F0-estimation-based primary ambient extraction for stereo signals

Hanxin Zhu, Chuang Shi, Yue Wang
School of Information and Communication Engineering
University of Electronic Science and Technology of China, Chengdu, China

***Abstract*—Primary-ambient extraction (PAE) plays an increasingly important role in spatial audio reproduction to achieve an immersive listening experience. The existing PAE algorithms produce notable extraction errors, especially when the primary components are relatively small in magnitude as compared to the ambient components. In this paper, an F0-estimation-based PAE method is proposed. This method explores harmonic structures of the primary components to tap the full potential and utilize the sparsity constraint. The experiment results validate that the F0-estimation-based PAE method achieves 5 dB lower extraction errors than the principal component analysis (PCA) method and the ambient phase estimation with a sparsity constraint (APES) method.**

***Index Terms*—Spatial audio reproduction; Primary ambient extraction; Harmonic structure; F0 estimation**

I. Introduction

Spatial audio reproduction has gained increasing importance in the entertainment industry these recent years. Audio files usually comprises both point-like directional sound sources and diffused environmental sound, which are usually referred to as primary components and ambient components, respectively [1]. They are perceived differently by human hearing. Hence, it is essential to make use of different rendering schemes to gain an optimal listening experience [2]. However, the primary and ambient components are usually mixed in the existing mainstream audio formats (e.g., stereo, multichannel signals) [3], which suggests that an extraction of the primary and ambient components becomes a necessity. Many fields in spatial audio processing such as spatial audio coding, audio up-mixing and immersive 3D sound systems have witnessed the applications of PAE [4]–[7].

The basic signal model of PAE is illustrated in Fig. 1. The primary components in different channels are correlated with each other while the ambient component in each channel is uncorrelated with the primary components and other ambient components [8]. Based on such a signal model, several PAE methods have been introduced [9]. The least-squares (LS) method extracts primary and ambient components on the basis of least squares criterion [10]. The PCA method calculates the correlation values between different channels to evaluate the correlation components of the input signals as primary components [11]. Jot *et al.* use the time-frequency masking method to extract environmental sound components from stereo signals [12]. In many cases of PAE, since the primary components are essentially speech-like signals, they can be considered to be sparse in the time-frequency domain. He *et al.* estimates the phases of the ambient components by enhancing the sparsity of the primary components and thus puts forward the state-of-the-art APES method [13]. Among the four methods mentioned above, the APES method shows the best overall performance, while the other three methods result in larger extraction errors.

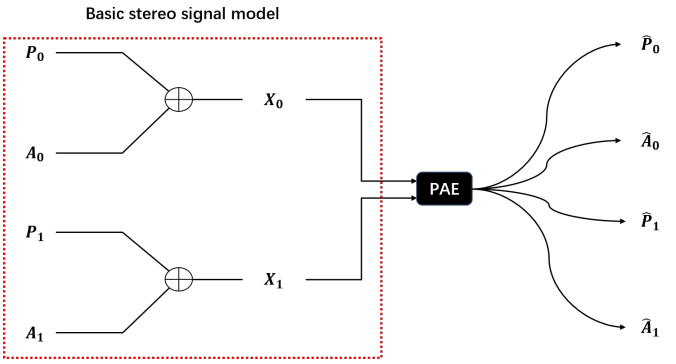


Fig. 1. Basic stereo signal model of PAE, where $\mathbf{X}_0$ and $\mathbf{X}_1$ are the input signals; $\mathbf{P}_0$ and $\mathbf{P}_1$ are the pure primary components; $\mathbf{A}_0$ and $\mathbf{A}_1$ are the pure ambient components; $\widehat{\mathbf{P}}_0$ and $\widehat{\mathbf{P}}_1$ are the extracted primary components; $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{A}}_1$ are the extracted ambient components

The APES method is still not the optimal method. At each time-frequency point, the APES method uses only the criterion of minimum amplitude of the primary components. Clearly such a method does not make full use of the sparsity of the primary components. The primary components extracted by the APES method still suffer from major leakage from the ambient components. This paper proposes to utilize more sparsities of the primary components to reduce the leakage of ambient components to a larger extent to ultimately extract purer primary components. A novel method to achieve a more accurate PAE result by means of F0 estimation, referred to as the PAEF method in short, is thus put forward.

II. Basic Stereo Signal Model

The input signal of each channel is firstly converted into time-frequency domain by means of short-time Fourier transform (STFT). For each frequency band or subband within

This manuscript is prepared based on the research work supported jointly by the National Natural Science Foundation of China and the Civil Aviation Administration of China (Joint Grant No. U1933127).

Corresponding author's email: shichuang@uestc.edu.cn

a time frame, it is usually assumed that only one dominant directional sound source constitutes the primary components [1]. After performing STFT on the input signals, every time-frequency point is denoted as $X_c(m,n)$ where m represents the index of time frames, n represents the index of frequency bands and c represents the index of channels ($c \in \{0,1\}$ for stereo signals). Thus, for one subband b which includes frequency points from $n_{b-1}+1$ to n_b (n_b is the upper boundary of frequency index of subband b) in time frame m, the subband signal can be donated as $\mathbf{X}_c[m,b] = [X_c(m,n_{b-1}+1), X_c(m,n_{b-1}+2), \cdots, X_c(m,n_b)]^T$.

Therefore, the stereo signal model can be written as

$$\mathbf{X}_c[m,b] = \mathbf{P}_c[m,b] + \mathbf{A}_c[m,b] \ \forall\ c \in \{0,1\}, \tag{1}$$

where $\mathbf{P}_c$ is the primary component and $\mathbf{A}_c$ is the ambient component of the input signal. In the following part of this paper, the subscript $[m,b]$ is omitted.

Previously, it is assumed in PAE that the primary components are correlated. They are localized as a result of the inter-channel level difference (ICLD) and inter-channel time difference (ICTD) [8]. The primary components between different channels have relatively more obvious ICLD in stereo recording that uses coincident microphone techniques and sound mixes that uses pan-pot stereo techniques. Therefore, this paper follows the classic simplification of PAE that only ICLD is taken into account, *i.e.* $\mathbf{P}_1 = k\mathbf{P}_0$ and k is called the primary panning factor. The mixed input signals of the primary and ambient components do not explicitly show the real value of k. The estimate of k is then carried out on a frame-by-frame basis by

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

where r_{11}, r_{00} and r_{01} are denoted as the auto-correlations and cross-correlation of the input signal [5].

The ambient components of two channels usually have low cross-correlation because of the diffuseness of environmental sounds. Therefore, decorrelation techniques are commonly applied to produce diffuse ambient components from raw recordings [9]. As the decorrelation techniques produce equal magnitude of ambient components in the two channels of the stereo signals, the ambient components are considered to have equal power in different channels, *i.e.* $|\mathbf{A}_1| = |\mathbf{A}_0|$, and to be uncorrelated with the primary components. The primary power ratio γ is defined as the ratio of total primary power to total signal power in two channels, where γ is used to quantify the power difference between the primary and ambient components ($\gamma \in [0,1]$). Previous studies have revealed that the performance of PAE is positively related with the value of γ. However, a PAE method that can achieve significant performance irregardless of the value of γ is highly sought after.

III. F0-Estimation-Based Primary Ambient Extraction

According to the basic stereo signal model, the PAE problems can be formulated as

$$\mathbf{X}_c = \mathbf{P}_c + \mathbf{A}_c\ , \forall\ c \in \{0,1\}, \quad s.j. \begin{cases} \mathbf{P}_1 = k\mathbf{P}_0 \\ |\mathbf{A}_1| = |\mathbf{A}_0| \end{cases} \tag{3}$$

in which the ambient components are further expressed as

$$\mathbf{A}_c = |\mathbf{A}_c| \odot \mathbf{W}_c \ \ \forall\ c \in \{0,1\}, \tag{4}$$

where $\odot$ donates element-wise Hadamard product and the element in the time-frequency point (m,n) of $\mathbf{W}_c$ is donated as $\mathbf{W}_c(m,n) = e^{j\theta_c(m,n)}$. $\theta_c(m,n) = \angle A_c(m,n)$ is the phase of $A_c(m,n)$.

Since $\mathbf{P}_1 = k\mathbf{P}_0$ and $\mathbf{X}_1 - k\mathbf{X}_0 = \mathbf{A}_1 - k\mathbf{A}_0$, substituting (4) yields

$$|\mathbf{A}_1| = |\mathbf{A}_0| = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0), \tag{5}$$

where $./$ is the element-wise division. On account that $|\mathbf{A}_1|$ is real and non-negative, the relationship between the phases of the ambient components in two channels is given by

$$\boldsymbol{\theta}_0 = \boldsymbol{\theta} + arcsin[sin(\boldsymbol{\theta} - \boldsymbol{\theta}_1)/k] + \pi, \tag{6}$$

where $\boldsymbol{\theta} = \angle(\mathbf{X}_1 - k\mathbf{X}_0)$ [13].

Substituting (5) and (6) into (3) yields

$$\mathbf{A}_c = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c \tag{7}$$

and

$$\mathbf{P}_c = \mathbf{X}_c - \mathbf{A}_c = \mathbf{X}_c - (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \odot \mathbf{W}_c. \tag{8}$$

If $\boldsymbol{\theta}_1$ is known, $\boldsymbol{\theta}_0$ can be calculated by (6). $\mathbf{W}_1 = e^{j\boldsymbol{\theta}_1}$ and $\mathbf{W}_0 = e^{j\boldsymbol{\theta}_0}$ are thereafter readily written out. By (7) and (8), $\mathbf{A}_c$ and $\mathbf{P}_c$ can then be obtained. The key to PAE is now converted to the estimation of $\boldsymbol{\theta}_1$.

When the primary components have an obvious harmonic structure in a duration of T_1 and the fundamental frequency is denoted as f_0, $\mathbf{P}_c$ can be expressed as

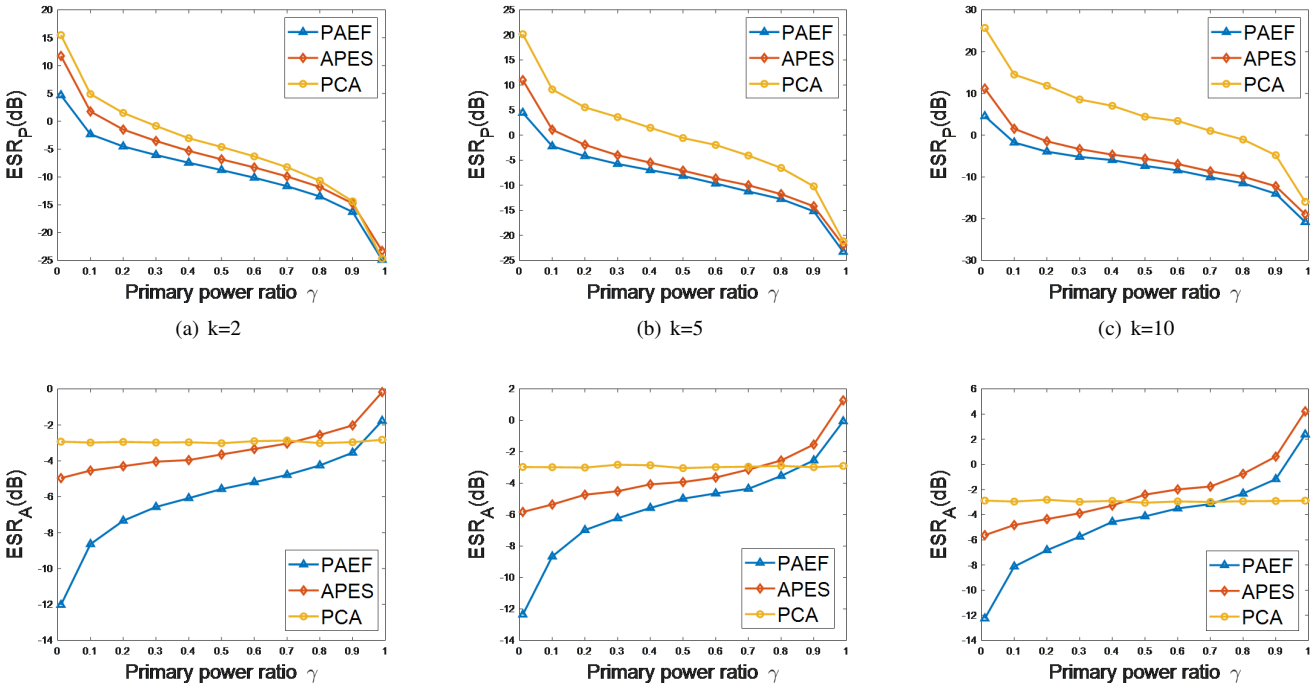
$$\mathbf{P}_c = \frac{2sin(fT_1/2)}{f} \times \sum_{r=-\infty}^{+\infty} P_{f_r}\delta(f - rf_0), \forall\ c \in \{0,1\}, \tag{9}$$

where P_{f_r} represents the weight of the primary components in the time-frequency point $(m,\ rf_0)$, $r \in (-\infty, \infty)$ and r is an integer.

Within the duration of T_1, the primary components have zero energy for all frequency points except rf_0. Estimating the optimal value of θ_1 is only requested at rf_0, which can be carried out by minimizing the sum of the modulus of the ambient components in the two channels at the frequency point rf_0, *i.e.*

$$\widehat{\theta}_1 = argmin(|A_1| + |A_0|), f = rf_0. \tag{10}$$

When it is outside the rectangular window, similar to the APES method, the values of $\boldsymbol{\theta}_1$ can be estimated by minimizing the



(a) k=2 (b) k=5 (c) k=10

(d) k=2 (e) k=5 (f) k=10

Fig. 2. ESR of (a)–(c) extracted primary component and (d)–(f) extracted ambient component using PAEF, APES, and PCA.

TABLE I
ACCURACY OF F0 ESTIMATION WHEN $\gamma = 0.1$

k	2	5	10
Correct rate of first experiment	92.91%	90.20%	87.50%
Correct rate of second experiment	90.54%	87.84%	91.22%
Correct rate of third experiment	90.88%	89.86%	88.51%
Average	91.44%	89.30%	89.08%

sum of the modulus of the primary components in the two channels, *i.e.*

$$\widehat{\theta}_1 = argmin(|P_1| + |P_0|). \tag{11}$$

Combining (10) and (11), $\boldsymbol{\theta}_1$ is proposed in this paper to be estimated by

$$\widehat{\theta}_1 = \begin{cases} argmin(|A1| + |A0|), f = rf0 \\ argmin(|P1| + |P0|), f \neq rf0. \end{cases} \tag{12}$$

This proposed method requests the estimate of f_0 [14]–[17]. Thus, it is called the F0-estimation-based PAE method and abbreviated as the PAEF method.

IV. RESULTS AND ANALYSIS

In this section, the PAEF method is compared with the state-of-the-art methods, PCA and APES. A speech signal is selected as the primary component with each time frame comprising 1024 samples. The amplitude panning factor k of the primary components is set to 2, 5 and 10. The ambient components are wave lapping sounds, which are decorrelated by all-pass filters with random phases. The input signal that uses different values of γ ranging from 0 to 1 with an interval of 0.1 is obtained by mixing the primary and ambient components together. Since the objective functions in (12) are not convex, a direct searching (DS) method is implemented to estimate the phase of the ambient components [13]. The optimal phase estimation can be selected from an array of phase values $\widehat{\theta}_1(d) = (2\pi d/D - \pi)$, where $d \in (1, 2, \cdots, D)$ with D being the total number of phase values to be considered.

A. Accuracy of F0 Estimation

As shown in Fig. 2, since the traditional methods of PAE have good extraction performance when the power of the primary component is large, the case when $\gamma = 0.1$ is specifically investigated to compare the accuracy of F0 estimation. The accuracy of F0 estimation is quantified by the percent correct rate, which is defined as

$$Correct\ rate = \frac{N_{est,pure}}{N_{pure}} \times 100\%, \tag{13}$$

where N_{pure} is the number of frames of the pure primary component and $N_{est,pure}$ is the number of frames in which the estimated F0 equals to the ground truth. To ensure the robustness, the experiment is repeated thrice. From Table I, it is found that even when γ is very small, the accuracy is satisfactorily high, achieving about 90% on an average.

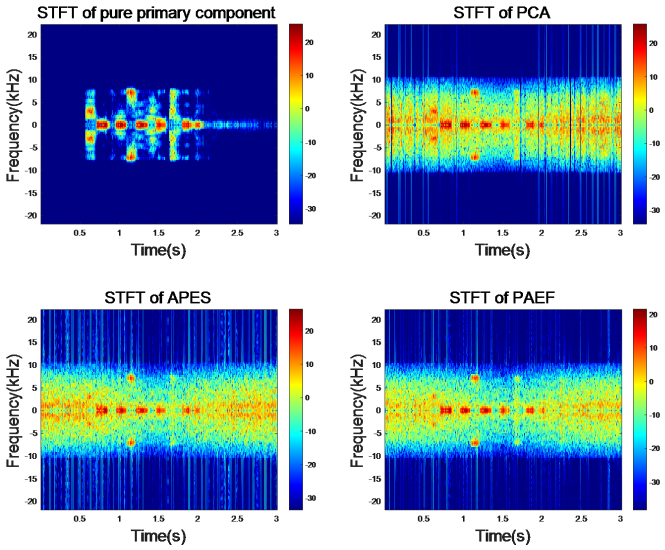


Fig. 3. STFT of pure primary component and STFT of extracted primary component using the PCA, APES and PAEF methods when k=2, γ=0.1

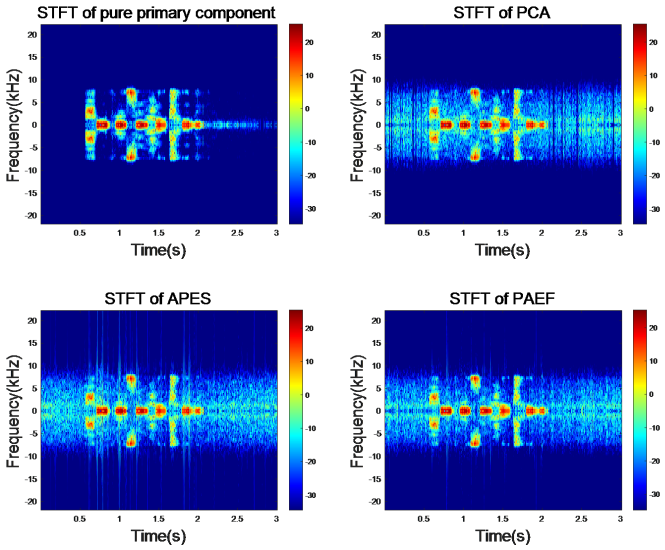
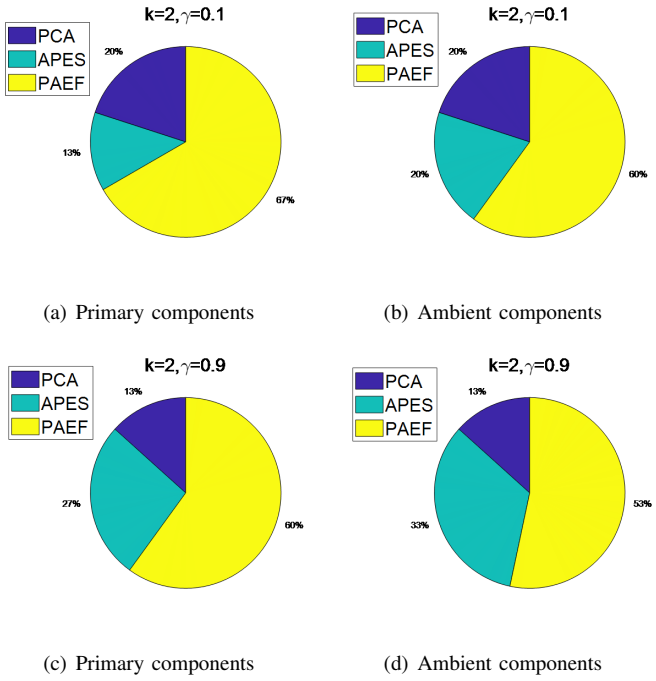


Fig. 4. STFT of pure primary component and STFT of extracted primary component using the PCA, APES and PAEF methods when k=2, γ=0.9



(a) Primary components (b) Ambient components

(c) Primary components (d) Ambient components

Fig. 5. Subjective listening test results of PAE using the PCA, APES and PAEF methods when k=2, γ=0.1, 0.9

B. Objective Comparison of the PCA, APES and PAEF Methods

The PCA, APES and PAEF methods are compared by the error-to-signal ratio (ESR, in dB) of the extracted primary and ambient components. Lower ESR indicates better PAE performance. The ESR for the primary and ambient components are calculated as

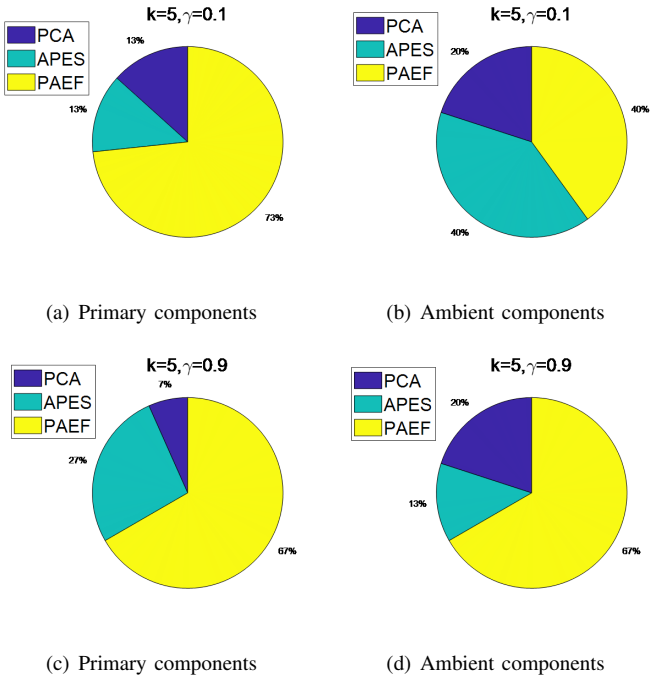
$$ESR_q = 10log_{10}\left\{\sum_{c=0}^{1}\frac{||\boldsymbol{q}_c - \hat{\boldsymbol{q}}_c||_2^2}{2||\boldsymbol{q}_c||_2^2}\right\}, \forall \boldsymbol{q} \in \boldsymbol{P},\ or\ \boldsymbol{A}. \quad (14)$$

According to Fig. 2, the PAEF method outperforms the PCA and APES methods in terms of both ESR_P and ESR_A. First of all, unlike the PCA method, the PAEF method inherits the good stability of APES and thus its primary component extraction performance will not change with k. Secondly, when compared to the APES method, the extraction error of the PAEF method is reduced by about 2-7dB. The smaller the value of γ is, the less the extraction error will be. However, because the PAEF method requires estimation of the fundamental frequency, it needs twice the computing time as compared to the APES method.

In Fig. 3 and Fig. 4, the STFT of the pure primary component and STFT of the extracted primary component using the PCA, APES and PAEF methods when k=2, γ=0.1 or 0.9 are plotted. When γ is small, the PAEF method performs the best among the three methods. When γ is close to 1, performances of the three methods are similar.

C. Subjective Comparison of the PCA, APES and PAEF Methods

In this subsection, 15 participants are invited to evaluate the PAE performance of the PCA, APES and PAEF methods by subjective listening tests. The pure primary and ambient components and the primary and ambient components extracted by the three methods are used. Participants are asked to choose their favourite version from the extracted signals when k=2, 5 and γ=0.1, 0.9. The subjective listening test results are shown in Fig. 5 and Fig. 6. About 60% participants prefer the primary and ambient components extracted by the PAEF method. Among the remaining participants, more than half of them prefer the APES method in most test cases.



(a) Primary components (b) Ambient components

(c) Primary components (d) Ambient components

Fig. 6. Subjective listening test results of PAE using PCA, APES and PAEF when k=5, γ=0.1, 0.9

V. Conclusions

This paper proposes the PAEF method based on the fact that most primary components have an obvious harmonic structure. Compared with the state-of-the-art PAE methods such as the APES and PCA methods, the PAEF method makes better use of the sparsity of the primary components. Objective comparison reveals that the PAEF method outperforms the APES and PCA methods, especially in the presence of relatively strong ambient components. The PAEF method achieves 5dB less extraction error than the other two methods on an average. Moreover, subjective comparison also validates the advantage of the PAEF method. When there are no obvious harmonic structures in the primary components, future works should develop a hybrid of the PAEF and APES methods for robust spatial audio coding.

References

[1] M. M. Goodwin and J. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, 2007.

[2] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, 2020.

[3] M. M. Goodwin, "Primary-ambient decomposition and dereverberation of two-channel and multichannel audio," *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2008, pp. 797-800.

[4] M. R. Bai and G. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 3, pp. 1011-1019, Aug. 2007

[5] G. Del Galdo, F. Kuech, M. Kallinger and R. Schultz-Amling, "Efficient merging of multiple audio streams for spatial sound reproduction in Directional Audio Coding," *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 265-268.

[6] C. Shi, H. Nomura, T. Kamakura, W. S. Gan, Spatial aliasing effects in a steerable parametric loudspeaker for stereophonic sound reproduction, *IEICE Trans. Fund. Electron. Commun. Computer Sci.*, **E97-A**, 1859–1866 (2014).

[7] K. M. Ibrahim and M. Allam, "Primary-Ambient Source Separation for Upmixing to Surround Sound Systems," *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, 2018, pp. 431-435.

[8] J. He, W. Gan and E. Tan, "Primary-Ambient Extraction Using Ambient Spectrum Estimation for Immersive Spatial Audio Reproduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1431-1444, Sept. 2015.

[9] L. Chen, C. Shi, and H. Li, "Primary ambient extraction for random sign Hilbert filtering decorrelation," *Proceedings of the 23rd International Congress on Acoustics*, Aachen, Germany, 2019.

[10] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Societ*, vol. 59, no. 6, p. 431, 2011.

[11] J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings," *Proceedings of the 123th Audio Engineering Society Convention*, New York, NY, USA, Oct. 2007.

[12] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," *Proceedings of the 123th Audio Engineering Society Convention*, San Francisco, CA, USA, 2012.

[13] J. He, W. Gan and E. Tan, "Primary-ambient extraction using ambient phase estimation with a sparsity constraint," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1127-1131, 2015.

[14] K. Miwa and M. Unoki, "Study on Method for Estimating F0 of Steady Complex Tone in Noisy Reverberant Environments," *Proceedings of the 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Beijing, China, 2013, pp. 456-459.

[15] M. G. Christensen, "On the estimation of low fundamental frequencies," *Proceedings of the 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2011, pp. 169-172.

[16] P. S. Rathore and R. B. Pachori, "Instantaneous fundamental frequency estimation of speech signals using DESA in low-frequency region," *Proceedings of the 2013 International Conference on Signal Processing and Communication*, Noida, India, 2013, pp. 470-473.

[17] H. Yang, L. Qui, and S. N. Koh, "Application of instantaneous frequency estimation for fundamental frequency detection," *Proceedings of IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, 1994, pp. 616-619.