Fast Non-Uniform Searching Strategy for Ambient Phase Estimation in Stereo Recordings with Sparse Primary Components

Chuang Shi[a,*], Jiangnan Liang[a], Yingzi Liu[a], Haocong Yang[a]

[a] *School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China*

Abstract

A stereo recording can be considered to consist of primary and ambient components in both left and right channels. Decomposing a stereo recording into primary and ambient components is a crucial step in upmixing to retain its spatial information. The state-of-the-art algorithm to carry out the primary ambient extraction (PAE), namely the ambient phase estimation with a sparsity constraint (APES), utilized the sparsity of the primary components in the time-frequency domain to formulate the ambient phase estimation as a non-convex optimization problem. Hence, the discrete searching method was adopted, resulting in a computationally complicated solving process. In this paper, a fast non-uniform searching strategy is proposed to improve the efficiency of the APES, resulting in the ambient phase estimation with a sparsity constraint and non-uniform searching (APEN). Objective and subjective results validate that the extraction error of the APEN is almost the same as that of the APES, while the algorithm complexity of the APEN is reduced to one third that of the APES.

Keywords: Primary ambient extraction, stereo recording, ambient phase estimation, sparsity constraint, non-uniform searching

[*] Corresponding Author's Email: shichuang@uestc.edu.cn

1. Introduction

Spatial audio is an important technology in virtual reality, immersive communication, interactive media, and the emerging metaverse [1, 2, 3, 4]. There are two distinct approaches of creating spatial sound experience [5]. One approach's emphasis is on the manipulation of sound pressures at the listener's ears [6]. The sound transmitted from a point source is scattered by the head, torso, and pinnae of the listener before entering both ears. Spectral cues, such as interaural time difference (ITD) and interaural level difference (ILD) are hence incurred in the binaural sound pressures [7]. They are essential for the auditory system to localize the point source [8]. The relation between the sound pressure of the point source and the binaural sound pressures can be further described by head related-transfer functions (HRTFs) in the frequency domain [9]. For the same listener, the exact reproduction of the binaural sound pressures should lead to identical spatial sound experience when there is no head movement [10]. Considering the difference among individuals, customized HRTF processing is likely to be the next key procedure [11].

The other approach of generating spatial sound is to apply a multi-channel system that can reconstruct a sound field approximating that of the original recording, even though this is often only feasible within certain "sweet spots" [12]. The multi-channel sound systems are continuously evolving in consumer electronics. Presently, the International Telecommunication Union (ITU) recommends the 5.1 channel surround sound system as the standard for multichannel stereophonic sound systems [13]. The 5.1 channel system represents a satisfactory compromise between system complexity and desired spatial sound experience. It employs three front loudspeakers preserving only the approximate directional information in the horizontal plane, two surround loudspeakers to play back the ambient sound, and one optional loudspeaker to reproduce the low frequency range of 20 to 120 Hz. With more complicated configurations, gaming headphones have been adopting 7.1 channel systems [14]. According to the rule of summing localization, projecting a virtual sound source at an arbitrary direction can be achieved by adjusting the relative amplitudes of multi-channel signals in an appropriate configuration of loudspeakers [15]. The 22.2 channel system has been promoted during the Tokyo 2020 Summer Olympics [16]. As compared to the 5.1 channel system, the 22.2 channel system provides the description of the up and down movement of sound and the sense of height by adding in an upper layer of channels. More channels are expected to produce higher spatial fidelity, but they also result in a more complicated system. The need to simplify the system often conflicts with the wish for better reproduction outcome. In the development of multi-channel sound systems, the number of required channels has been a debatable issue for some time [17].

However, most of the recordings till date have been carried out using two channels, which are also called the stereo recordings. Playing the channel-based audio requires the configuration of channels in advance [18]. Thus, the channel-based audio format often needs adaptability of audio playback systems by upmixing and downmixing [19]. In the upmixing process, surround channels can be created from stereo recordings using two methods. One method uses the decorrelating components of the stereo recording as the surround channels, whereas the other method generates the surround channels by simulating the reverberant sound field [20, 21]. In contrast to the upmixing process, the downmixing process is employed to cater for a decreased number of channels due to practical reasons such as availability of loudspeakers. Downmixing can be accomplished by simple mixing or more complex HRTF processing [22]. Upmixing and downmixing can address the mismatch of the number of channels, yet the spatiality of the sound should be dealt with cautiously [23].

As stereo recordings are still the most preferred choice for audio data storage, upmixing from the stereo recording is often the main research focus, in which the primary ambient extraction (PAE) is one of the most representative methods. PAE regards the audio scene as a linear combination of foreground and background sound, which are referred to as the primary and ambient components, respectively [24]. Till now, the PAE has several implementation frameworks in literature. Many of them have been developed based on the linear estimation framework, whereby the extracted components are weighted sums of the stereo recordings [25]. Faller proposed the least squares (LS) algorithm to estimate the primary and ambient components of stereo signals by minimizing the estimation error in the linear estimation framework [26]. Goodwin and Jot proposed the PAE algorithm based on the principal component analysis (PCA), which was dedicated to finding a common basis vector of the primary components in either the time domain or the time-frequency domain [27]. He *et al.* proposed the minimum leakage least squares (MLLS) and the minimum distortion least squares (MDLS) algorithms [28]. The MLLS and MDLS redefine the extraction error as a combination of distortion, interference and leakage, and then minimize the extraction error under the premise of minimum leakage and minimum distortion, respectively. Avendano and Jot proposed a time-frequency domain masking algorithm that judged whether a time-frequency bin belonged to the primary or ambient component based on the cross-channel correlation [29].

In addition, PAE may also be developed based on the ambient spectrum estimation framework, which has been validated to achieve higher accuracy than the linear estimation framework. He *et al.* proposed the ambient phase estimation with a sparsity constraint (APES) algorithm, in which the primary components were presumed to distribute sparsely in the time-frequency domain. [30]. The APES adopts an angle-by-angle searching strategy to try out all the possible phase angles of ambient components, among which the final solution is picked up to maximize the sparsity of the extracted primary component. This sparsity constraint can be adjusted based on prior information about the stereo recordings and leads to specific improvements to the original APES [31, 32]. The angular resolution in the solving process is significant to the extraction error of the APES and its variants. Higher angular resolution results in less extraction error, but increases the algorithm complexity. In order to simplify the solving process, the APEX algorithm was introduced along with the APES. The APEX takes the phase angle of one channel in the stereo recording straightforwardly as the estimated phase angle of the ambient component. By doing so, the APEX can hasten the solving process, with acceptable accuracy degradation in several exemplified circumstances [33].

In this paper, a fast non-uniform searching strategy is proposed to accelerate the solving process of the ambient phase estimation, resulting in the ambient phase estimation with a sparsity constraint and non-uniform searching (APEN). Differing from the APES that adopts a constantly high angular resolution in the full range of searching, the APEN combines a fine searching in a small range with high confidence and a fast searching in a large range with low confidence. The APEN is several times faster than the APES, while providing equivalent accuracy to the APES in objective and subjective tests.

2. Ambient phase estimation with a sparsity constraint and non-uniform searching

The PAE decomposes a stereo recording into primary and ambient components, which is a key step in upmixing while keeping the spatial information. Such decomposition requires several assumptions to make it a determined mathematical problem to solve [34]. Typically, the primary components in the left and right channels are assumed to be differentiated by just a panning factor. The ambient components in the left and right channels are assumed to have equal magnitudes but random phases. Therefore, in every channel of the stereo recording, there is a primary component and an ambient component, which are written as

$$\mathbf{x}_c(n) = \boldsymbol{p}_c(n) + \boldsymbol{a}_c(n), \quad \forall c \in \{0, 1\}, \tag{1}$$

where $c \in \{0, 1\}$ is the channel index and the notation (n) is omitted for brevity in the latter. The primary components $\boldsymbol{p}_c$ are assumed to be correlated across channels and only differentiated in the amplitude with a panning factor k, *i.e.* $\boldsymbol{p}_1 = k\boldsymbol{p}_0$. The ambient components $\boldsymbol{a}_c$ are assumed to have the same energy but to be uncorrelated to each other. Each ambient component should ideally be uncorrelated with every primary component. However, taking the PCA algorithm as an example, the extracted components are expressed as

$$\hat{\boldsymbol{p}}_1 = k\hat{\boldsymbol{p}}_0 = \frac{k}{1+k^2}\left(\boldsymbol{x}_0 + k\boldsymbol{x}_1\right) \tag{2}$$

and

$$\hat{\boldsymbol{a}}_0 = -k\hat{\boldsymbol{a}}_1 = \frac{k}{1+k^2}\left(k\boldsymbol{x}_0 - \boldsymbol{x}_1\right), \tag{3}$$

where the extracted ambient components are correlated with the extracted primary components. This is a common problem faced by the linear estimation framework.

In the ambient spectrum estimation framework, the left and right channels of the stereo recording are rewritten in the time-frequency domain as

$$\mathbf{X}_c(m,f) = \mathbf{P}_c(m,f) + \mathbf{A}_c(m,f), \tag{4}$$

where m is the index of frame and f is the index of frequency bin. The spectra of ambient components are further expressed by their magnitudes and phase angles as

$$\mathbf{A}_c(m,f) = \left|\mathbf{A}(m,f)\right| e^{j\boldsymbol{\theta}_c(m,f)}, \tag{5}$$

where $\boldsymbol{\theta}_c(m,f) = \angle\mathbf{A}_c(m,f)$ is the phase angle of $\mathbf{A}_c(m,f)$. The notations (m,f) are omitted for brevity in the latter part of this paper.

The assumptions between primary components in the time-frequency domain remains to be

$$\mathbf{P}_1 = k\mathbf{P}_0. \tag{6}$$

In this case, the panning factor can be accurately estimated from a stereo recording by

$$\hat{k} = \frac{\boldsymbol{X}_1^H\boldsymbol{X}_1 - \boldsymbol{X}_0^H\boldsymbol{X}_0}{2\boldsymbol{X}_1^H\boldsymbol{X}_0} + \sqrt{\left(\frac{\boldsymbol{X}_1^H\boldsymbol{X}_1 - \boldsymbol{X}_0^H\boldsymbol{X}_0}{2\boldsymbol{X}_1^H\boldsymbol{X}_0}\right)^2 + 1}. \tag{7}$$

Substituting (5), (6) into (4) and eliminating primary components yield

$$|\mathbf{A}| = \frac{\mathbf{X}_1 - k\mathbf{X}_0}{e^{j\boldsymbol{\theta}_1} - ke^{j\boldsymbol{\theta}_0}}. \tag{8}$$

Let $\boldsymbol{\theta} = \angle\left(\mathbf{X}_1 - k\mathbf{X}_0\right)$, which can be directly calculated from the stereo recording. Since $|\mathbf{A}|$ is real, $\tan\boldsymbol{\theta} = (\sin\boldsymbol{\theta}_1 - k\sin\boldsymbol{\theta}_0)/(\cos\boldsymbol{\theta}_1 - k\cos\boldsymbol{\theta}_0)$ must hold. This relationship between the phase angles of the ambient components is further manipulated as

$$\sin\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) = k^{-1}\sin\left(\boldsymbol{\theta} - \boldsymbol{\theta}_1\right), \tag{9}$$

which indicates that only one phase angle $\boldsymbol{\theta}_1$ needs to be estimated in the ambient spectrum estimation framework.

For the APES, with a sparsity constraint, the PAE is transformed into a non-convex optimization problem, whereby the ambient phase is estimated as

$$\boldsymbol{\theta}_1^* = \arg\min_{\hat{\boldsymbol{\theta}}_1} |\hat{\mathbf{P}}_1|, \tag{10}$$

where

$$\hat{\mathbf{P}}_1 = k\hat{\mathbf{P}}_0 = k\frac{\mathbf{X}_0 e^{j\hat{\boldsymbol{\theta}}_1}}{e^{j\hat{\boldsymbol{\theta}}_1} - ke^{j\hat{\boldsymbol{\theta}}_0}} \tag{11}$$

and

$$\sin\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0\right) = k^{-1}\sin\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_1\right). \tag{12}$$

The standard implementation of the APES adopts a discrete searching method with a uniform searching strategy, which is tedious and extremely time-consuming [30]. Instead, an approximate solution to (10) can be analytically obtained, which is also known as the APEX. For $k > 1$, the phase angle of $\mathbf{X}_1$ is straightforwardly used as the estimate of the ambient phase, *i.e.*

$$\hat{\boldsymbol{\theta}}_1^* = \angle\mathbf{X}_1. \tag{13}$$

However, for $k = 1$, the APEX employs a different approximation that is given by

$$\hat{\boldsymbol{\theta}}_1^* = \angle\left(\mathbf{X}_1 - \mathbf{X}_0\right) = \boldsymbol{\theta}. \tag{14}$$

The APEX often causes a notable loss of accuracy. Therefore, a fast non-uniform searching strategy is proposed in this paper to accelerate the solving process of the ambient phase estimation, resulting in the APEN. Differing from the APES that adopts a constantly high angular resolution in the full range of searching, the APEN combines a fine searching within a small range and a fast searching within remaining ranges.

The fine searching range is centered with the approximate solution to (10) that is provided by the APEX. Therefore, the fine searching range of the APEN is written as $[\hat{\boldsymbol{\theta}}_1^* - \beta, \hat{\boldsymbol{\theta}}_1^* + \beta]$, while the fast searching range is given by $[\hat{\boldsymbol{\theta}}_1^* +$

$\beta, \hat{\boldsymbol{\theta}}_1^* - \beta + 2\pi]$. Here, β is a parameter to set the boundary of the fine searching range, which is determined by

$$\beta = E[\Delta\boldsymbol{\theta}] + \sigma \cdot D[\Delta\boldsymbol{\theta}], \tag{15}$$

where $\Delta\boldsymbol{\theta}$ denotes the phase difference between the extracted ambient components of the APEX and the APES; $E[\cdot]$ and $D[\cdot]$ denote the mean and variance of a random variable, respectively; σ is a factor to adjust the confidence interval of $\Delta\boldsymbol{\theta}$ with respect to its variance.

According to (13) and (14), the phase difference between the extracted ambient components of the APEX and the APES are written as

$$\Delta\boldsymbol{\theta} = \angle\mathbf{X}_1 - \boldsymbol{\theta}_1^* = \cos^{-1}\left(\frac{|\mathbf{A}_1|}{|\mathbf{X}_1|}\cos\boldsymbol{\theta}_1 + \frac{|\mathbf{P}_1|}{|\mathbf{X}_1|}\cos\boldsymbol{\theta}_p\right) - \boldsymbol{\theta}_1^* \tag{16}$$

and

$$\Delta\boldsymbol{\theta} = \angle\left(\mathbf{X}_1 - \mathbf{X}_0\right) - \boldsymbol{\theta}_1^* = \boldsymbol{\theta} - \boldsymbol{\theta}_1^* \tag{17}$$

for $k > 1$ and $k = 1$, respectively. Here, $\boldsymbol{\theta}_p$ denotes the phase angle of $\mathbf{P}_1$. Since $\boldsymbol{\theta}_1$ is uniformly distributed in $[-\pi, \pi]$, $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_p$ are assumed to be distributed in the same way. Therefore, the mean of $\Delta\boldsymbol{\theta}$ is estimated to be 0 in both (16) and (17).

The variance of $\Delta\boldsymbol{\theta}$ is relatively simpler to estimate for $k = 1$, since $\boldsymbol{\theta}$ can be straightforwardly obtained from the stereo recording. Therefore, the variance of $\Delta\boldsymbol{\theta}$ is given by

$$D[\Delta\boldsymbol{\theta}] = D[\boldsymbol{\theta}] + \frac{4\pi^2}{3}, \tag{18}$$

for $k = 1$.

For $k > 1$, as the primary components are sparsely distributed in the time-frequency domain, the probability of a negligible primary component is defined as

$$\alpha = P(1 - \varepsilon < \frac{|\mathbf{X}_1|}{|\mathbf{X}_0|} < 1 + \varepsilon), \tag{19}$$

where ε is a small threshold. Since α is only related to $\mathbf{X}_0$ and $\mathbf{X}_1$, it can be straightforwardly obtained from the stereo recording. The ratio of $|\mathbf{X}_1|$ to $|\mathbf{X}_0|$ can be equivalently viewed as $|\mathbf{P}_1 + \mathbf{A}_1|/|\mathbf{P}_0 + \mathbf{A}_0|$. It approaches 1, when the primary components approach 0. At the same time, $|\mathbf{A}_1|/|\mathbf{X}_1|$ also approaches 1, and therefore, $\Delta\boldsymbol{\theta}$ becomes 0 in (16).

When the primary component is not negligible, the variances of $|\mathbf{A}_1|/|\mathbf{X}_1|$ and $|\mathbf{P}_1|/|\mathbf{X}_1|$ are respectively estimated by

$$D(\frac{|\mathbf{A}_1|}{|\mathbf{X}_1|}) = \frac{\sqrt{\frac{(1-PPR)(1+k^{-2})}{2PPR+(1-PPR)(1+k^{-2})}} - \alpha}{1-\alpha} \tag{20}$$

and

$$D(\frac{|\mathbf{P}_1|}{|\mathbf{X}_1|}) = \frac{\sqrt{\frac{2PPR}{2PPR+(1-PPR)(1+k^{-2})}}}{\alpha} \tag{21}$$

where Poisson distributions are assumed for $|\mathbf{A}_1|/|\mathbf{X}_1|$ and $|\mathbf{P}_1|/|\mathbf{X}_1|$ in order for estimates of variances to be simplified to estimates of means; and the primary power ratio (PPR) calculates the proportion of the primary component energy in the total energy of the stereo recording, *i.e.*

$$PPR = \frac{\|\mathbf{P}_0\|_2^2 + \|\mathbf{P}_1\|_2^2}{\|\mathbf{X}_0\|_2^2 + \|\mathbf{X}_1\|_2^2} = \frac{\left(1+k^{-2}\right)\|\mathbf{P}_1\|_2^2}{(1+k^{-2})\|\mathbf{P}_1\|_2^2 + 2\|\mathbf{A}_1\|_2^2}. \tag{22}$$

Furthermore, as $\cos^{-1} x \approx \frac{\pi}{2} - \frac{\pi}{2}x$ for $-1 < x < 1$, (16) can be approximated by

$$\Delta\boldsymbol{\theta} = \frac{\pi}{2} - \frac{\pi}{2}\left(\frac{|\mathbf{A}_1|}{|\mathbf{X}_1|}\cos\boldsymbol{\theta}_1 + \frac{|\mathbf{P}_1|}{|\mathbf{X}_1|}\cos\boldsymbol{\theta}_p\right) - \boldsymbol{\theta}_1^*. \tag{23}$$

Since $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_p$ are treated as independent uniform distributions in $[-\pi, \pi]$, the variance of $\Delta\boldsymbol{\theta}$ can be finally estimated for $k > 1$ as

$$D[\Delta\boldsymbol{\theta}] = \frac{\pi^2}{8}D[\frac{|\mathbf{A}_1|}{|\mathbf{X}_1|}] + \frac{\pi^2}{8}D[\frac{|\mathbf{P}_1|}{|\mathbf{X}_1|}] + \frac{4\pi^2}{3}. \tag{24}$$

Substituting (24) and (18) into (15) results in the boundaries of the fine searching range of the APEN for $k > 1$ and $k = 1$, respectively.

3. Performance evaluation

In this section, three combinations of primary and ambient sound clips are employed for performance evaluation, despite previous practices using only one combination to demonstrate the effectiveness of the APES and the APEX [30]. The first combination includes a male voice and a wave lapping sound as the

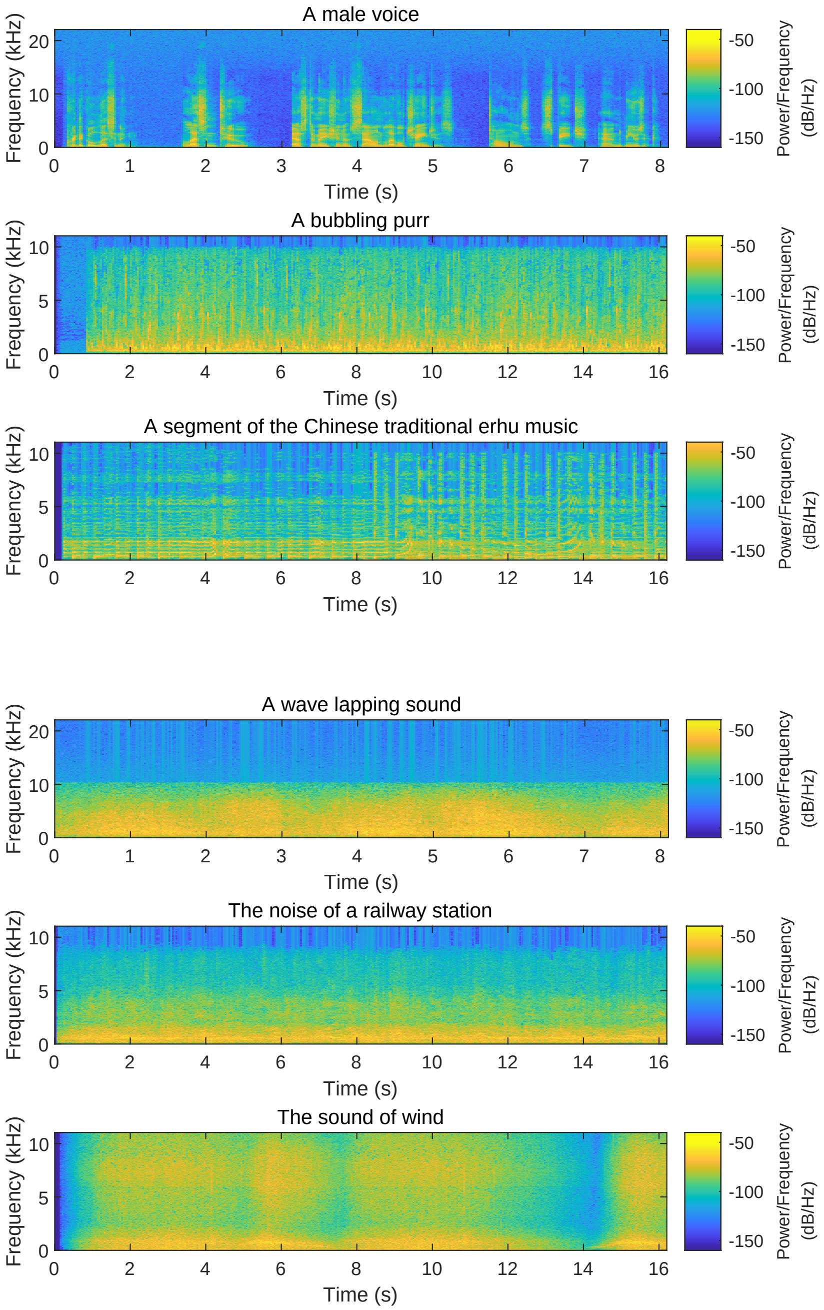


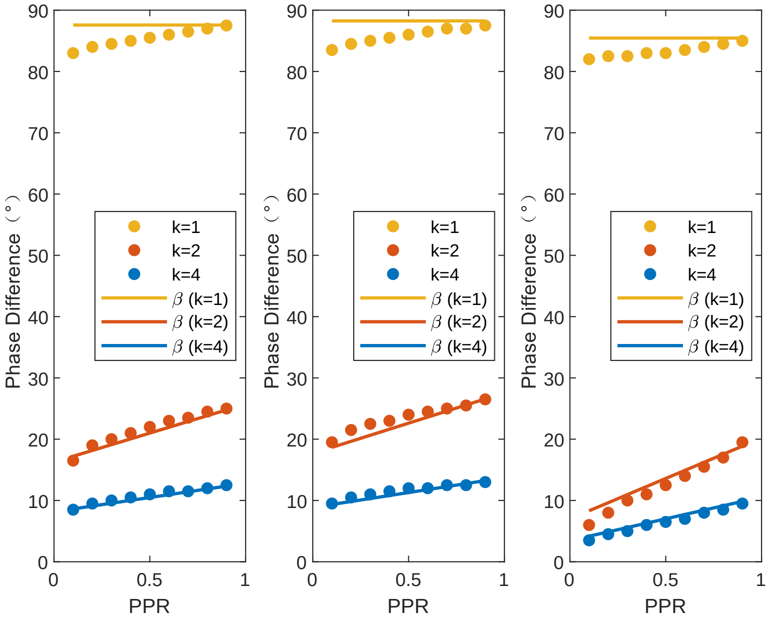
Figure 1: Spectrograms of the primary components and ambient sound clips used for performance evaluation.

sources of the primary and ambient components, respectively. It is provided in [22] and referred to as "speech+wave" in the latter part of this paper. The second combination comprises of a bubbling purr and the noise of a railway station, and is thus referred to as "bubble+station". The third combination is referred to as "erhu+wind", which consists of a segment of Chinese traditional erhu music and the sound of wind. The sound clips of the first combination are sampled at 44.1 kHz, while the sound clips of the other two combinations are sampled at 22.05 kHz. All the sound samples are processed with a frame size of 4096 samples. A random phase filter is implemented to process one ambient source into two uncorrelated ambient components appearing in the left and right channels.

Fig. 1 shows the spectrograms of the primary components and ambient sound clips used for performance evaluation. The three sources of the primary component present a variety of time-frequency characteristics. The speech signal is more sparsely distributed in the time-frequency domain than the erhu music and the bubbling purr. The three sources of the ambient components are chosen to address the diversity of audio scenes, which includes both nature and urban sounds. The sound of wind has a wider frequency bandwidth, as compared to the bubbling purr and the wave lapping sound.

3.1. Algorithm complexity

Firstly, the boundary of the fine searching range is examined. Fig. 2 shows the phase difference between the extracted ambient components of the APEX and the APES under a confidence level of 90% and different settings of the PPR. The solid lines indicate β calculated by (15), when ε and σ are set to 0.01 and 0.6, respectively. For $k = 1$, the phase difference between the extracted ambient components of the APEX and the APES increases with the PPR. Even though β is not affected by the PPR, it still defines a good angular range for fine searching. For $k > 1$, β matches the phase difference between the extracted ambient components of the APEX and the APES. When k increases, β decreases and the algorithm complexity of the APEN is likely to be reduced with a smaller



(a) speech+wave (b) bubble+station (c) erhu+wind

Figure 2: Phase difference between the extracted ambient components of the APEX and the APES.

fine searching range.

Secondly, the number of search points in the fine and fast searching ranges are denoted as D_0 and D_1. In order to figure out the appropriate settings of D_0 and D_1, the error to signal ratio (ESR) is introduced as the objective evaluation index. The ESRs of the primary and ambient components are defined as

$$ESR_p = 10\log_{10}\sum_{c=0,1}\frac{\left(\boldsymbol{p}_c-\hat{\boldsymbol{p}}_c\right)^T\left(\boldsymbol{p}_c-\hat{\boldsymbol{p}}_c\right)}{2\boldsymbol{p}_c^T\boldsymbol{p}_c} \tag{25}$$

and

$$ESR_a = 10\log_{10}\sum_{c=0,1}\frac{\left(\boldsymbol{a}_c-\hat{\boldsymbol{a}}_c\right)^T\left(\boldsymbol{a}_c-\hat{\boldsymbol{a}}_c\right)}{2\boldsymbol{a}_c^T\boldsymbol{a}_c}, \tag{26}$$

respectively.

Fig. 3 shows ESR$_p$ and ESR$_a$ with respect to the PPR by the APEN when $k = 2$. The results of the three combinations have been averaged to avoid

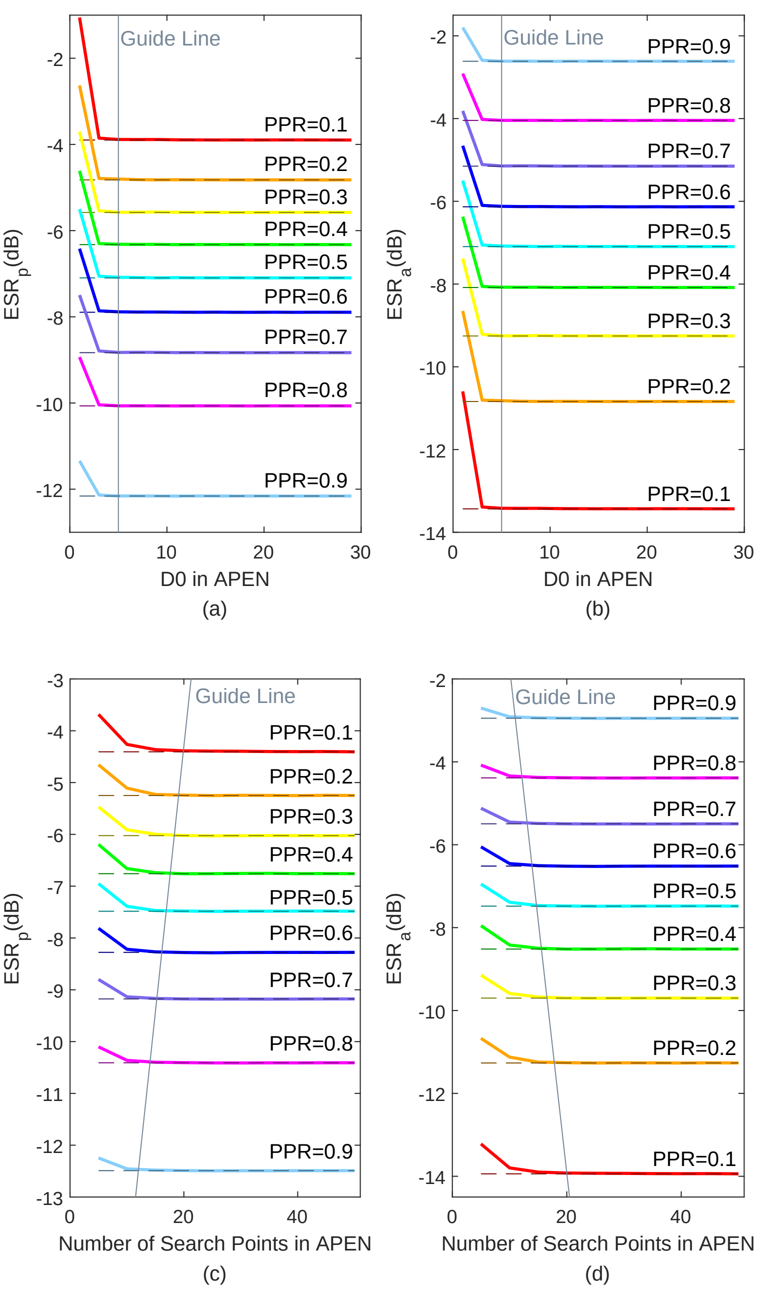


Figure 3: ESR with respect to the PPR by the APEN when $k = 2$. In (a) and (b), the number of search points in the fine searching range (*i.e.* D_0) varies, when the fast searching range is skipped (*i.e.* $D_1 \triangleq 0$). In (c) and (d), the total number of search points (i.e. $D_0 + D_1$) varies.

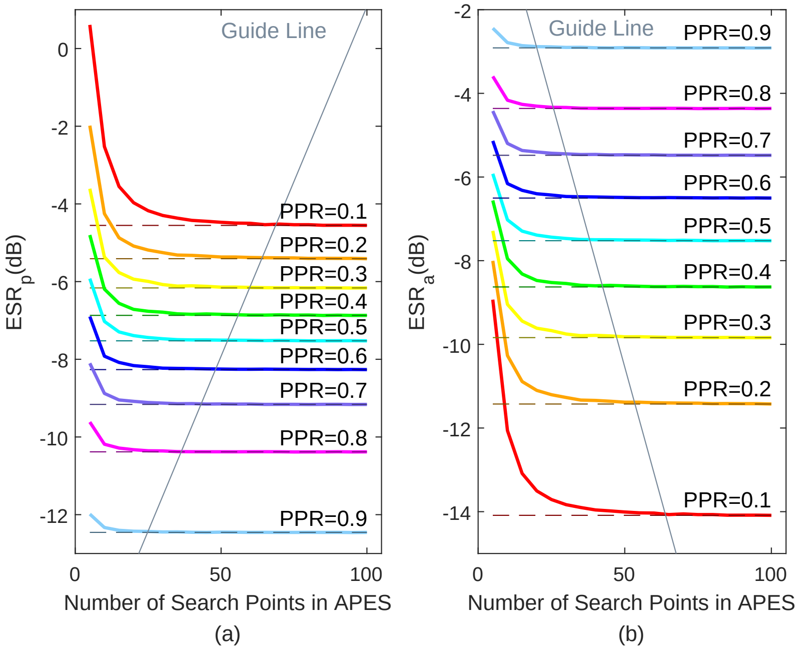


Figure 4: ESR with respect to the number of search points in the APES.

redundant plots, for their individual trends are similar. In Figs. 3(a) and 3(b), D_1 is set to 0 and D_0 varies. It is noted that when $D_0 > 5$, ESR$_p$ and ESR$_a$ no longer improves with more search points. In Figs. 3(c) and 3(d), the total number of search points ($i.e.$ $D_0 + D_1$) varies. Both ESR$_p$ and ESR$_a$ converge when there are 18 search points for the PPR of 0.1 and 10 search points for the PPR of 0.9, respectively. By comparison, as shown in Fig. 4, the APES has to search for at least 70 points and 30 points to achieve their steady-state solutions when the PPR is 0.1 and 0.9, respectively. Hence, the guideline numbers of search points are listed in Table 1. A smaller number of search points is required for larger values of k and PPR. The APEN reduces more than two thirds of the algorithm complexity as compared to the APES.

Table 1: Number of search points suggested with respect to the PPR.

PPR	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
APEN ($k = 2$)	18	14	13	13	11	11	10	10	10
APES ($k = 2$)	70	70	65	55	50	45	35	35	30
APEN ($k = 4$)	15	15	13	13	11	10	8	5	5
APES ($k = 4$)	60	55	50	45	40	40	35	30	25

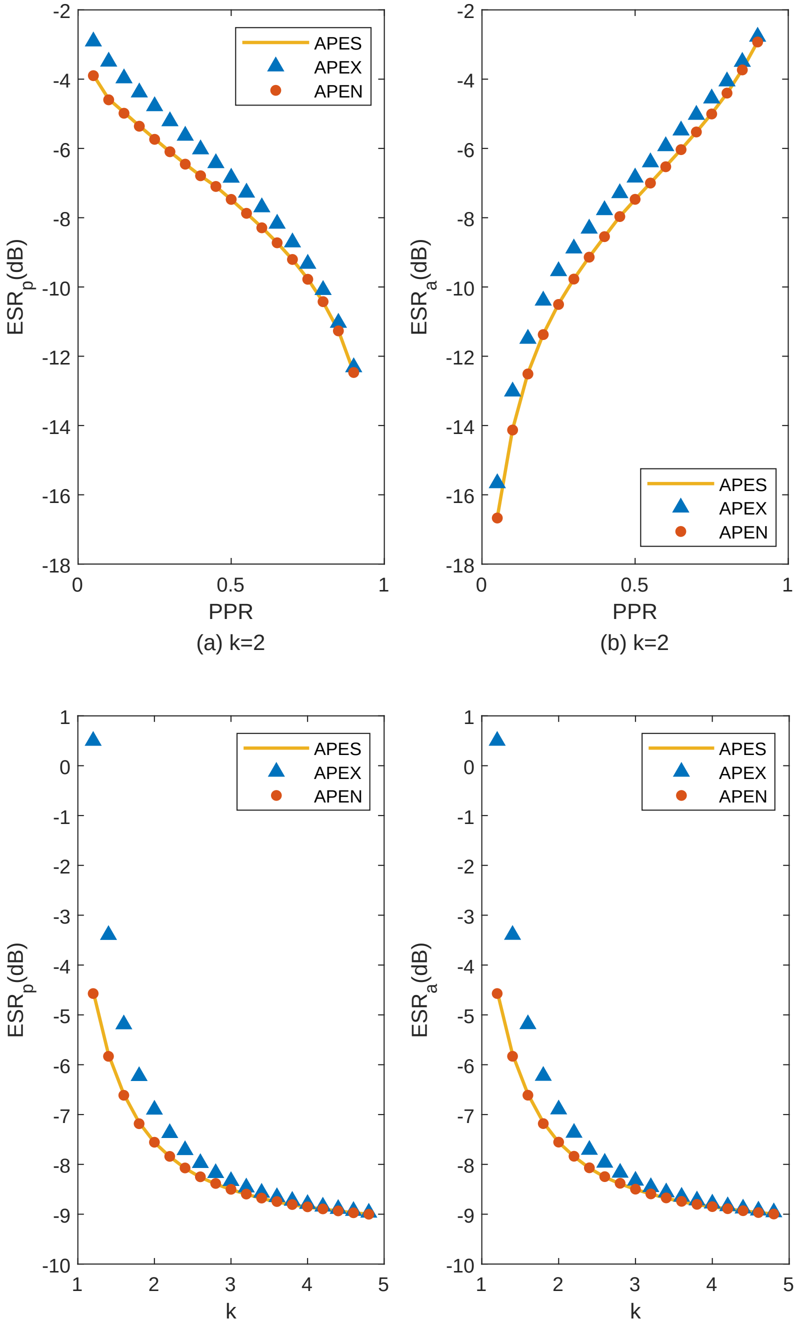
3.2. Objective evaluation

The performance of the APEN is compared with the other PAE algorithms under the ambient spectrum estimation framework by a series of objective evaluation indices [35]. Similarly, the results of the three combinations have been averaged to avoid redundant plots.

Firstly, the ESR curves are plotted in Fig. 5 by using the APES, APEX and APEN. In Figs. 5(a) and 5(b), the PPR varies from 0.1 to 0.9, while the panning factor is fixed at $k = 2$. With the increase of the PPR, the primary component can be extracted more accurately, at a cost of increased error in the ambient component extraction. The APEN achieves almost the same extraction accuracy as compared to the APES. In Figs. 5(c) and 5(d), the PPR is set to 0.5, while the panning factor varies with an interval of 0.2. It is noted that the APEN can extract both the primary and ambient components as accurately as the APES. Both of the APEN and the APES outperforms the APEX in terms of the ESR.

Secondly, the perceptual evaluation of audio quality (PEAQ) is employed to evaluate the performance of the APES, APEX and APEN. The PEAQ is a standard recommended by ITU that provides a computational model of the human auditory system to measure the perceived audio quality [36, 37]. The output of the PEAQ follows the subjective difference grade (SDG), of which the values range from -4 (very annoying) to 0 (imperceptible).

In Fig. 6, the PEAQ grades are plotted for the extracted primary compo-



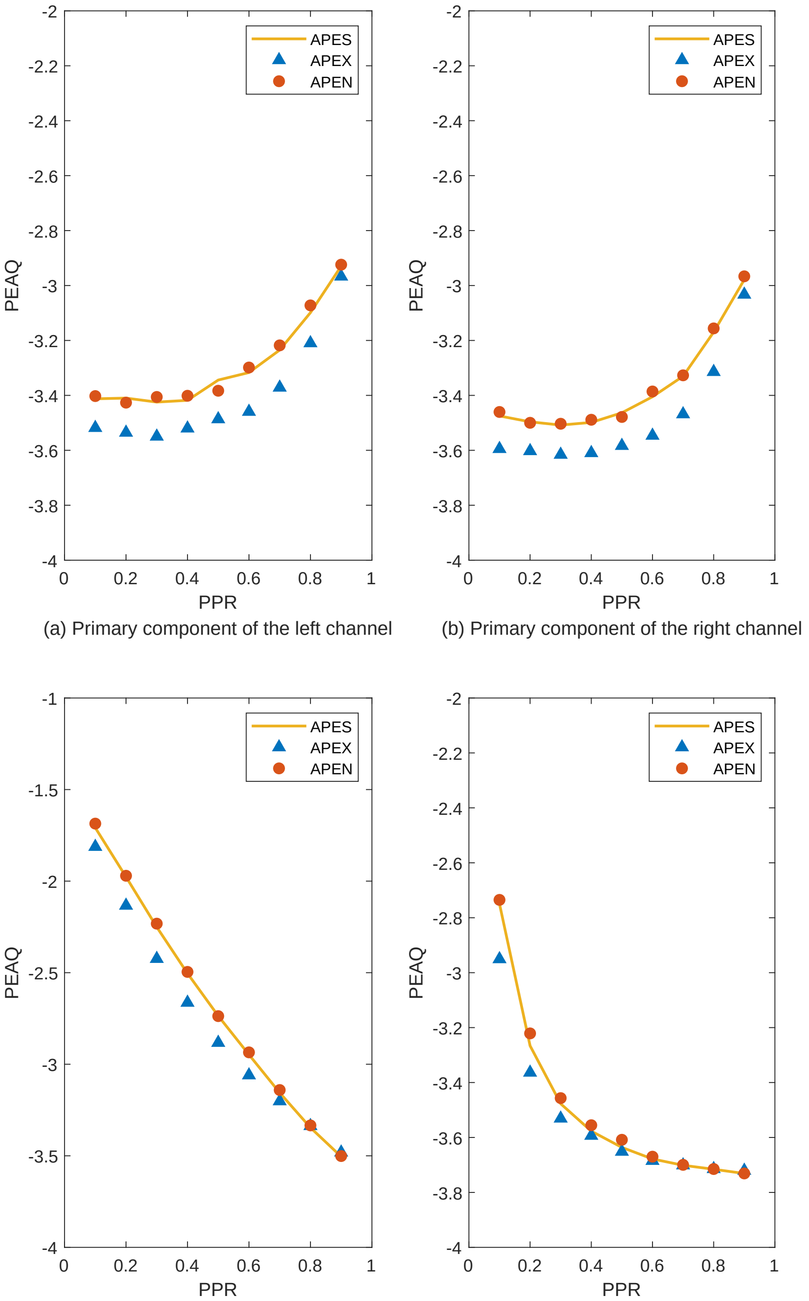
(a) k=2

(b) k=2

(c) PPR=0.5

(d) PPR=0.5

Figure 5: ESR of the primary and ambient components extracted by the APES, APEX and APEN.



(a) Primary component of the left channel

(b) Primary component of the right channel

(c) Ambient component of the left channel

(d) Ambient component of the right channel

Figure 6: PEAQ of the primary components and ambient components extracted by the APES, APEX and APEN.

nents and ambient components separately. The PPR varies from 0.1 to 0.9, while the panning factor is fixed at $k = 2$. Fig. 6(a) shows that the APES achieves the best extraction quality of the primary components. The APEN is close to the APES, while the APEX produces the worst result. Fig. 6(b) shows that the APES and the APEN can achieve the same extraction quality of the ambient components, while the APEX is not much behind the APES and the APEN.

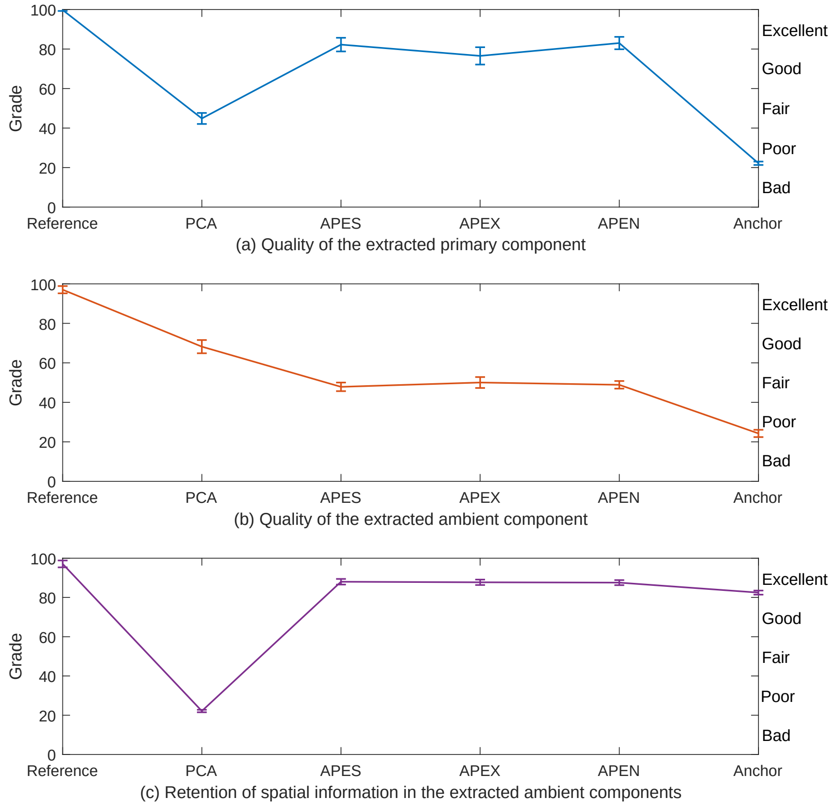
In general, the ambient spectrum estimation framework is better at extracting the ambient components rather than the primary components when the PPR is a relatively small value. The APEN can obtain faster calculation speed and equivalent accuracy of PAE, as compared to the APES. Both the APES and the APEN have higher accuracy of PAE than the APEX.

3.3. Subjective evaluation

The multi stimulus with hidden reference and anchors (MUSHRA) listening test was conducted to evaluate the primary components and the ambient components extracted by the PCA, APES, APEX and APEN [38]. The panning factor was set as $k = 2$. The PPR was set to 0.4 and 0.8. The original sound clips were used as the reference and hidden reference stimulus. The primary and ambient components extracted by the masking algorithm was considered as the anchor.

Before the listening test, there was a briefing session to explain the principle and steps of the MUSHRA listening test to the participants. During the listening test, participants were asked to focus on the similarity of the extracted primary component as compared to the reference, the similarity of the extracted ambient component as compared to the reference and the retention of spatial information in the extracted ambient components. Every stimulus was graded from 0 to 100. The hidden reference stimuli were expected to receive a near perfect score of 100, while the anchor stimulus would probably receive the lowest scores.

The listening test was conducted in the scientific research building at the University of Electronic Science and Technology of China. An audio-technica



(a) Quality of the extracted primary component

(b) Quality of the extracted ambient component

(c) Retention of spatial information in the extracted ambient components

Figure 7: Subjective listening test results of the PCA, APES, APEX and APEN.

ATH-M30x headphone was used. In total, there were 11 male and 6 female participants. Their ages ranged from 21 to 26. All the participants were confirmed to have normal hearing. The listening test results of the three combinations were statistically analyzed into mean values and 95% confidence intervals, which were plotted in Fig. 7.

In Fig. 7(a), the primary components extracted by the PAE algorithms under the ambient spectrum estimation framework demonstrate better audio quality than those extracted by the PCA. The audio quality of the primary components extracted by the APEN is very close to the APES and slightly better than the APEX. This is consistent with the objective evaluation results. Fig. 7(b) shows that the PCA results in the best audio quality of the ambient components. A similar observation was previously reported in [22]. It was explained that during the listening test, participants paid extra attention to the similarity between the stimulus and the reference, without too much consideration to the spatial information. Therefore, Fig. 7(c) further presents the retention of the spatial information in the extracted ambient components. The PCA results in the worst spatial perception, while the APES, APEX and APEN manage to achieve similar results that are nearer to the reference. The listening test results validate that through the non-uniform searching strategy, the APEN can obtain faster calculation speed and ensure the equivalent accuracy of PAE, as compared to the APES.

4. Conclusions

In this paper, a fast non-uniform searching strategy is proposed to accelerate the solving process of the ambient phase estimation, resulting in the APEN. Differing from the state-of-the-art algorithm that adopts a constantly high angular resolution in the full range of searching, the APEN combines a fine searching in a small range with high confidence and a fast searching in the rest with low confidence. The APEN is only one third as computationally complicated as the APES, while providing equivalent accuracy to the APES in objective and subjective evaluations.

5. Acknowledgements

This manuscript is supported jointly by the National Natural Science Foundation of China and the Civil Aviation Administration of China (Joint Grant No. U1933127). The authors would like to thank Ms. Lu Chen for her assistance with the implementation of algorithms.

References

[1] D. R. Begault and L. J. Trejo, "3-D sound for virtual reality and multimedia," *Technical Memorandum*, NASA/TM-2000-209606, 2000.

[2] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—The new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770-779, 2015.

[3] D. Rojas *et al.*, "The effect of sound on visual fidelity perception in stereoscopic 3-D," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1572-1583, 2013.

[4] J. M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, "Rendering spatial sound for interoperable experiences in the audio metaverse," in Proceedings of the International Conference on Immersive and 3D Audio: from Architecture to Automotive, 2021.

[5] H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. O. Smith III, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36-54, 2017.

[6] A. Roginska and P. Geluso, *Immersive sound: The art and science of binaural and multi-channel audio*, Abingdon, UK: Taylor & Francis, 2017.

[7] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77,2010.

[8] Blauert and Jens, *Spatial hearing: The psychophysics of human sound localization*, Cambridge, MA: MIT Press, 1983.

[9] B. S. Xie, *Head-related transfer function and virtual auditory display*, Plantation, FL: J. Ross Publishing, 2013.

[10] H. Tatsuya, S. Hiroyuki, T. Iwaki, and O. Makoto, “Head movement during head-related transfer function measurements,” *Acoustical Science and Technology*, vol 31, pp. 165-171, 2010.

[11] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, “HRTF personalization using anthropometric measurements,” In Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003, pp. 157-160.

[12] J. Herre and S. R. Quackenbush, “MPEG-H 3D audio: Immersive audio coding,” *Acoustical Science and Technology*, vol. 43, no. 2, pp. 143-148, 2022.

[13] International Telecommunication Union, “Multichannel stereophonic sound system with and without accompanying picture,” *International Telecommunication Union*, ITU-R-Rec-BS.775-3, 2012. [Online]. Available: https://www.itu.int/rec/R-REC-BS.775/. [Accessed: July., 2022].

[14] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio—The new standard for universal spatial/3D audio coding,” *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2014.

[15] P. Damaske and Y. Ando, “Interaural crosscorrelation for multichannel loudspeaker reproduction,” *Acta Acustica united with Acustica*, vol 27, pp. 232-238, 1972.

[16] T. Sugimoto, Y. Nakayama, and T. Komori, “22.2 ch audio encoding/decoding hardware system based on MPEG-4 AAC,” *IEEE Transactions on Broadcasting*, vol. 63, no. 2, pp. 426-432, 2017.

[17] C. Eaton and H. Lee, “Subjective evaluations of three-dimensional, surround and stereo loudspkeare reproductions using classical music recordings,” *Acoustical Science and Technology*, vol. 43, no. 2, pp. 149-161, 2022.

[18] F. Rumsey, “Time-frequency processing of spatial audio,” *Journal of the Audio Engineering Society*, vol. 58, no. 7, pp. 655–659, 2010.

[19] M. R. Bai and G. Y. Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *Journal of the Audio Engineering Society*, vol. 53, no. 3, pp. 1011–1019, 2007.

[20] M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," In Proceedings of the 123rd Audio Engineering Society Convention, 2007, pp. 7277.

[21] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," In Proceedings of the 128th Audio Engineering Society Convention, 2010, pp. 7986.

[22] J. He, *Spatial audio reproduction with primary ambient extraction*, Singapore: Springer Publishing Company, 2016.

[23] M. A. Gerzon, "Optimum reproduction matrices for multispeaker stereo," *Journal of the Audio Engineering Society*, vol. 40, no. 7/8, pp. 571–589, 1992.

[24] N. Stefanakis and A. Mouchtaris, "Foreground suppression for capturing and reproduction of crowded acoustic environments," In Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, 2015, pp. 51-55.

[25] K. M. Ibrahim and M. Allam, "Primary-ambient source separation for upmixing to surround sound systems," In Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018, pp. 431-435.

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

[27] M. Goodwin and J. M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," In Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, pp. I-9-I-12.

[28] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 505–517, 2014.

[29] C. Avendano and J. M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," In Proceedings of the 27th IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, pp. II-1957-II-1960.

[30] 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.

[31] L. Chen, C. Shi, and H. Y. Li, "Primary ambient extraction for random sign Hilbert filtering decorrelation," In Proceedings of the 23rd International Congress on Acoustics, 2019, pp. 7239-7246.

[32] H. Zhu, C. Shi, and Y. Wang, "F0-estimation-based primary ambient extraction for stereo signals," In Proceedings of the 29th European Signal Processing Conference, 2021.

[33] 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.

[34] M. Goodwin and J. M. Jot, "Spatial audio scene coding," In Proceedings of the 125th Audio Engineering Society Convention, 2008, pp. 7507.

[35] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques—A review, and recent developments," *Signal Processing*, vol. 89, pp. 1489-1500, 2009.

[36] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, 2008.

[37] International Telecommunication Union, “Method for objective measurements of perceived audio quality,” *International Telecommunication Union*, ITU-R-Rec-BS.1387-1, 2001. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1387/en. [Accessed: Mar., 2022].

[38] M. Schoeffler, A. Silzle, and J. Herre, “Evaluation of spatial/3D audio: Basic audio quality versus quality of experience,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 75-88, 2017.