PRIMARY-AMBIENT EXTRACTION IN AUDIO SIGNALS USING ADAPTIVE WEIGHTING AND PRINCIPAL COMPONENT ANALYSIS

Karim M. Ibrahim
Nile University
k.magdy@nu.edu.eg

Mahmoud Allam
Nile University
mallam@nu.edu.eg

ABSTRACT

Most audio recordings are in the form of a 2-channel stereo recording while new playback sound systems make use of more loudspeakers that are designed to give a more spatial and surrounding atmosphere that is beyond the content of the stereo recording. Hence, it is essential to extract more spatial information from stereo recording in order to reach an enhanced upmixing techniques. One way is by extracting the primary/ambient sources. The problem of primary-ambient extraction (PAE) is a challenging problem where we want to decompose a signal into a primary (direct) and ambient (surrounding) source based on their spatial features. Several approaches have been used to solve the problem based mainly on the correlation between the two channels in the stereo recording. In this paper, we propose a new approach to decompose the signal into primary and ambient sources using Principal Component Analysis (PCA) with an adaptive weighting based on the level of correlation between the two channels to overcome the problem of low ambient energy in PCA-based approaches.
Key words: Audio Source Separation, Primary/ambient Separation, Surrounding Sound Systems, Upmixing.

1. INTRODUCTION

Currently, most audio recordings are available as 2-channel stereo recordings. For a long time, this has been considered sufficient to give the listener a pleasant experience. However, with new sound systems that give a better sense of surrounding and enclosing atmosphere, older recordings fail to utilize the capabilities of these new systems. Thus, it is important to develop methods of extracting additional spatial information from these recordings to enhance the experience of listening to them: this process is called upmixing [1, 2]. One approach is applying audio source separation to extract the original sources from the mixture, which are then rendered for the new playback system [3]. An important distinction between the different audio sources that can be used as a base for separating the sources is the ability to localize the sound sources. Separating sources based on their directional and diffuse features can be used in upmixing to create an immersive feeling.

5.1 surround systems [4] are an example of a multi-channel sound system commonly used in home theaters that are often used to play stereo recordings. A practical method of upmixing the stereo sound to the 5.1 system is by separating the primary (localizable) and ambient (non-localizable) sources and playing the primary sources on the two front channels to recreate the direct sources as it was intended in the original recording while playing the ambient sources on all channels to give a better feeling of surround sound.

Such applications call for advanced audio source separation methods. Hence, such methods have increasingly gained attention in the research community. Audio source separation can generally be categorized into two main challenges: blind audio source separation (BASS), where the goal is to extract the different sound sources in the mix, and primary-ambient extraction (PAE), where the goal is to separate between primary (direct) sources and ambient (diffuse) sources.

Several approaches have been proposed to extract the primary and ambient sources from a mixed-down recording. A commonly used approach is using Principal Component Analysis (PCA) as in [5, 6], which is investigated in detail later in this paper as it is the basis for the proposed approach. A different approach for the problem is using the least square method to estimate the primary and ambient sources as proposed by Faller in [7] by minimizing the errors between the extracted signals and the original stereo input.

In Avendano's work [8], the approach is to calculate a band-wise inter-channel short-time coherence from the cross- and autocorrelation between the stereo channels which is then used as the basis for the estimation of a panning and ambiance index. In Kraft's approach [9], the proposed method is based on the mid-side decomposition of stereo signals where the two-channel recording is split into "mid" signal that captures the centered content of the recording and a "side" signal that captures the content panned to the left and right side.

The focus of this paper lies in developing a new technique for primary-ambient extraction in stereo signals and to introduce an evaluation method for PAE to compare between the different commonly used approaches and our new proposed method.

The paper is structured as follows: Section 2 explains the problem definition of audio source separation and primary ambient extraction, the possible application for these tech-

niques and the constraints for an ideal extraction.

Section 3 explains our proposed method to improve the separation based on Principal Component Analysis (PCA). Finally, Section 4 shows the evaluation between the proposed method and the previous methods from the literature.

1.1 Notation

The convention in this paper is to express signals in the time domain in lower case letter as x, while signals in the STFT domain are in upper case as X. Scalar variables are expressed in normal italic font as X while column vectors are expressed in bold italic font as $\boldsymbol{X}$ and matrices are expressed in bold non-italic font as $\mathbf{X}$.

Table 1 shows the commonly used symbols in this paper:

$\boldsymbol{x}$	Mixed stereo signal
$\boldsymbol{x}_l$, $\boldsymbol{x}_r$	left and right channels of a sound mixture
$\boldsymbol{p}_l$,$\boldsymbol{p}_r$	Left and right primary components
$\boldsymbol{a}_l$,$\boldsymbol{a}_r$	Left and Right ambient components
n	Discrete time index
m	Frequency index
k	Frame index
w_{pl},w_{pr}	weighting factor of the primary source
$\boldsymbol{v}$	Normalized unit vector of 1^{st} Principal component

Table 1: Symbols used in this paper

2. PRIMARY-AMBIENT EXTRACTION

One of the key characteristics in spatial audio is whether an audio source is localizable or not. A localizable source is perceived as coming from a certain direction and the listener can determine this direction, also called primary or directional source. A non-localizable source is perceived as a surrounding sound, coming from all around, also called an ambient or diffuse sound. Ambient sources usually describe the surrounding atmosphere of the recording . Methods for separating these two types of sources have been receiving increasing attention for applications such as upmixing [10, 11], multichannel format conversion and headphone reproduction [12, 13].

2.1 Signal model for PAE

When approaching the problem of primary-ambient extraction, we consider the input signal as a mix of two sources; a primary and an ambient source. In this paper, we only approach the problem of separating the mixture of a stereo signal.

Stereo recordings consist of two channels that contain both the primary and ambient sources mixed together and the goal is to separate them. The signals can be expressed as follows:

$$x_l[n] = p_l[n] + a_l[n] \quad (1)$$

$$x_r[n] = p_r[n] + a_r[n] \quad (2)$$

where x_l, x_r are the left and right channels of the stereo recording respectively, p_l, p_r are the primary component in each channel, a_l, a_r are the ambient component and n is the time index of the discrete signals.

Most PAE approaches are applied in the STFT domain as it is safer to assume there is only one primary source and one ambient source in each frequency-frame sub-band. The signals are expressed then in the form:

$$X_l[m,k] = P_l[m,k] + A_l[m,k] \quad (3)$$

$$X_r[m,k] = P_r[m,k] + A_r[m,k] \quad (4)$$

where m, k are the frame and frequency index respectively.

2.2 Sound localization and human auditory system

To be able to precisely separate the primary and ambient sources, it is necessary to understand how the human auditory system works and how it determines the location of a sound source and then use the same characteristics in the separation process.

The human auditory system uses several cues to localize a sound source, including inter-channel time difference (ICTD), also referred to as inter-aural time difference (ITD), inter-channel level difference (ICLD), also referred to as inte-aural level difference (ILD), spectral information and correlation analysis [14].

A comparison between the two channels should be sufficient to extract the directional information of an audio source. The correlation between the two channels plays a significant role in determining the location of the source, i.e., an ambient source shows no correlation between the two channels, making it impossible for the human auditory system to determine the direction of the sound. Hence, calculating the correlation between the two channels is usually a necessary step in extracting the primary and ambient sources.

2.3 PAE applications: upmixing to 5.1 systems

A common application for PAE is upmixing from n to m channels, where $m > n$. Here, we explain how to use PAE in upmixing to one of the commonly used systems, the 5.1 surround system. By separating the primary and ambient sources using one of the PAE methods, the extracted sources are re-panned in a way that the left primary sources $p_l[n]$ are played on the front left and center channels, $x_{lf}[n]$ and $x_c[n]$ while the right primary sources are played on the front right and center channels $x_{rf}[n]$ and $x_c[n]$ and the ambient sources are played throughout the five speakers. This way, the directionality of the primary sources are kept as originally intended while the surrounding sound is enhanced by the ambient sources. Figure 1 shows the block diagram of the upmixing technique.

2.4 PAE assumptions

To accurately separate between the primary and ambient components, we need to define the constraints that achieve the right separation. By definition, the primary sources are localizable while the ambient sources are non-localizable.

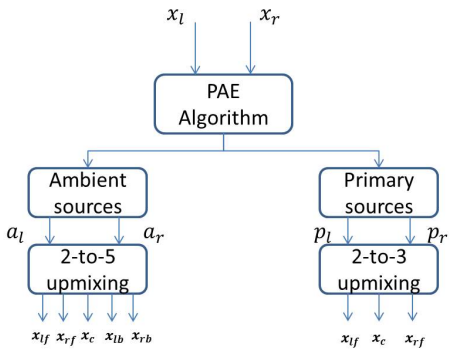

Figure 1: Block diagram of the stereo to 5.1 upmixing using PAE

To find a mathematical representation for this definition, we need to review the sound localizing process in the human auditory system mentioned in section 2.2. The key characteristic in localizing the sound sources is the correlation between the two signals reaching the left and right ears. In the case of a complete non-localizable diffuse source, the two signals are expected to be orthogonal in a way that the brain fails to detect any similarity between the left and right signals to extract location information. Similarly, primary sources are expected to be partially or fully correlated. Based on the representation of the stereo signal in equation (3) and assuming that the left and right primary components are P_l, P_r respectively, where P_l, P_r are vectors of adjacent STFT frames, Similarly the left and right ambient components are A_l, A_r, ω is the scaling factor between the primary components in the two channels due to ICLD and $\boldsymbol{A}^H$ is the Hermitian transpose of the vector $\boldsymbol{A}$, these constraints are defined according to [5] as:

1. The primary components are correlated

$$\boldsymbol{P}_l = \omega \boldsymbol{P}_r \tag{5}$$

2. The ambient components are orthogonal (fully uncorrelated)

$$\boldsymbol{A}_l^H \boldsymbol{A}_r = 0 \tag{6}$$

3. The ambient and primary components are orthogonal to each other

$$\boldsymbol{P}_l^H \boldsymbol{A}_l = 0 \qquad \boldsymbol{P}_r^H \boldsymbol{A}_r = 0 \tag{7}$$

4. The two ambient components have almost the same energy level

$$\boldsymbol{A}_l^H \boldsymbol{A}_l \approx \boldsymbol{A}_r^H \boldsymbol{A}_r \tag{8}$$

Figure 2 shows the assumed constraints between the different components.

2.5 PCA-Based PAE

Many of the approaches of PAE are based on the Principal Component Analysis (PCA) as in [5,6,15–18]. PCA is widely used since the common signal model assumes that the stereo signal is composed of primary sources that are

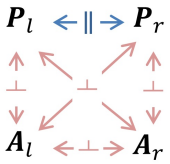

Figure 2: Constraints on the primary and ambient components

highly correlated and ambient diffuse sources. It is suitable to use a decomposition method such as PCA to extract the correlated primary sources and to assume the ambient sources are the residuals. The work in [15] is also based on the PCA but with an important modification, it takes into consideration the Inter-Channel Time Difference (ICTD) by using a time-shifting technique to improve the extraction of the primary sources.

One major drawback of methods based on PCA is the assumption that there is always a primary source in each frequency-frame sub-band and that it is never too weak. This is evident from the extraction of the primary sources as the first principal component. In case of absence of any primary sources, the method would still assign the first principal component, the one with the highest energy, to the primary source, which clearly produces a significant error in this particular case.

3. IMPROVING PCA-BASED APPROACH

As described in Section 2.5, the PCA-based approach has a number of drawbacks that impairs its accuracy. The solution we propose is to add an adaptive weighting to increase the amount of energy the ambient signal. The concept of adaptive weighting in PCA was previously introduced by Goodwin [19] with a different weighting scheme. The weighting we propose is based on the relation between the two channels of the signal in a way that supports the ambient extraction by detecting the level of presence of the primary sources. One way to do this is by considering the second dominant eigenvalue and comparing its value to the dominant eigenvalue. In the case of high correlation, the first (dominant) eigenvalue will be considerably larger than the second eigenvalue. In this case it would be safe to decompose the signal into primary and ambient components. However, in the case of having a more dominant ambient source, the ratio between the first and second eigenvalues will be relatively small.

The PAE using our weighting scheme is applied as follows:

1. We start with the original 2-channel signals, $x_l[n]$ and $x_r[n]$. We apply the STFT on the signals to get $X_l[m,k]$ and $X_r[m,k]$, where m is the frame index and k is the frequency index. We calculated the STFT using $\frac{3}{4}$ overlapping Hamming windows of Length 4096 samples, corresponding to a duration of 92.8 milliseconds at a sampling frequency of 44.1 kHz.

2. For each frequency-frame bin we define a vector with

the STFT values of the M adjacent frames:

$$\boldsymbol{X}_{l,r}[m,k] = \begin{bmatrix} X_{l,r}[m-M,k] \\ \vdots \\ X_{l,r}[m+M,k] \end{bmatrix} \quad (9)$$

For brevity, the index $[m,k]$ is dropped in the following equations.

3. The decomposition is then applied per frame-frequency index to extract the primary and ambient components in each frame-frequency using these two vectors as shown in Figure 3.

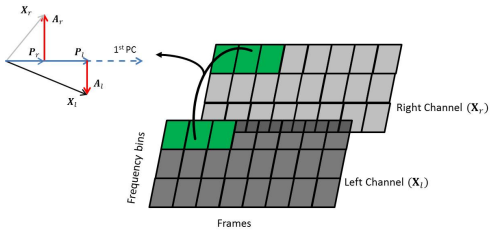

Figure 3: PAE using PCA

4. Using the Eigendecomposition on the covariance matrix $\mathbf{C}$ of the two vectors $\boldsymbol{X}_l, \boldsymbol{X}_r$, we get the normalized dominant Eigenvector $\boldsymbol{V}$ and the first two dominant Eigenvalues λ_1, λ_2.

5. Next we calculate the weighting factor ω based on the ratio between the two eigenvalues. The primary weights are defined as:

$$\omega = 1 - \frac{\lambda_2}{\lambda_1} \quad (10)$$

This weighting ensures that in cases of low correlation between the two channels, a higher weight is given to the ambient component.

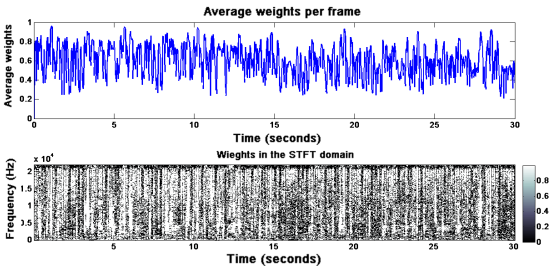

Figure 4: Sample of the weights extracted from an audio file

6. In order to enhance the ambient extraction, we define a threshold θ. The goal is to detect the cases where there is no strong presence of a primary source, so all the content is put into the ambient component. The primary component is still weighted by ω in case of passing the threshold to support extracting the ambient component.

$$\boldsymbol{P}_{l,r} = \begin{cases} \omega(\boldsymbol{V}^H \boldsymbol{X}_{l,r})\boldsymbol{V}, & \omega > \theta \\ 0, & \omega < \theta \end{cases} \quad (11)$$

$$\boldsymbol{A}_{l,r} = \boldsymbol{X}_{l,r} - \boldsymbol{P}_{l,r} \quad (12)$$

where P_l, P_r, A_l, A_r are the primary and ambient components of the right and left channel respectively. Figure 4 shows an example of the weights extracted from an audio file.

7. Finally, two schemes can be used to extract the information from each frame; either to merge the extracted vectors by averaging them or to take out the center point of each vector as shown in figure 5. However, the results of both schemes are very similar, so we can use the "take center" scheme to reduce the computation.

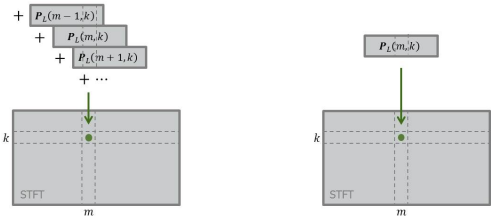

Figure 5: Different schemes of merging the output vectors

4. EVALUATION

Our objective evaluation is based on the work done in [20] which is intended to evaluate blind audio source separation (BASS). It can be used for primary-ambient separation, as well, by assuming that the mixture of sources is made out of only two sources, an ambient and a primary one. Ideally the extraction methods should output two sources that are identical to the originals ones. However, due to the limitations of the extraction methods, there is interference between the two sources.

In the following we would like to compare the following approaches:

1. PCA-based PAE without weighting, referred to as "PCA without weighting".

2. The weighted PCA method by Goodwin in [5, 6]. Referred to as (PCA Goodwin).

3. The extraction method by Avendano and Jot in [8]. referred to as (Avendano)

4. The modified PCA method with the weighting scheme proposed in this paper with two different threshold values: $\theta = 0.5$ and $\theta = 0.9$.

The evaluation was performed using two databases, one is made out of all ambient sources, consisting of strong ambient sources as sounds of crowd, forest, rain and echoes, and the second is made of all primary sources, consisting of strong primary sources as vocal recordings, solo instruments and dialogs. Each of the two data sets consist of 40 different recordings that are mixed together to compose 40 mixed recordings. We used the Matlab toolbox "BSS Eval" [21] for calculating the errors. The evaluation is as follows:

1. Mixing one ambient source with one primary source after normalizing the two of them, by ensuring the highest energy level of the two sources is the same, so no source would be more prominent than the other.

2. Applying the five different PAE methods to extract the primary and ambient sources.

3. Use the extracted outputs and the original sources to evaluate each method using BSS Eval.

4. A baseline is defined by comparing the original ambient or primary sources to the mixture without any separation. This is used to define the improvement of each extraction method over the original mixture.

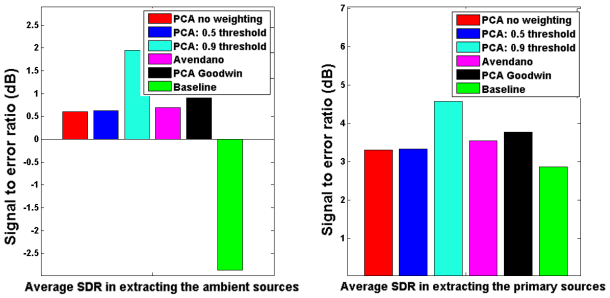

Figure 6: Average SDR in primary and ambient extraction

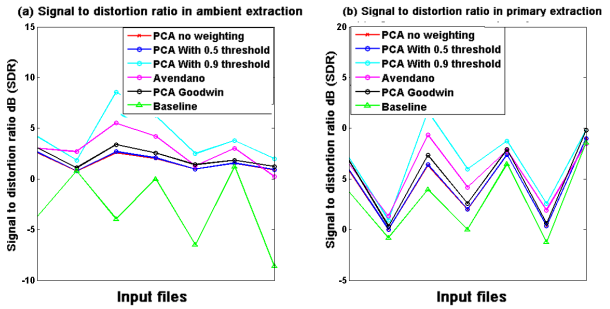

Figure 7: SDR values for a sample of five mixtures

Figure 6 shows the average Signal to Distortion ratio (SDR) in extracting both the primary and the ambient sources for different methods. By analyzing the graph, we find that the proposed weighting shows an improvement in the separation over the other methods. We find that using a higher threshold of 0.9 gives much better separation than using a lower threshold or no threshold. This shows how the weighting improves the accuracy of extraction over both the original PCA and the weighted PCA introduced by Goodwin in [19].

Figure 7 shows the exact SDR values of a sample of five mixtures with comparison to the baseline in both the primary and ambient extraction. We find that all the methods improve clearly over the baseline without separation. In general the SDR values for the primary extraction is higher than the ambient because the primary sources tend to have higher energy.

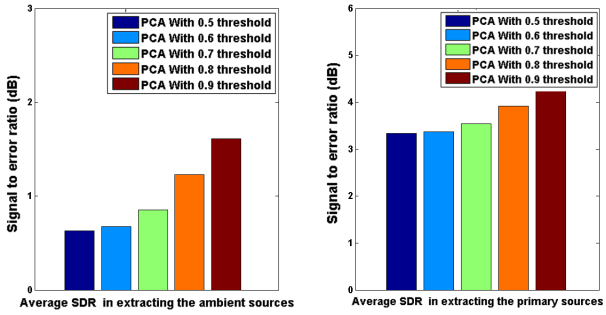

Figure 8: Average SDR for different thresholds

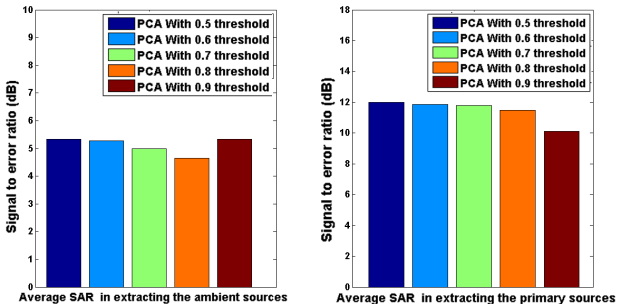

Figure 9: Average SAR for different thresholds

Figure 8 shows how using different thresholds affects the extraction quality. Using higher thresholds gives higher accuracy in the separation, however, extreme weights result in higher distortion caused by the artifacts in the separation process, especially in the extracted primary sources, as shown in Figure 9. Hence, there is a trade-off between sharp separation and artifact distortion. Typically, a threshold in the range $\theta \in [0.6, 0.8]$ would give a proper trade-off between separation quality and artifact distortion.

5. CONCLUSIONS

Separating the primary and ambient sources from an audio mixture shows potential for applications including upmixing an audio recording. In this paper, we explained the need for this separation technique and proper ways of using it in upmixing techniques. We presented a method of extracting the sources using an adaptive Principal Component Analysis (PCA) to solve the common problem of the dominant primary source. The adaptive weighting tests the level of presence of primary sources and ensures to give a proportional weight to both of the sources based on this estimate. The method shows higher separation quality compared to the classic PCA-based separation methods and other methods from the literature. However, this method still shows correlation between the two ambient components leaving room for further improvement in future work. Future work could also include a subjective

evaluation by performing listening test with the different separation methods to ensure the user's experience coincide with the results of the objective evaluation.

6. REFERENCES

[1] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Proc. Int. Conf. Audio Engineering Society: 28th International Conference: The Future of Audio Technology–Surround and Beyond*. Audio Engineering Society, 2006.

[2] M. R. Bai and G.-Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *J. Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, 2007.

[3] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *Proc. Int. Conf. Digital Signal Processing (DSP), 2011*. IEEE, 2011, pp. 1–7.

[4] B. Xie, "Signal mixing for a 5.1-channel surround sound system'analysis and experiment," *J. Audio Engineering Society*, vol. 49, no. 4, pp. 263–274, 2001.

[5] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2007, pp. I–9.

[6] M. M. Goodwin, "Geometric signal decompositions for spatial audio enhancement," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 409–412.

[7] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, 2006.

[8] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Engineering Society*, vol. 52, no. 7/8, pp. 740–749, 2004.

[9] S. Kraft and U. Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 2015.

[10] M. R. Bai and G.-Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *J. Consumer Electronics*, vol. 53, no. 3, pp. 1011–1019, 2007.

[11] C. Faller and J. Breebaart, "Binaural reproduction of stereo signals using upmixing and diffuse rendering," in *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[12] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *J. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1503–1511, 2008.

[13] W.-S. Gan, E.-L. Tan, and S. M. Kuo, "Audio projection," *J. Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 43–57, 2011.

[14] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[15] J. He, E.-L. Tan, and W.-S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 266–270.

[16] S.-W. Jeon, D. Hyun, J. Seo, Y.-C. Park, and D.-H. Youn, "Enhancement of principal to ambient energy ratio for pca-based parametric audio coding," in *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 385–388.

[17] J. Merimaa, M. M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[18] S. Dong, R. Hu, W. Tu, X. Zheng, J. Jiang, and S. Wang, "Enhanced principal component using polar coordinate pca for stereo audio coding," in *Proc. Int. Conf. Multimedia and Expo (ICME)*. IEEE, 2012, pp. 628–633.

[19] M. M. Goodwin, "Adaptive primary-ambient decomposition of audio signals," Jun. 19 2012, uS Patent 8,204,237.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *J. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.