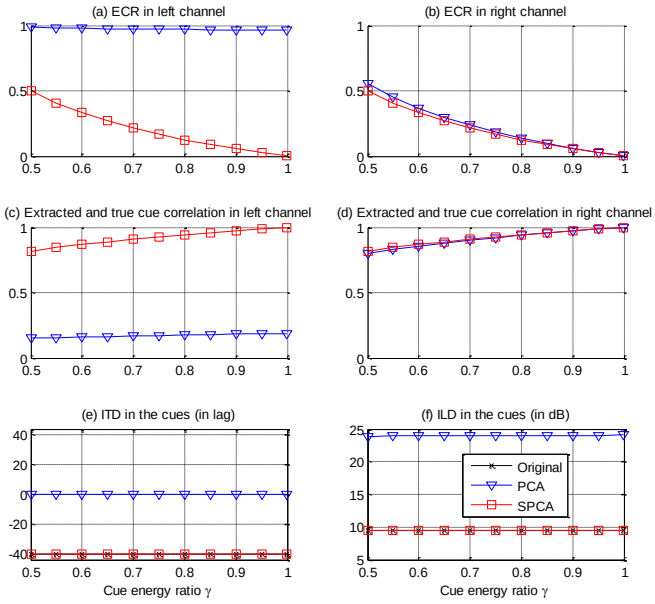

Fig. 4. A performance comparison of cue extraction using PCA and SPCA. The x-axis label and legend in (f) apply to all the plots.

cues, SPCA outperforms PCA as it produces cues having ITD and ILD values closer to the true cues, as shown in the Figs. 4(e) and 4(f). In addition to these objective measures, our informal subjective listening tests also revealed that SPCA performs better than PCA in terms of accuracy in extraction and localization of the cues.

5. CONCLUSION

In this paper, we revisited the problem of cue extraction from stereo audio signals using PCA [16]. Inspired by the discussion of PCA based extraction in [1] and [15], we extended the analysis by relaxing the assumptions of the input signal model, and introduced two groups of performance measures. The performance of PCA based cue extraction degrades drastically when the input cue is not completely correlated at zero lag. The proposed shifted PCA method overcomes the problem by strategically time-shifting input signals prior to PCA decomposition. This approach extracted cue having the correct ITD and ILD, and increases the similarity of the extracted cue to the original cue. Simulation results and informal subjective listening tests verified the improved performance of SPCA over PCA for cue extraction in practical cases. Although several other methods have also been utilized to improve the extraction [12], [17], [19]-[22], the proposed shifted PCA based cue extraction in this paper is simple and effective.

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] M. M. Goodwin and J. M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," *in IEEE Int. Conf. on Acoust., Speech, and Sig. Process.*, Hawaii, Apr. 2007.

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

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

[4] 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.

[5] 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.

[6] C. Faller and J. Breebaart, "Binaural reproduction of stereo signals using upmixing and diffuse rendering," *in 131th Audio Eng. Soc. Conv.*, New York, Oct. 2011.

[7] 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, 2003.

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

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

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

[11] 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.

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

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

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York: John Wiley & Sons, 2001.

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

[16] 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.

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

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

[19] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio mixer," *IEEE Tran. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2141-2150, Sept. 2007.

[20] A. Härmä, "Stereo audio classification for audio enhancement," *in IEEE Int. Conf. on Acoust., Speech, Sig. Process.*, Prague, Czech, May 2011.

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

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

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

[24] A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35-39, Feb. 1948.

[25] W. A. Yost, "Perceptual models for auditory localization," *in 12th Audio Eng. Soc. Int. Conf.*, Copenhagen, Denmark, June 1993.

[26] 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.

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

[28] 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.

[29] C. Tournery and C. Faller, "Improved time delay analysis/synthesis for parametric stereo audio coding," *in 120th Audio Eng. Soc. Conv.*, Paris, France, May 2006.

[30] J. Allen, D.A. Berkeley and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals." *J. Acoust. Soc. Am.*, Vol. 62, No.4, pp. 912-915, October 1977.

[31] J. He. (2012 Nov. 30). A comparative testing of cue extraction using PCA and shifted PCA [Online]. Available: http://eeeweba.ntu.edu.sg/dsplab/audiobeam/jhe007/.