Comments on "Primary-Ambient Extraction Using Ambient Spectrum Estimation for Immersive Spatial Audio Reproduction"

Haisheng Lu, Jiangnan Liang, and Chuang Shi

***Abstract*—In the above paper, He *et al.* propose a primary-ambient extraction method using ambient phase estimation with a sparsity constraint (APES). The primary-ambient extraction problem is formulated as a $L1$-norm minimization of the primary component with respect to the phase of the ambient component, based on a stereo signal model they have assumed. Discrete searching (DS) is proposed as the approach for solving the APES in each time-frequency bin, which is computationally complicated and potentially imprecise. This correspondence provides an analytical solution to the APES by exploring the geometric relation between the primary and ambient components in a rotated coordinate.**

***Index Terms*—Primary-ambient extraction, ambient phase estimation with a sparsity constraint, analytical solution.**

I. Introduction

Primary-ambient extraction of stereo content is an important preprocessing task in spatial audio reproduction. The stereo signal model presented in [1] assumes that primary components are only different in amplitude by a constant panning factor k, while ambient components are uncorrelated and have equal magnitude. The primary-ambient extraction problem is firstly converted to ambient phase estimation by eliminating the primary component in the right channel. Given that the ambient magnitude is real and non-negative, the relation between the phases of the two ambient components is derived. A sparsity constraint is introduced to ensure a determined result. Hence, the primary-ambient extraction problem is formulated as a $L1$-norm minimization of the primary component with respect to the phase of the ambient component. As mentioned in [2], the magnitude of the primary component at one time-frequency bin is solely determined by the phase of the ambient component in the same time-frequency bin. DS is suggested to minimize the magnitude of the primary component independently for each time-frequency bin, despite it being computationally complicated and potentially imprecise. Therefore, this correspondence provides an analytical solution to the APES by exploring the geometric relation between the primary and ambient components in a rotated coordinate.

II. Ambient Phase Estimation with a Sparsity Constraint

In [1], denoting the short-time Fourier transform of the stereo signal at time index n and frequency bin index l as $X_c(n,l)$, where $c \in \{0,1\}$ is the channel index, the stereo signal model is written as

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

where $\boldsymbol{X}_c[n,b]$ consists of a column of time-frequency bins of the stereo signal at time index n within a given subband b; $\boldsymbol{P}_c[n,b]$ and $\boldsymbol{A}_c[n,b]$ are the primary and ambient components, respectively. The indices $[n,b]$ are often omitted for brevity in discussions about APES. It is also assumed that $\boldsymbol{P}_1 = k\boldsymbol{P}_0$, $k \geq 1$, and $|\boldsymbol{A}_1| = |\boldsymbol{A}_0|$.

The APES formulates the primary-ambient extraction problem as

$$\hat{\boldsymbol{\theta}}_1^* = \arg\min_{\hat{\boldsymbol{\theta}}_1} \|\hat{\boldsymbol{P}}_1\|_1, \tag{2}$$

Haisheng Lu and Jiangnan Liang are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Chuang Shi is with the Institute of Sound and Vibration Research, University of Southampton, Southampton, UK.

Corresponding author's email: chuang.shi@soton.ac.uk

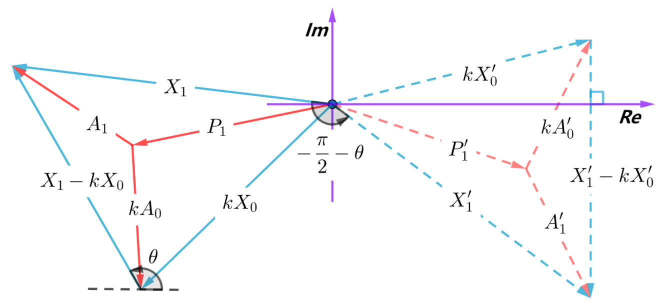

Fig. 1. Geometric illustration of (4).

where $\hat{\boldsymbol{P}}_1$ is the primary component subject to a possible phase of the ambient component $\hat{\boldsymbol{\theta}}_1$. $\hat{\boldsymbol{\theta}}_1^*$ is the optimum phase of the ambient component for the APES. When DS is applied, $\hat{\boldsymbol{\theta}}_1$ traverses a linearly spaced vector bounded from $-\pi$ to π.

III. Analytical Solution to the APES

Similar to [2], the stereo signal model in (1) is rewritten for each time-frequency bin, with the indices (n,l) omitted, as

$$X_1 = P_1 + A_1,\ kX_0 = kP_0 + kA_0. \tag{3}$$

When $X_1 = kX_0$, the solution to the APES is straightforwardly provided by $P_c = X_c$ and $A_c = 0$.

When $X_1 \neq kX_0$, let

$$X_c' = X_c e^{j\left(-\frac{\pi}{2}-\theta\right)} = P_c e^{j\left(-\frac{\pi}{2}-\theta\right)} + A_c e^{j\left(-\frac{\pi}{2}-\theta\right)}, \tag{4}$$

where θ is the phase of $X_1 - kX_0$ and $j = \sqrt{-1}$. The rotation in (4) ensures that $X_1' - kX_0'$ is purely imaginary, as shown in Fig. 1.

Denote $P_c' = P_c e^{j\left(-\frac{\pi}{2}-\theta\right)}$ and $A_c' = A_c e^{j\left(-\frac{\pi}{2}-\theta\right)}$. Therefore, $X_c' = P_c' + A_c'$ and

$$k|X_1' - P_1'| = |kX_0' - P_1'|. \tag{5}$$

Supposing that $P_1' = x + jy$, $kX_0' = x_0 + jy_0$ and $X_1' = x_0 + jy_1$, the locus of all the possible $\hat{P}_1'$ can be derived from (5) as

$$\begin{aligned}(k^2-1)(x-x_0)^2 + (k^2-1)y^2 - \\ 2(k^2y_1 - y_0)y + k^2y_1^2 - y_0^2 = 0.\end{aligned} \tag{6}$$

A. When $k > 1$

We can divide both sides of (6) by the non-zero factor $k^2 - 1$, yielding a circle formula:

$$(x-x_0)^2 + \left(y - \frac{k^2y_1 - y_0}{k^2-1}\right)^2 = \frac{k^2}{(k^2-1)^2}(y_0 - y_1)^2, \tag{7}$$

of which the radius is written as

$$r = \frac{k}{k^2-1}|y_0 - y_1|. \tag{8}$$

The distance from the origin to the center of the circle is given by

$$d = \sqrt{x_0^2 + (\frac{k^2y_1 - y_0}{k^2-1})^2}. \tag{9}$$

When $d > r$, the origin is outside the circle. Otherwise, the origin is on or inside the circle. Nevertheless, the range of $|\hat{P}_1'|$ must satisfy

$$|r - d| \leq |\hat{P}_1'| \leq d + r, \tag{10}$$

and the lower bound is achieved by

$$\hat{P}_1'^* = \left(x_0 + j\frac{k^2y_1 - y_0}{k^2-1}\right)\frac{d-r}{d}. \tag{11}$$

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2023.3328286

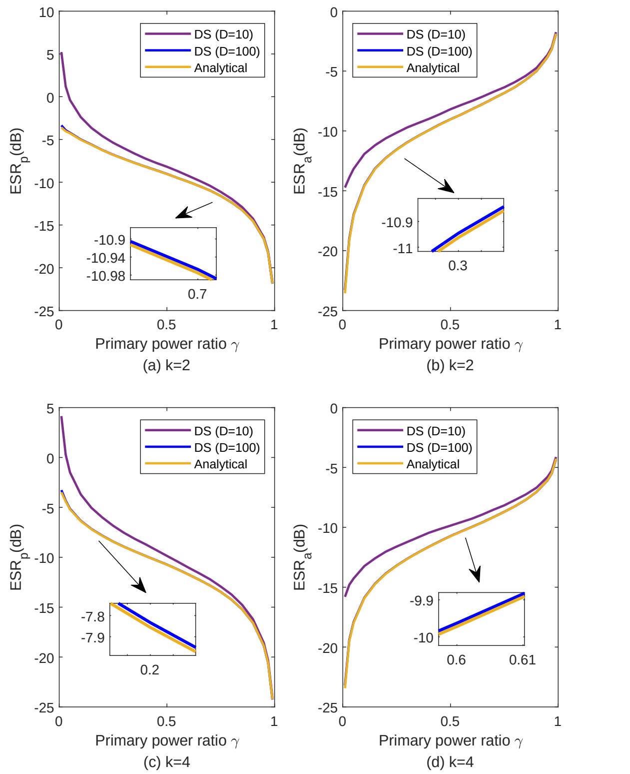

Fig. 2. ESR of the primary and ambient components extracted by the analytical solution and DS when $k = 2$ and $k = 4$.

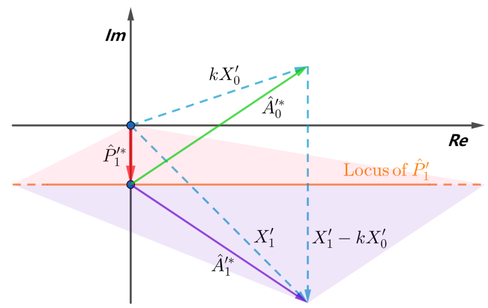

Fig. 3. Locus of all possible $\hat{P}_1'$ and corresponding analytical solution when $k = 1$.

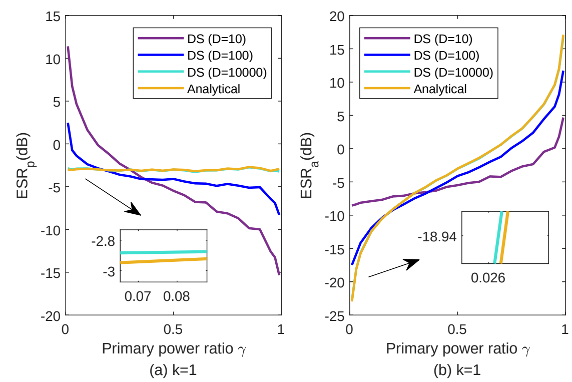

Fig. 4. ESR of the primary and ambient components extracted by the analytical solution and DS when $k = 1$.

Upon $\hat{P}_1'^*$ is determined, the analytical solution to the APES can be obtained by

$$\hat{P}_1^* = \hat{P}_1'^* e^{j\left(\frac{\pi}{2}+\theta\right)},\ \hat{P}_0^* = k^{-1}\hat{P}_1^*,\ \hat{A}_c^* = X_c - \hat{P}_c^*. \tag{12}$$

In Fig. 2, the error-to-signal ratio (ESR) results in [1] and [2] for $k = 2$ and $k = 4$ are replicated and compared with the results of the analytical solution. The primary power ratio is defined as the ratio of the total power of the primary component to the total signal power in two channels [1]. The analytical solution achieves more accurate primary-ambient extraction than DS, while reducing the computational complexity from $O(D)$ to $O(1)$. Here, D denotes the number of search points in DS.

B. When $k = 1$

The locus of all the possible $\hat{P}_1'$ becomes a line on the complex plane (see Fig. 3), of which the expression is given by

$$y = \frac{y_0 + y_1}{2}. \tag{13}$$

In this case, the range of $|\hat{P}_1'|$ satisfies

$$|\frac{y_1 + y_0}{2}| \leq |\hat{P}_1'|. \tag{14}$$

Taking the lower bound of the above range, the analytical solution to the APES is obtained from

$$\hat{P}_1'^* = j\frac{y_1 + y_0}{2}. \tag{15}$$

It should be noted in Fig. 3 that $\hat{\theta}_1'$, *i.e.* the phase of $\hat{A}_1'$, can only be found from $-\pi$ to 0, since the locus is located in the lower half-plane. Moreover, because the phase of $\hat{A}_1$ is given by $\hat{\theta}_1 = \hat{\theta}_1' + \frac{\pi}{2} + \theta$, the search range of $\hat{\theta}_1$ should not be set to from $-\pi$ to π when $k = 1$. Although this issue has been overlooked previously, it does not significantly affect the results in most of the time-frequency bins. A more critical issue that demands attention is that a large number of search points is needed, since the locus of all the possible $\hat{P}_1'$ forms a line.

In Fig. 4, the ESR results in [1] and [2] for $k = 1$ are replicated and compared with the results of the analytical solution. Only when D is extremely large, the results of DS closely approximate the results of the analytical solution. This occurs because, when D is small, $\hat{P}_1'^*$ is skipped over during the process of DS, leading to only a local minimum of the $L1$-norm of the primary component.

IV. Conclusion

The APES, proposed in [1], formulates the primary-ambient extraction problem by minimizing the $L1$-norm of the primary component with respect to the phase of the ambient component. Consequently, DS has been employed to solve the APES in each time-frequency bin. This correspondence highlights that when $k > 1$, DS encounters high computational complexity and can be completely replaced by the analytical solution proposed in Section III. However, when $k = 1$, DS tends to yield imprecise results. In this case, the analytical solution reveals that the sparsity constraint does not perform effectively. Therefore, future research should explore alternative approaches, such as modifying (2) in a statistical or data-driven manner.

References

[1] J. He, W. S. Gan, and E. L. 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, 2015.

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