ICA 2019 AACHEN

PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Primary ambient extraction for random sign Hilbert filtering decorrelation

Lu CHEN; Chuang SHI; Huiyong LI

University of Electronic Science and Technology of China, Chengdu, China

Corresponding email: shichuang@uestc.edu.cn

ABSTRACT

The primary ambient extraction (PAE) decomposes a stereo mixture into separated primary and ambient components, providing a crucial step in spatial audio analysis and reproduction. The necessary spatial assumptions of PAE include (i) uncorrelated balanced ambient components, (ii) uncorrelated ambient and primary components, and (iii) primary components with the only difference in the panning factor. Hence, different PAE methods can be developed for different decorrelation processes that are used to produce the uncorrelated ambient components. This paper investigates a new decorrelation process, namely the random sign Hilbert filtering decorrelation process. This decorrelation process ensures that the ambient components consistently have ±90° phase difference between the left and right channels. The PAE method catering for the random sign Hilbert filtering decorrelation process is called the uncorrelated-ambient PAE (UAPAE). This paper expounds the UAPAE and demonstrate a comparison between the UAPAE and existing PAE methods.

Keywords: Primary ambient extraction, Decorrelation, Hilbert filtering

1. INTRODUCTION

Channel-based audio formats are widely used in consumer electronics, such as mobile phones, tablets and personal computers. Most of the channel-based audio files simply contain two channels that can be directly played back with headphones. In order for an immersive listening environment to be constructed with arbitrary playback systems, Goodwin and Jot proposed the idea of PAE that intended to change the number of channels without losing the spatial information (1).

Considering the audio signal as a linear combination of the primary and ambient components, there is a high possibility that the PAE would face an underdetermined problem, resulting in a necessity for three spatial assumptions (2). Firstly, the ambient components are assumed to be uncorrelated between any two channels and have equal energy. Secondly, the ambient components are assumed to be uncorrelated with the primary components. Thirdly, the primary components are almost identical. The only difference between any two primary components is just a panning factor.

When the primary components have higher energy than ambient components, principal component analysis (PCA) is dedicated to find the common unit vector of the primary components that maximizes the energy of the primary components (1). The result of PCA implies that the primary components are vertical to the ambient components. When ambient components are extracted from a stereo mixture, one ambient component pans to the opposite direction of the other, which doesn't obey with the assumptions that the ambient components are uncorrelated and balanced (3).

Ambient phase estimation with a sparsity constraint (APES) is suitable for the uncorrelated ambient components produced by the random phase decorrelation process (4). Magnitudes of ambient components are further assumed to be equal in every frequency bin. An additional sparse constraint is required to solve the underdetermined problem. APES achieves better accuracy when the primary components are dominant in the stereo mixture (5). The optimization that minimizes the norm of the primary components is non-convex in APES. Therefore, APES introduces an angle-by-angle search algorithm to find out the correct phases of ambient components. To reduce the computation complexity of APES, APEX has been proposed as a simplified algorithm with little loss in extraction accuracy (6).

Different PAE methods can be developed for different decorrelation processes that are used to produce the uncorrelated ambient components. The random sign Hilbert filtering decorrelation process ensures the ambient components have ±90° phase difference between the left and right channels.

Correlation coefficients between the processed ambient components are constantly nearly 0. In comparison, the random phase decorrelation process leads to correlation coefficients with more variations. Thus, this paper investigates the random sign Hilbert filtering decorrelation process and proposed the UAPAE method.

2. RANDOM SIGN HILBERT FILTERING DECORRELATION PROCESS

Highly correlated audio signals often lead to unnatural listening experience. When they are played back using headphones, listeners feel the sound artificially streaming inside of their heads (7, 8). When broadcasted in a reverberating room, highly correlated audio signals cause the combing phenomenon, which is destructive most of the time (9). The decorrelation process is a very important tool in audio engineering that can eliminate those unnatural listening sensations and the combing phenomenon.

The random phase decorrelation process can be carried out in either the time domain or frequency domain. Taking the frequency domain process as an example, the short time Fourier transform of the input signal $y_1(t)$ is firstly calculated. Secondly, a random phase shift is added into every time-frequency bin. Thirdly, the inverse Fourier transform is obtained to recover the output signal $y_2(t)$. The random sign Hilbert filtering decorrelation process is different from the random phase decorrelation process only in the second step, where $+90°$ or $-90°$ phase shift instead of the random phase shift is added.

To evaluate the performance of decorrelation processes, the correlation coefficient $\Omega(\Delta\mathrm{t})$ of two signals $y_1(t)$ and $y_2(t)$ with equal energy is defined as

$$\Omega(\Delta\mathrm{t}) = \left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_1(t)y_2(t+\Delta\, t)dt\right] \Big/ \left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_1(t)y_1(t)dt\right] \qquad (1)$$

or equivalently

$$\Omega(\Delta\mathrm{t}) = \left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_1(t)y_2(t+\Delta\, t)dt\right] \Big/ \left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_2(t)y_2(t)dt\right], \qquad (2)$$

where Δt represents a temporal offset.

Figure 1 presents correlation coefficients of the random sign Hilbert filtering and random phase decorrelation processes. It shows that correlation coefficients of the random sign Hilbert filtering decorrelation process are almost constantly 0, which are significantly smaller than those of the random phase decorrelation process.

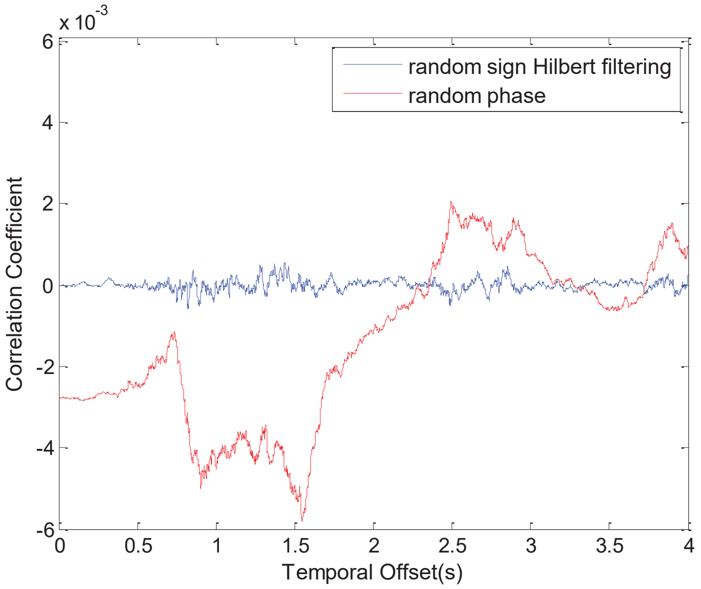

Figure 1 – Correlation coefficients of decorrelation processes

The combing phenomenon testing system is another approach to evaluate the effectiveness of decorrelation processes. A block diagram of the combing phenomenon testing system is shown in Figure 2.

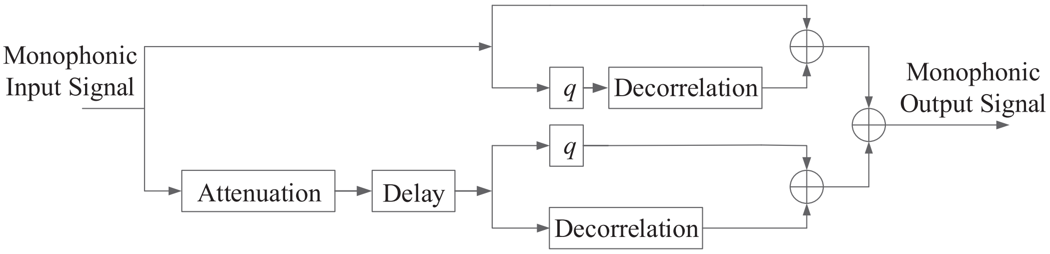

Figure 2 – Combing phenomenon testing system

Firstly, the factor q is set to 0. When there is no decorrelation process, the overall system behaves like a comb filter. Setting the attenuation gain to 0.7 and the delay amount to 0.04s leads to the result plotted in Figure 3(a). The magnitude response clearly demonstrates that the frequency interval between neighboring troughs is reciprocal of the delay amount. Figures 3(b) and 3(c) shows the results when the random sign Hilbert filtering and random phase decorrelation processes are taken into account. There is no obvious pattern in both magnitude responses. This validates that the random sign Hilbert filtering decorrelation process can make similar decorrelation effect as compared to the random phase decorrelation process.

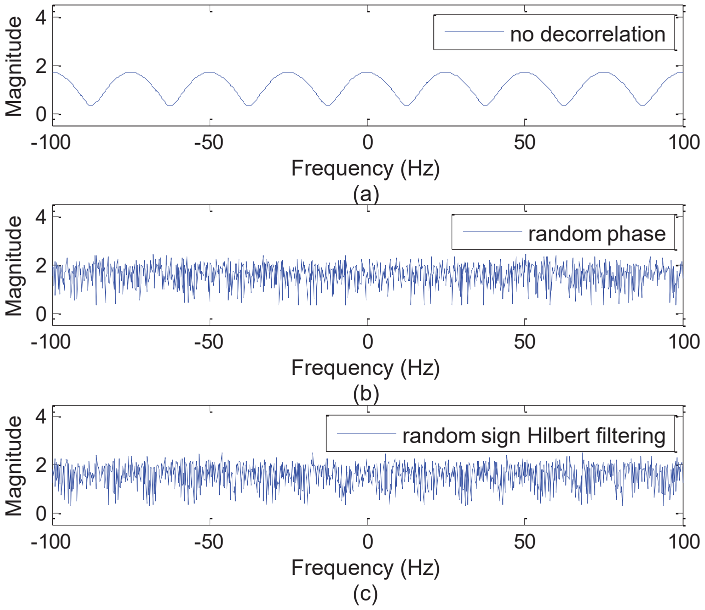

Figure 3 – Magnitude response of the combing phenomenon testing system when $q = 0$ and (a) no decorrelation; (b) the random phase decorrelation process; (c) the random sign Hilbert filter is carried out

Furthermore, setting q to 0.25, 0.5 and 1, the magnitude response when the random sign Hilbert filtering decorrelation process is considered are plotted in Figure 4. As q increases, the combing phenomena gradually recovers. The effect of the random sign Hilbert filtering decorrelation process is controllable by the factor q in the combing phenomenon testing system (4, 10).

3. UNCORRELATED-AMBIENT PAE

3.1 Description

In PAE, a stereo mixture is considered as a linear combination of primary and ambient components, which are written as

$$x_L(t) = p_L(t) + a_L(t) \tag{3}$$

and

$$x_R(t) = p_R(t) + a_R(t), \tag{4}$$

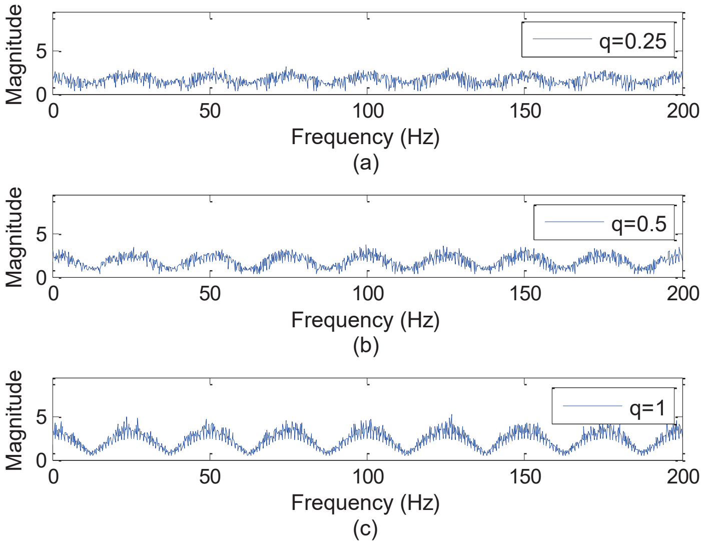

Figure 4 – Magnitude response of the combing phenomenon testing system when the random sign Hilbert filter is carried out and (a) $q = 0.25$; (b) $q = 0.5$; (c) $q = 1$, respectively

where x_L and x_R are the left and right channels of the stereo mixture. The primary components p_L and p_R are correlated and different only in a panning factor k as $p_R = kp_L$. The ambient components a_L and a_R are uncorrelated and have equal power. They are also uncorrelated with the primary components. These spatial assumptions are crucial in PAE.

After the short-time Fourier transform, equations (3) and (4) are rewritten as

$$X_L[m,f] = P_L[m,f] + A_L[m,f] \tag{5}$$

and

$$X_R[m,f] = P_R[m,f] + A_R[m,f], \tag{6}$$

where m is the frame index and f is the index of the frequency bin. PAE is thus carried out in every time-frequency bin. In the latter part of this paper, the notation $[m,f]$ is abbreviated unless necessary.

As shown in Figure 5, X_L, X_R, P_L, P_R, A_L, and A_R are all complex and can be represented as vectors in the complex plane. Denote the coordinate of X_L, X_R, A_L, and A_R as (x_1, y_1), (x_2, y_2), (a_1, b_1) and (a_2, b_2), respectively. The random sign Hilbert filtering decorrelation process suggests that $A_L \perp A_R$ and $|A_L| = |A_R|$, *i.e.*

$$a_1 a_2 + b_1 b_2 = 0 \tag{7}$$

and

$$a_1^2 + b_1^2 = a_2^2 + b_2^2. \tag{8}$$

Apparently, $a_2 = \mp b_1$ and $a_1 = \pm b_2$. The sign is associated with the random sign in the decorrelation process, which represent either A_L is behind A_R in phase by 90° or A_L is ahead A_R in phase by 90°

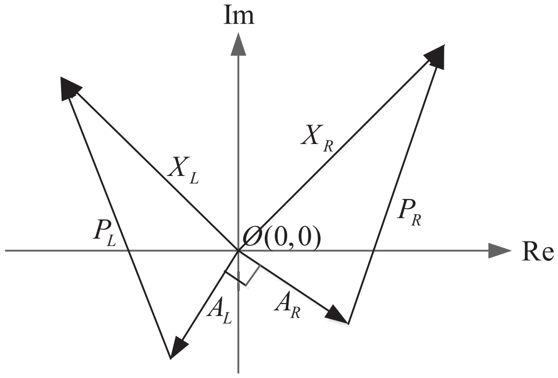

Figure 5 – PAE in one time-frequency bin illustrated on the complex plane

Since $P_R = kP_L$, knowing that

$$P_L = X_L - A_L = (x_1 - a_1, y_1 - b_1) \tag{9}$$

and

$$P_R = X_R - A_R = (x_2 - a_2, y_2 - b_2) \tag{10}$$

yields

$$(x_2 - a_2, y_2 - b_2) = k(x_1 - a_1, y_1 - b_1). \tag{11}$$

Given that $a_2 = -b_1$ and $a_1 = b_2$, equation (11) determines the ambient components by

$$\begin{cases} a_1 & = & (-ky_1 + k^2x_1 - kx_2 + y_2)/(k^2+1) \\ b_1 & = & (k^2y_1 + kx_1 - kx_2 + x_2)/(k^2+1) \\ a_2 & = & -(k^2y_1 + kx_1 - ky_2 - x_2)/(k^2+1) \\ b_2 & = & (-ky_1 + k^2x_1 - kx_2 + y_2)/(k^2+1) \end{cases}. \tag{12}$$

Alternatively, given that $a_2 = b_1$ and $a_1 = -b_2$, equation (11) determines the ambient components by

$$\begin{cases} a_1 & = & (ky_1 + k^2x_1 - kx_2 - y_2)/(k^2+1) \\ b_1 & = & (k^2y_1 - kx_1 - ky_2 + x_2)/(k^2+1) \\ a_2 & = & (k^2y_1 - kx_1 - ky_2 + x_2)/(k^2+1) \\ b_2 & = & -(ky_1 + k^2x_1 - kx_2 - y_2)/(k^2+1) \end{cases}. \tag{13}$$

After ambient components are determined, primary components are readily obtained by equations (9) and (10).

The panning factor k of a stereo mixture is estimated by

$$k = (\boldsymbol{x}_R^H\boldsymbol{x}_R - \boldsymbol{x}_L^H\boldsymbol{x}_L)/(2\boldsymbol{x}_L^H\boldsymbol{x}_R) + \sqrt{[(\boldsymbol{x}_R^H\boldsymbol{x}_R - \boldsymbol{x}_L^H\boldsymbol{x}_L)/(2\boldsymbol{x}_L^H\boldsymbol{x}_R)]^2 + 1} \tag{14}$$

where

$$\boldsymbol{x}_L = [x_L(t_0)\ x_L(t_1)\ x_L(t_2)\ \cdots x_L(t_N)] \tag{15}$$

and

$$\boldsymbol{x}_R = [x_R(t_0)\ x_R(t_1)\ x_R(t_2)\ \cdots x_R(t_N)] \tag{16}$$

are segments of the left and right channels.

However, there are two different solutions as provided by equations (12) and (13) in every time-frequency bin. It is necessary to select one of them. The rule of thumb used in the APES method is considered, *i.e.* the solution with the minimum magnitude or energy of $\hat{P}_L$ is selected. The flow of the UAPAE method is depicted in Figure 6

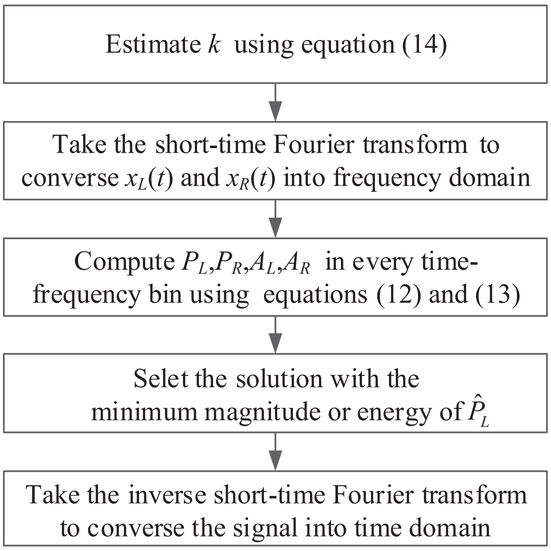
Estimate k using equation (14)
Take the short-time Fourier transform to converse $x_L(t)$ and $x_R(t)$ into frequency domain
Compute P_L,P_R,A_L,A_R in every time-frequency bin using equations (12) and (13)
Selet the solution with the minimum magnitude or energy of $\hat{P}_L$
Take the inverse short-time Fourier transform to converse the signal into time domain

Figure 6 – Flow chart of the UAPAE method

3.2 Experimental Results

In this experiment, a monophonic male speech is used as the primary component of the left channel. The panning factor k is set to 2, in order for the primary component of the right channel to be generated. A wave lapping sound recorded at the beach is selected as the ambient component of the left channel, which is also used as the input of the random sign Hilbert filtering decorrelation process to generate the ambient component of the right channel. When the primary and ambient components are mixed up, the power ratio of the primary components to the stereo mixture (abbreviated as PPR) is a useful measure to describe the relative energy of the primary and ambient components in the time domain. The audio files of the male speech and the wave lapping sound are extracted from (6).

The primary extraction error between the estimate of the primary component and the ground truth is computed by

$$err_p = \|p_L - \hat{p}_L\|_2^2/2\|p_L\|_2^2 + \|p_R - \hat{p}_R\|_2^2/2\|p_R\|_2^2. \tag{17}$$

Similarly, the ambient extraction error between the estimate of the ambient component and the ground truth is computed by

$$err_a = \|a_L - \hat{a}_L\|_2^2/2\|a_L\|_2^2 + \|a_R - \hat{a}_R\|_2^2/2\|a_R\|_2^2. \tag{18}$$

The primary and ambient extraction errors are plotted with respect to PPR in Figure 7, where the UAPAE method is compared with the APES, APEX, and PCA methods.

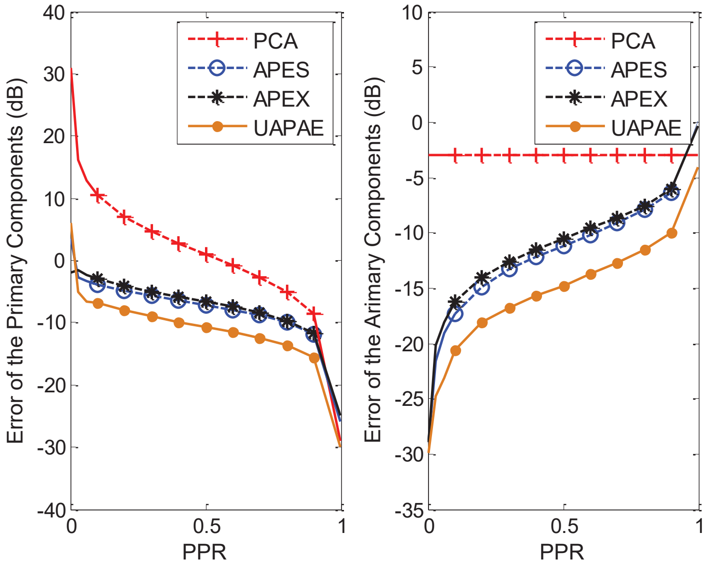

Figure 7 – Extraction error of various PAE methods with respect to PPR

Figure 7 shows that the primary extraction error decreases with PPR for every PAE methods and the ambient extraction error increases with PPR except for the PCA method. The UAPAE method achieves the lowest extraction error for both the primary and ambient components when the random sign Hilbert filtering decorrelation process has been taken.

Besides the extraction error, the time-domain segmental SNR (SNRseg),the frequency-weighted segmental SNR (fwSNRseg), the log likelihood ratio (LLR), perceptual evaluation of speech quality (PESQ) and perceptual evaluation of audio quality (PEAQ) are also calculated for comparison (11, 12).

The SNRseg is defined as

$$\mathrm{SNRseg} = \frac{10}{M}\sum_{m=0}^{M-1} \log_{10}\left[\sum\nolimits_{n=Nm}^{Nm+N-1} x^2(n)\right]\Big/\sum\nolimits_{n=Nm}^{Nm+N-1}[x(n)-\hat{x}(n)]^2, \tag{19}$$

where $x(n)$ is the ground truth of the primary or ambient components; $\hat{x}(n)$ is the estimate provided by PAE; N is the frame size and M is the number of frames. The frame size is set to 1323 in this experiment.

The fwSNRseg is defined as

$$\text{fwSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\left\{\sum_{f=1}^{L} w[m,f]\log_{10}\frac{\{X[m,f]\}^2}{\{X[m,f]-\hat{X}[m,f]\}^2}\right\} \Bigg/ \left[\sum_{f=1}^{L} W[m,f]\right], \tag{20}$$

where $X[m,f]$ and $\hat{X}[m,f]$ are the ground truth and estimate of the primary or ambient components; $L = 25$ is the number of critical bands; $W[m,f]$ is the energy of ground truth in the f-th critical band.

The linear predictive coding (LPC) represent the spectral envelope of a speech or an audio signal. By using LPC, LLR measures the spectral envelope difference between the ground truth and estimate of the primary or ambient components. It is defined as

$$d_{LLR} = \log\left[(\boldsymbol{a}_p\boldsymbol{R}_c\boldsymbol{a}_p^T)/(\boldsymbol{a}_c\boldsymbol{R}_c\boldsymbol{a}_c^T)\right] \tag{21}$$

where $\boldsymbol{a}_p$ is the LPC vector of the estimate; $\boldsymbol{a}_c$ and $\boldsymbol{R}_c$ are the LPC vector and the autocorrelation matrix of the ground truth, respectively. LLR of a clip is calculated frame by frame and bounded within 0 to 2. Only the lowest 95% LLR values are averaged to result in a single value.

The PESQ and PEAQ imitate the auditory system of human ear and predict the subjective score of a stimulus (13, 14).The range of PESQ is generally between 1 and 4.5 and a higher score represents better speech quality. The PEAQ uses the subjective difference grade (SDG) to represent audio quality. Although the value of SDG is from -4 to 4, only negative SDG values are obtained when the reference audio and the testing audio are distinguishable. In this experiment, PESQ is used to evaluate extracted primary components and PEAQ is used to score extracted ambient components.

When PPR is set to 0.8, evaluation results of various PAE methods are listed in Table 1. The best result in each category is in bold. The UAPAE method achieves the best performance, while the PCA method is worse than the other PAE methods.

Table 1 – Objective scores of primary and ambient components extracted with different PAE methods

		SNRseg (dB)	FwSNRseg (dB)	LLR	PESQ	PEAQ
$\hat{p}_L$	PCA	4.9855	6.4170	1.2766	2.0308	\
	APEX	5.4509	11.1996	0.3148	2.5319	\
	APES	5.5786	11.4584	0.2873	2.6277	\
	UAPAE	**8.3941**	**16.4009**	**0.2009**	**2.9283**	\
$\hat{p}_R$	PCA	4.9886	6.2481	1.2731	2.0255	\
	APEX	5.7545	11.3184	0.3261	2.5519	\
	APES	6.0492	12.3706	0.3039	2.6617	\
	UAPAE	**9.1755**	**16.9130**	**0.1968**	**2.9388**	\
$\hat{a}_L$	PCA	6.9556	18.4903	0.0602	\	-3.412
	APEX	12.2110	24.3520	0.0517	\	-2.593
	APES	12.8904	24.7844	0.0495	\	-2.551
	UAPAE	**16.1388**	**30.3395**	**0.0289**	\	**-1.094**
$\hat{a}_R$	PCA	0.9709	12.7951	0.1561	\	-3.713
	APEX	8.2178	22.1753	0.0456	\	-3.536
	APES	8.6797	22.6839	0.0464	\	-3.514
	UAPAE	**10.7547**	**27.1677**	**0.0185**	\	**-2.367**

4. CONCLUSIONS

PAE decomposes a stereo mixture into separated primary and ambient components and forms an underdetermined problem that always incur errors in the solution. Based on spatial assumption of PAE that the ambient components are uncorrelated, the random sign Hilbert filtering decorrelation process is investigated in this paper. It is relatively easy to implement and achieves similar decorrelation performance as compared to the random phase decorrelation process. To further support the random sign Hilbert filtering decorrelation process, this paper proposes the UAPAE method, which has been proved to outperform the other PAE methods in the experiment. The computation complexity of the UAPAE method is much lower than the APES method and with similar level of the frequency domain PCA and APEX methods.

ACKNOWLEDGEMENTS

This work is jointly supported by the National Science Foundation of China (Grant No. 61671137 and Grant No. 61701090) and Sichuan Science and Technology Program (Project No. 2018JY0218 and 2019YJ0182).

REFERENCES

1. Goodwin M, Jot JM. Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement. Proc ICASSP; 15-20 April 2007; Honolulu, HI, USA. p. 9–12.
2. Avendano C, Jot JM. A frequency-domain approach to multichannel upmix. J Audio Eng Soc. 2004; 52(7): 740-749.
3. He J, Gan WS, Tan EL. A study on the frequency-domain Primary-ambient extraction for stereo audio signals. Proc ICASSP; 04-09, May, 2014; Florence, Italy 2014.
4. Kendall G. The decorrelation of audio signals and its impact on spatial imagery. Computer Music Journal. 1995; 19(4): 71–87.
5. He J, Gan WS, Tan EL. Primary-ambient extraction using ambient phase estimation with a sparsity constraint. IEEE Sig Process Letters. 2015; 22(8):1127-1131.
6. He J. Spatial audio reproduction with primary ambient extraction. Singapore: Springer Publishing Company; 2016.
7. Xie BS. Head related transfer function and virtual auditory. Guangzhou, China: National Defense Industry Press; 2008.
8. Blauert J, Butler RA. Spatial hearing: The psychophysics of human sound localization by Jens Blauert. J Acoust Soc Am. 1985; 77(1): 334-335.
9. Bilsen FA. On the interaction of a sound with its repetitions. Waltman. 1968.
10. Xiang N, Xie B, Shi B. Decorrelating audio signals for stereophonic and surround sound using coded and maximum-length-class sequences. US Patent No. 9025776. 2015.
11. Hu Y, Loizou PC. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process. 2008;16(1): 229-238.
12. Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Am. 2009; 125(5): 3387-3405.
13. Beerends JG, Hekstra AP, Rix AW, et al. Perceptual evaluation of speech quality (PESQ) - The new ITU standard for end-to-end speech quality assessment - Part II - Psychoacoust model. J Audio Eng Soc. 2002; 50(10): 765-778.
14. ITU-R Recommendation BS.1387-1. Method for objective measurements of perceived audio quality. 2001.