Gao *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:13
DOI 10.1186/s13636-016-0091-z

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

JND-based spatial parameter quantization of multichannel audio signals

Li Gao[1,2,3], Ruimin Hu[1,2,3*], Xiaochen Wang[1,2,3], Gang Li[1,2,3], Yuhong Yang[1,2,3] and Weiping Tu[1,2,3]

Abstract

In multichannel spatial audio coding (SAC), the accurate representations of virtual sounds and the efficient compressions of spatial parameters are the key to perfect reproduction of spatial sound effects in 3D space. Just noticeable difference (JND) characteristics of human auditory system can be used to efficiently remove spatial perceptual redundancy in the quantization of spatial parameters. However, the quantization step sizes of spatial parameters in current SAC methods are not well correlated with the JND characteristics. It results in either spatial perceptual distortion or inefficient compression. A JND-based spatial parameter quantization (JSPQ) method is proposed in this paper. The quantization step sizes of spatial parameters are assigned according to JND values of azimuths in a full circle. The quantization codebook size of JSPQ was 56.7 % lower than one of the quantization codebooks of MPEG surround. Average bit rate reduction on spatial parameters for standard 5.1-channel signals reached up to approximately 13 % compared with MPEG surround, while preserving comparable subjective spatial quality.

Keywords: Spatial audio coding, Multichannel signals, Spatial parameters, Just noticeable difference

1 Introduction

Along with the trend towards high-quality audio, audio systems have evolved through mono, stereo, to multichannel audio systems. The multichannel audio systems developed from 5.1 to 22.2 (NHK) [1], 64 (Dolby Atmos [2]) channels loudspeaker systems and other multichannel audio systems with even more loudspeakers (e.g., loudspeaker systems with Ambisonics [3] or WFS [4]). With more loudspeakers configured in three-dimensional (3D) space, current multichannel audio systems can freely reproduce virtual sound image in 3D space. However, with the increasing number of loudspeakers, it will bring challenges to the efficient storage and transmission of large amounts of channel data in current multichannel audio systems. Take the 22.2-channel audio system for example, the data rate will reach 28 Mbps before data compression. Even with general efficient audio codec such as MP3 with data rate 128 kbps used to compress all the channel data, the data amounts will still reach up to 10 G bits for an hour's 22.2-channel signals, which are often not affordable for the storage device and transmission bandwidth in real application. So high-efficient compression schemes for multichannel audio signals play an important role on popularization of current 3D multichannel audio systems.

Spatial audio coding (SAC) [5] has been an ongoing research topic in recent years for the high-efficient compression performance. SAC represents two or more channel signals as one or two downmix signals, accompanied with spatial parameters extracted from channels to model spatial attributes of the auditory scene generated by original channel signals. With the increasing channel number, the total amount of coded information does not notably increase in SAC compared with conventional stereo or multichannel audio coding schemes.

In multichannel audio systems, the virtual sounds are widely distributed in space and a large amount of complex spatial information is contained. It is not that easy to find a few finite spatial parameters to comprehensively represent all spatial attributes of the auditory scene in multichannel system. Many aspects might be very important, such as the basic audio quality, auditory spatial image, auditory

*Correspondence: hurm1964@gmail.com
[1] State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China
[2] National Engineering Research Center for Multimedia Software, the Key Laboratory of Multimedia and Network Communication Engineering, Computer School, Wuhan University, Wuhan, China
Full list of author information is available at the end of the article

object location/width, ambience, listener envelopment, etc. [6]. Many processes in multichannel system might add artifacts to these aspects and need to be optimized, such as multichannel downmixing, quantization of the downmix, and 3D reproduction. For example, the downmix signals in SAC are closely related to both basic and spatial audio quality and often encoded by an outstanding perceptual encoder such as AAC [7] to ensure basic quality.

Although there may be many complex spatial parameters associated with spatial quality, and the exact spatial location of perceived sound may be not the most important aspect in a multichannel system, the accurate representation and efficient compression of spatial information of virtual sound is still very important to the perfect and vivid reconstruction of spatial sound effects in SAC. In the general SAC methods, spatial parameters mainly include spatial parameters between two channel signals such as inter-channel level difference (ICLD) and inter-channel time difference (ICTD) [5], as well as spatial parameters among more than two channel signals such as azimuths of virtual sound images, which are closely related to the spatial location of virtual sound.

In recent years, lots of research achievements in spatial psychological acoustics model have contributed a lot to the development of SAC. The spatial perceptual features of human auditory system, for example, the perceptual sensitivity limits of human auditory system, have played important roles in the perceptual redundancy removal of spatial parameter quantization in SAC. The change of sound property (for instance, the direction or location of sound) has to reach a certain amount to be detectable for the human auditory system, and such required minimum detectable change is often referred to as just noticeable difference (JND) [8–11]. Although not all the spatial perceptual aspects of the human auditory system have the JND characteristics, and for most spatial perceptual aspects the JND characteristics are still not very clear at present, many researches have been done on the directional perceptual JND characteristics of the human auditory system. And the relevant research results have been used in the perceptual redundancy removal of spatial parameter quantization in general SAC methods. Thus, with the perceptual characteristics of JND, if the spatial distortion caused by quantization errors of spatial parameters can be limited below the JND thresholds, then the quantization is almost perceptual lossless.

2 Related work

The spatial perceptual features of human auditory system have played important roles in the perceptual redundancy removal of spatial parameters in SAC. Early in 1877, Strutt has stated that binaural cues take the leading role in the human auditory system in the judgment of sound direction in horizontal plane [12]. Binaural cues are mainly due to the comprehensive filter action of reflection, absorption, and attenuation in the transmission process of sound; thus, the sounds that arrive at humans' two ears are different from each other and also different with the original sound. The major binaural cues include interaural level difference (ILD), and iInteraural time difference (ITD). In the case of playback with headphones, the signals played out of the headphones are almost the same with the signals that arrive at the human ears. Thus, the inter-channel cues (also called spatial parameters, such as ICLD) are almost the same with inter-aural cues (also called binaural cues).

As for the auditory scene reproduced by stereo loudspeaker system, without considering the transmitting procedures of sound from loudspeakers to ears, inter-channel cues can be regarded as approximations of binaural cues. For example, one of the most important spatial parameters ICLD can be regarded as an approximation of interaural level difference (ILD) [11]. Then the spatial perceptual features of binaural cues (such as JND of ILD) can be used as a reference in the quantization of spatial parameters (such as ICLD) to remove perceptual redundancy. These spatial hearing characteristics of human auditory system were used in the first SAC framework Binaural Cue Coding (BCC) to reconstruct spatial effect in stereo signal with a bit rate of only 2 kbps for spatial parameters [13, 14].

However, the spatial parameters are uniformly quantized in BCC, which means that the quantization step sizes between different quantization codewords are the same. But in fact, the perceptual sensitivities for different cues in human auditory system are not the same. For example, the JNDs of ILDs with different values are not the same [15]. When ILD is 0 dB, this ILD has the smallest JND that is about 0.5 dB. As absolute ILD increases, JND increases too. For ILD of 15 dB, the JND amounts to about 2 dB. So as not to introduce audible spatial distortion, the quantization errors of spatial parameters should be controlled below JNDs. Different spatial parameters with different JNDs should be assigned with different quantization step sizes.

Ignoring the transmitting procedures from sound source to ears with earphone or stereo audio system, ICLD can be regarded as an approximation of ILD. In this case, the JND characteristic of ILD can be referred in the quantization codebook design of ICLD. Thus, in parametric stereo (PS), the uniform quantization of spatial parameters are replaced by nonuniform quantization according to the perceptual sensitivity attributes of interaural cues. Different quantization step sizes are set between quantization codewords. For example, small quantization steps are assigned to ICLDs with small JNDs. For ICLD value of 0 dB, the quantization step size is the smallest. As for bigger ICLD values, the quantization step sizes are bigger.

With the nonuniform quantization of spatial parameters, PS was known as one of the most efficient stereo coding schemes and was standardized as MPEG-4 AAC Plus v2 or 3GPP Enhanced AAC+ (E-AAC+) [16, 17].

SAC was extended from stereo to multichannel signals and standardized as ISO/MPEG multichannel coding standard MPEG surround [15]. With two kinds of elementary coding blocks one-to-two (OTT) or two-to-three (TTT), multichannel signals are combined pairwise or per three channel signals to form coding tree structures. OTT/TTT referring to two/three channel signals are represented as one/two downmix accompanied with extracted spatial parameters. Each OTT block has the same coding procedure as in PS, as well as the same quantization methods of spatial parameters as in PS. All OTT blocks for different channel pair signals share the same quantization codebooks of spatial parameters.

Although ICLD can be regarded as an approximation of ILD ignoring the transmission filter procedures in earphone or stereo audio system, the transmission procedures are not ignorable in multichannel audio systems. Owing to the filter function of transmission process from multiple loudspeakers to ears, significant differences exist between inter-channel cues and interaural cues in the case of playback with loudspeakers. For example, with two loudspeakers with azimuths $-110°$ and $30°$, the difference of ICLD and corresponding ILD could be up to 55 dB [18].

The correlations of inter-channel parameters and interaural cues (such as ICLD and ILD) closely depend on the location of loudspeaker pairs. In the case of loudspeaker pairs with arbitrary locations in multichannel systems, there are significant differences between ICLD and ILD, which cannot be treated as the approximation correlation. It is obviously inappropriate to design quantization codebook for ICLDs of arbitrary channel pairs uniformly according to JND of ILD. The same quantization codebooks for spatial parameters of channel pairs in different directions will definitely degrade the spatial quality.

Besides JND of binaural cues often used to analyze the spatial perceptual sensitivity of human auditory system, the JND of sound angle change (also referred to as minimum audible angle, MAA) has also been investigated by lots of previous researchers. Early in 1958, Mills presented a classic method to measure the JND of human auditory system when sound angle changed [10]. The experimental measurements showed that for the sounds with frequencies from 500–3700 Hz located in front of subjects, when the sound azimuths changed about 1–3.1°, the subjects can just notice the change.

Supposing the JND values for all azimuths in a circle of 360° are a fixed value 3°, Choi designed a quantization codebook of spatial parameters. Owing to the different configurations of different loudspeaker pairs (such as the different interval angel between each two loudspeakers), Choi designed a specific quantization codebook for each specific loudspeaker pair in 5.1 loudspeaker system. However, the quantization codebooks had no essential differences with MPEG surround [19].

It is known that the ICLD value between two channel signals may be any value from 0 to infinite. When ICLD reaches some threshold, even if ICLD continues to increase, the spatial sound image will limitlessly approach one of the two loudspeakers but never go outside of the region between two loudspeakers. No matter how ICLD changes, the spatial sound image will always locate between two loudspeakers. Although ICLD may be infinite values without threshold, but the azimuth of virtual sound image has threshold and can only be value between $-180°$ and $180°$ in a circle.

Thus, instead of spatial parameter ICLD between channel pair to represent spatial direction of virtual sound, Cheng proposed to extract the azimuth of virtual sound as spatial parameter. Cheng designed three different quantization codebooks for azimuths in three regions: front, side, and rear regions [20]. The quantization resolutions of azimuths in three regions were different. But in each region, the azimuth quantization resolution was uniform. The front region (from $-30°$ to $30°$) had the smallest quantization step sizes with two kinds of quantization resolution $2°$ or $3°$. The quantization resolution for side regions (from $30°$ to $110°$, or from $-30°$ to $-110°$) were $6.6°/20°$. As for the rear region from $-110°$ to $110°$, the azimuth resolutions were fixed $17.5°/35°$. Since the azimuth quantization errors of Cheng's method were much bigger than the azimuthal JNDs, obvious perceptual distortion of spatial quality was introduced.

To improve the spatial quality of multichannel spatial audio coding, Elfitri presented a closed-loop encoding system applied on MPEG surround [21]. The closed-loop procedure inversely quantized downmix and spatial parameters at the encoder to get decoded signals and then obtained new residual signals between original and decoded signals. The improved spatial quality mainly benefited from the accurate extraction of residual signals, however, had nothing to do with the quantization method of spatial parameters. Meanwhile, the closed-loop procedure will significantly increase the computational complexity at the encoder.

The abovementioned spatial audio coding methods are mainly based on the elementary mechanism of parametric stereo and focus on the quantization of spatial parameters between two channel signals. The inter-channel spatial parameters such as ICLD can be extracted from only two channel signals. Cheng proposed a method to extract the azimuth of virtual sound from an arbitrary number of loudspeaker signals [22]. In Cheng's method, downmix signal of multichannel signals were obtained as the signal of virtual sound source. Then, the signal and the azimuth

of virtual sound source were coded. It seemed similar with what spatial audio object coding (SAOC) schemes do [23, 24]. But they are different in fact. The spatial parameters in SAC represent spatial information of virtual sound. However, the parameters in SAOC are correlation parameters between audio object signals, which have nothing to do with the spatial location information of sound. The quantization of parameters in SAOC is beyond the scope of this paper. Cheng designed uniform quantization codebooks with fixed step sizes of 2° or 3° for azimuths in horizontal plane. Since high quantization resolution of azimuths in Cheng's method, the quantization codebook sizes for azimuths in a full circle were almost twice or triple of that in MPEG surround and resulted in high bit rates of spatial parameters [22].

Since Mills' experiments conducted to get azimuthal JND in 1958, lots of experiments have been conducted by researchers to study the directional perceptual sensitivity of human auditory system. The results showed that when sound was located in front of subjects in a horizontal plane (where sound azimuth was 0°), azimuthal JND was the smallest; when sound azimuth increased, JND increased too; when sound azimuth was 90° or −90° (sound was located at the sides of subjects), JND was the biggest; azimuthal JND of the rear sound was almost twice that of the front sound [25–27].

With the different JNDs for different azimuths, different azimuths in a full circle should be assigned with different quantization step sizes according to azimuthal JND data. The azimuthal JND characteristics were used in the quantization method of spatial parameters as a first attempt in [28]. An adaptive quantization scheme of spatial parameter ICLD according to arbitrary loudspeaker configurations was proposed. However, the spatial parameter ICLD was still extracted from channel pair signals. The quantization method for spatial parameters extracted from multichannel signals was not discussed in detail. The objective and subjective experiments about quantization performance was limitedly discussed and mainly compared with MPEG surround.

In this paper, the quantization of spatial parameters between two or more channel signals are both discussed. The main contributions and works include the following: to accurately represent the virtual sound, a method to estimate spatial parameter azimuth and the signal of virtual sound from an arbitrary number of loudspeakers was proposed; an azimuthal JND based spatial parameters quantization method (JSPQ) was proposed; and the generation procedure of azimuth quantization codebook was elaborated in details. The quantization of spatial parameter is more consistent with the JND values of azimuths in a full circle, and thus, perceptual redundancy of spatial parameter can be efficiently removed with coding bit rate as low as possible. Objective experiments and subjective evaluation were conducted to confirm that the proposed JSPQ outperformed reference quantization methods of spatial parameters in the respects of codebook sizes, quantization errors, coding bit rates, and spatial qualities.

3 Spatial audio coding

3.1 Codec structures of SAC

According to the extraction method of spatial parameter, there are mainly two kinds of coding structures for high-efficient multichannel SAC. One typical representative is the tree-structured layer-by-layer coding schemes such as in MPEG surround [15] and MPEG 3D audio standard [24]. Take MPEG surround for example; after T/F transformation, multichannel signals are combined pairwise or per three channel signals using elementary coding blocks (OTT or TTT). As illustrated in Fig. 1, spatial parameters (ICLD, ICC, etc.) and downmix signals are acquired in each OTT blocks before quantization in the encoder. In the decoder, decoded downmix signals and spatial parameters are processed in inverse OTT blocks to get recovered multichannel T/F signals. In each OTT or TTT block in the cascaded tree structure, a set of spatial parameters are extracted from different channel signals or downmix signals. However, for each kind of spatial parameter (such as ICLD), the same quantizer is used as applied in parametric stereo coders.

The other type of spatial audio coding structure is based on virtual sound source information representation as in Fig. 2 [22]. Given that only one virtual sound source generated by loudspeakers exists for each time-frequency bin, multichannel signals can be represented with a virtual sound source with its spatial location information. The signal and location information of the virtual sound source can be extracted from an arbitrary number of loudspeaker signals in encoder. With inverse-quantized information of virtual sound in decoder, virtual sound can be reproduced by general panning techniques such as vector-based amplitude panning (VBAP) [29], or Ambisonics techniques such as higher order Ambisonics (HOA) [30] to get recovered multichannel T/F signals.

The proposed JND-based spatial parameter quantization methods (JSPQ) in this paper are independent of the SAC structure. Thus, the modified quantization methods for spatial parameters ICLD and azimuth can be used as substitutions respectively in these two types of SAC structures.

3.2 Quantization of spatial parameters

ICLD is one of the most important spatial parameters to represent the direction information of virtual sound image and is commonly used in SAC such as E-AAC+ [17] and MPEG surround [15]. In MPEG surround for coding multichannel signals, ICLD values are extracted from pairwise channel signals pair by pair among all channels.

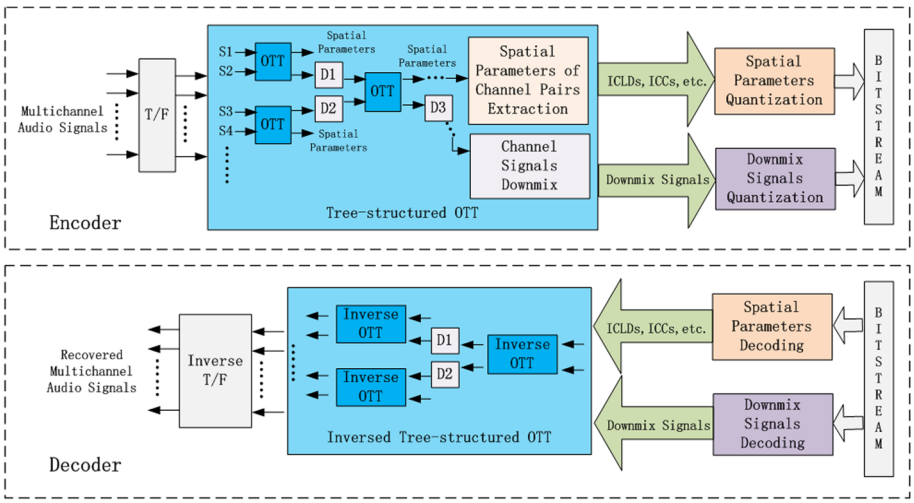

Fig. 1 Spatial audio coding structure in MPEG surround

Given the frequency amplitude of two loudspeakers, $X_1(k)$ and $X_2(k)$, respectively, k is the index of frequency bin and ICLD can be calculated in each frequency Bark band b ($b \in \{0, 1, 2, .., 23\}$) as:

$$ICLD(b) = 10\log_{10}\left(\frac{E_1(b)}{E_2(b)}\right), \tag{1}$$

$$E_1(b) = \sum_{k=k_b}^{k_{b+1}-1} X_1^2(k), \tag{2}$$

$$E_2(b) = \sum_{k=k_b}^{k_{b+1}-1} X_2^2(k) \tag{3}$$

Except ICLD, the azimuth of virtual sound is also commonly used as one of the spatial parameters as in [22], especially used for more than two loudspeakers. The azimuth of virtual sound extracted from two loudspeaker signals by:

$$\theta_0 = \arctan\left(\frac{g_L - g_R}{g_L + g_R}\tan\theta\right), \tag{4}$$

in which g_L and g_R are the gain parameters of two loudspeakers signals, and θ is half the intersection angle between two loudspeakers. Given there is only one sound source that can be perceived by human auditory system in each frequency bin, ICLD and azimuth can be converted to each other under the condition of two loudspeakers.

The quantization codewords distribution of spatial parameters ICLDs and azimuths in different methods

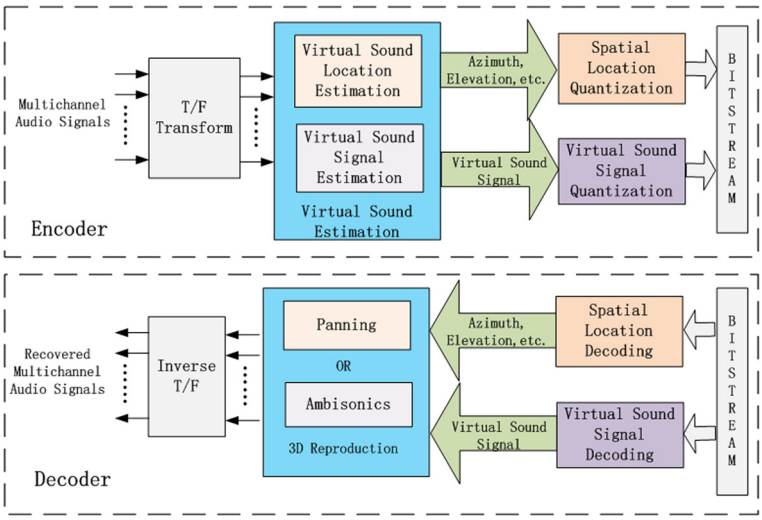

Fig. 2 Spatial audio coding structure based on virtual sound source information representation

(MPEG surround [15], Cheng's [20], Choi's [19] methods and SLQP [22]) are intuitively illustrated in Fig. 3.

MPEG surround and Choi's methods provide quantization of spatial parameters ICLDs between channel pairs. ICLDs extracted from four kinds of channel pairs under 5.1 loudspeaker configuration are converted to azimuths in Fig. 3a–c. Some methods have two kinds of quantization resolutions; thus, H and L, respectively, represent high and low resolution quantization methods. Each red point on the circles represents one quantization codeword of azimuth. Every two adjacent solid lines form the quantization interval region for the inside quantization codeword. That means the azimuths in each quantization interval region will be quantized as the quantization codeword inside the region. Then, are these distribution of azimuth codewords in Fig. 3a–f reasonable for azimuthal perceptual quantization or not?

Researches about the directional perceptual sensitivity of human auditory system showed that: humans have different azimuthal JNDs for sound in different directions. The smallest azimuthal JND corresponds to front directions, bigger JND for rear directions and biggest azimuthal JND for side directions [8–11]. According to the azimuthal JND features of human auditory system, the azimuths with small JNDs allow small quantization errors, the azimuths with big JNDs allow big quantization errors. The quantization step sizes of front azimuths should be the smallest. The quantization step sizes of side azimuths should be the biggest.

But it can be observed from Fig. 3 that the previous quantization methods of spatial parameters are not correlated well with the azimuthal JND. It is evident that there are four regions that are obviously filled with dense solid lines both in Fig. 3a, c, which correspond to the four loudspeakers's locations (L, R, Ls, Rs) in 5.1 audio system. The quantization codewords around the four loudspeakers are the densest in Fig. 3a, c. In Fig. 3a, b, d and e, the most sparse distribution of codewords all lie in rear regions. In Fig. 3f, all the azimuth codewords are equally spaced for the uniform quantization with fixed step sizes. Thus, the quantization step sizes of spatial parameters in current SAC methods are not in coincidence with the JND

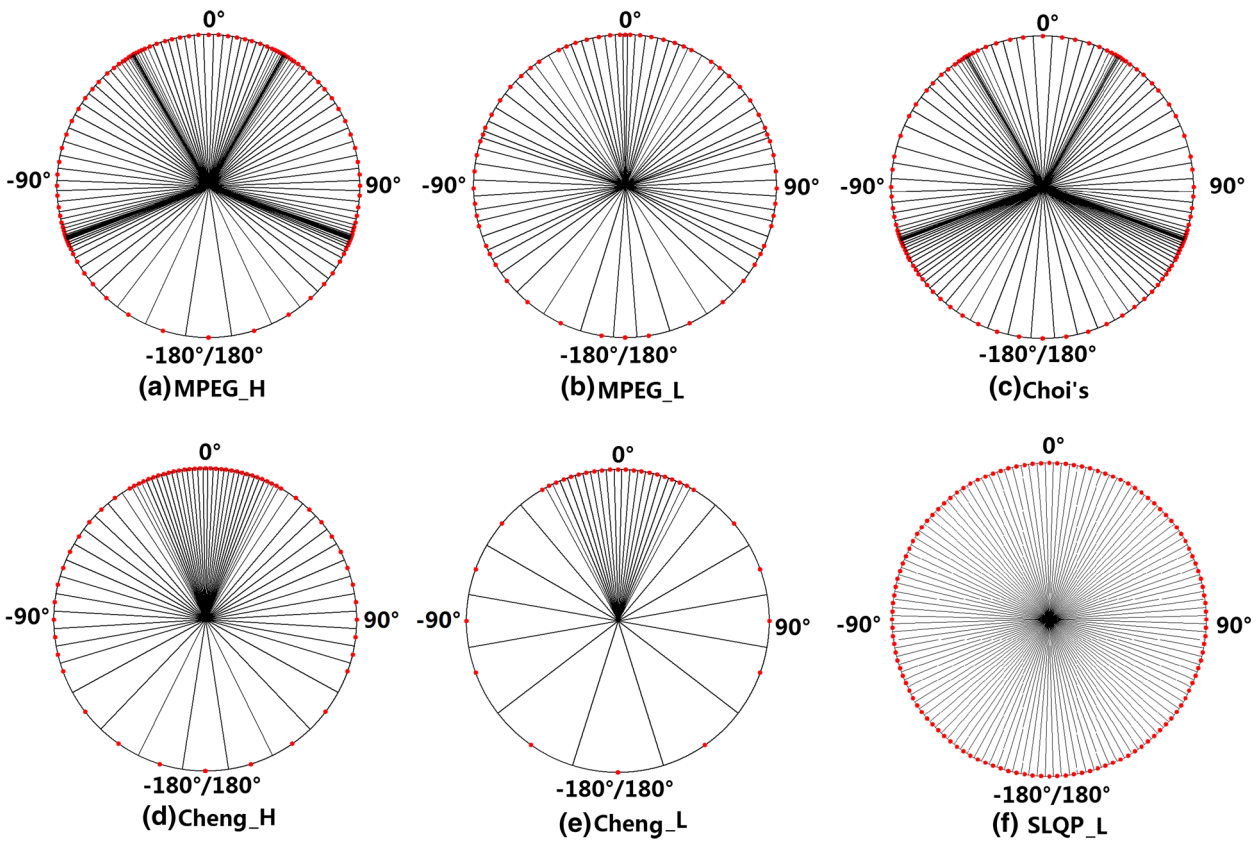

Fig. 3 Quantization codewords distributions in different methods. **a** High resolution quantization of spatial parameters in MPEG Surround [15], denoted as MPEG_H; **b** Low resolution quantization of spatial parameters in MPEG Surround [15], denoted as MPEG_L; **c** Choi's method [19]; **d** High resolution quantization in Cheng's method [20], denoted as Cheng_H; **e** Low resolution quantization in Cheng's method [20], denoted as Cheng_L; **f** Low resolution quantization in SLQP [22], denoted as SLQP_L

characteristics of different azimuths in a full circle. These quantization methods of spatial parameters will result in either spatial perceptual distortions of virtual sound images, or inefficient compressions of spatial parameters.

The intention of this paper is to effectively use the characteristic of azimuthal JND to design the quantization of spatial parameters. Thus, in the proposed quantization codebook, front azimuths with small JND will have small quantization steps, side azimuths with big JND will have big quantization steps, and rear azimuths will have medium quantization steps.

4 Proposed quantization methods of spatial parameters

In SAC, spatial parameters are extracted from channel signals to represent the spatial information of virtual sound source generated by loudspeakers. The most important spatial parameters extracted in SAC schemes often include azimuth of virtual sound generated by multiple loudspeakers and ICLD between channel pairs. The perceptual lossless coding of spatial parameters is the important guarantee of spatial quality for the spatial sound reproduction in decoder.

4.1 Derivation and quantization of azimuth

4.1.1 Estimation of virtual sound source

With multiple correlated sound signals simultaneously rendered by two or multiple loudspeakers, humans can only perceive a summing virtual sound. Regarding the human head as a rigid sphere and without considering the effect of the head on the sound transmission, the sound field around listener generated by loudspeakers can be approximatively equivalent to the sound field generated by a real sound at the location of virtual sound. Given a fixed frequency, if the sound field in the head region can be approximatively reproduced with multiple loudspeakers, then the sounds at a human's two ears in reproduced sound field will be approximatively equivalent to those in the original sound field. In sound reproduction, two essential properties, sound pressure and particle velocity [1], are often used to describe the properties of sound field. The total sound pressure and particle velocity at the listening point (center of listener's head) generated by multiple loudspeakers are approximatively equivalent to those by the single sound (or loudspeaker) at the location of virtual sound. With the peer properties of reproduced sound field by multiple loudspeakers and the sound field by original single sound source, the virtual sound generated by multiple loudspeakers can be taken as the original single sound and estimated from multichannel signals. Since particle velocity is highly correlated with the direction of arrival of sound, the azimuth of virtual sound source can be estimated from loudspeaker signals based on the equations of particle velocity in two sound fields [1].

After 1024-point short time Fourier transform (STFT) with 50 % overlapped window is applied in each frame of each channel signal, the virtual sound source is estimated in each frequency bin. Given loudspeakers with azimuths and distances to the coordinate system's origin in Fig. 4, the total sound particle velocity vector at the coordinate system's origin as the listening point generated by M loudspeakers of frequency bin k can be written as:

$$u = A \begin{pmatrix} \sum_{m=1}^{M} \frac{e^{-ik_{wn}r_m}}{r_m} X_m \cos\theta_m \\ \sum_{m=1}^{M} \frac{e^{-ik_{wn}r_m}}{r_m} X_m \sin\theta_m \end{pmatrix}, \tag{5}$$

where A is the proportionality coefficient relevant to sound transmission, X_m is the amplitude of loudspeaker S_m of frequency k, m is loudspeaker index,$m \in \{1, 2, .., M\}, M \geq 2$, θ_m is the azimuth of loudspeaker S_m, r_m is the distance from loudspeaker S_m to the receiving point, i is imaginary unit, k_{wn} is the wave number, $k_{wn} = \frac{2\pi k}{c}$, k is frequency, and c is sound speed.

Given a sound source S_0 with azimuth θ_0 and distance r_0 to the coordinate system's origin in Fig. 4, the particle velocity vector generated at the coordinate system's origin as the receiving point is written as:

$$u_0 = A \begin{pmatrix} \frac{e^{-ik_{wn}r_0}}{r_0} \cos\theta_0 \\ \frac{e^{-ik_{wn}r_0}}{r_0} \sin\theta_0 \end{pmatrix} X_0. \tag{6}$$

If the virtual sound source generated by loudspeakers is located with azimuth θ_0 and distance r_0 to the coordinate system's origin, then

$$u_0 = u \tag{7}$$

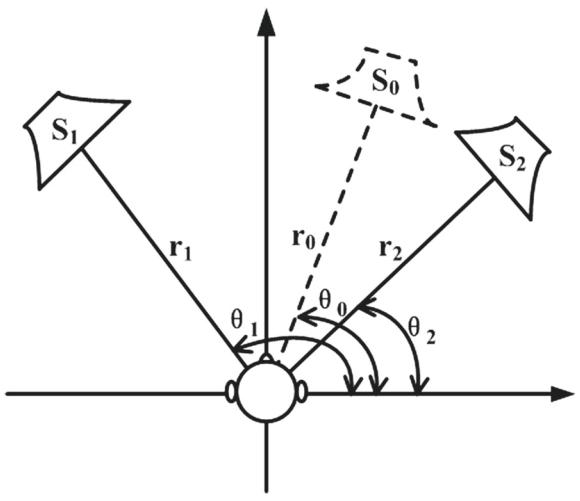

Fig. 4 Virtual sound image generated by loudspeaker pair in the coordinate system

holds. Given the loudspeakers arranged on a sphere surface with the same distance to the coordinate system's origin as the receiving point, the azimuth of the virtual sound can be obtained as [31]:

$$\theta_0 = \arctan\left(\frac{\sum_{m=1}^{M} X_m \sin\theta_m}{\sum_{m=1}^{M} X_m \cos\theta_m}\right). \tag{8}$$

The estimation of virtual sound includes azimuth, as well as the signal estimation of virtual sound. To keep the same energy of virtual sound signal with total energy of original loudspeaker signals, the signal of virtual sound can be estimated as:

$$X_0 = \sum_{m=1}^{M} X_m \frac{\sqrt{\sum_{m=1}^{M} X_m^2}}{\left|\sum_{m=1}^{M} X_m\right|}. \tag{9}$$

Figure 5 illustrated the sound fields (expressed with sound pressure) around listener's head region generated respectively by two loudspeakers and single loudspeaker at the location of estimated virtual sound. It can be deduced from Fig. 5 that the sounds at the listener's two ears in the two sound fields are approximate.

Thus, the multichannel signals are expressed with the virtual sound signal as well as spatial information. The virtual sound signal can be coded with conventional mono perceptual codec, such as AAC, and then transmitted to the decoder. The spatial parameter azimuth will be coded as side information and transmitted to the decoder. The coding method of azimuth is mainly concerned and analyzed in the following subsections.

4.1.2 Quantization of azimuth

For any azimuth in a circle of 360°, if the azimuth is quantized with quantization error below the corresponding azimuthal JND (or MAA), then the azimuth is perceptually losslessly quantized. This is the original intention of the proposed quantization method.

Given the quantization value of some azimuth is C, and the JND of C, then the quantized interval region of C can be computed. All azimuths in the quantized interval region should be quantized as C and the biggest quantization error of all azimuths would not be bigger than their JNDs. If the whole circle of 360° is divided to these quantization interval regions adjacent to each other, then all azimuths can be perceptually losslessly quantized.

The azimuthal JND data can only be obtained by subjective listening tests. The JND data obtained by different listeners are different. The listening environments also influence the test results for the same listener. Thus, the final JND data is often an average among multiple listeners. Meanwhile, owing to the operation limitation, it is impractical to get JND data of all azimuths by practical listening tests. A few of typical azimuths with limited resolution are often used for listening tests. Thus, the JND data obtained from the listening tests often contain a finite number of sparsely distributed values and need to be interpolated to get higher resolution JND data. JND data in our previous research [32] are used in the following experiments. Since the JND data obtained from the tests are angles with one decimal place, JND data are interpolated at a resolution of 0.1° for azimuths from 0°

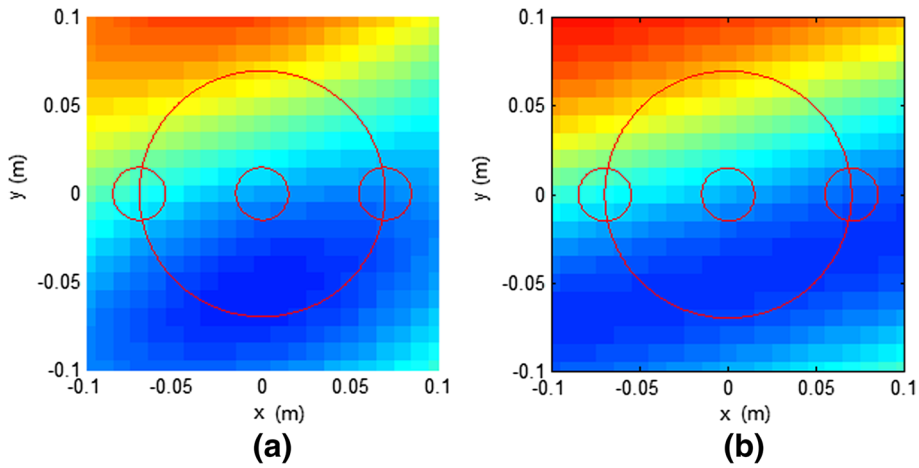

Fig. 5 Sound fields (expressed with sound pressure) in the zone of 0.1 m*0.1 m around mimic human heads (*two big red circles*) in the horizontal plane, different color levels represent different sound pressure levels, and *small red circles* correspond to the center of head and two ears. **a** Sound field generated with two loudspeakers (azimuth 60°, −60°; elevation 0°, 0°; radius 2.5 m, 2.5 m; 1000 Hz signal with gains 0.518 and 1.316, respectively, allocated to two loudspeakers). **b** Sound field generated with one loudspeaker at the estimated location of virtual sound generated by two loudspeakers in (**a**)

(front) to 180° (rear) in the horizontal plane before the quantization codebook design.

JND data of azimuths were often incrementally obtained in subjective audiometries from 0° to 180°. Given an azimuth A_n between 0° and 180°, with the increase of A_n, if the directional difference between the sound with azimuth A_n and $A_{n+1} = A_n + a$ was just noticeable by the listener in the subjective audiometry, JND of azimuth A_n is $JND\_A_n = a$. This can be expressed as $A_{n+1} = A_n + JND\_A_n$ or $A_n = A_{n+1} - JND\_A_n$. With these, we can get the quantization codebook of azimuths with the procedures as follows.

At first, the first quantization value is selected as a start point value. One of the azimuths whose JND is the biggest is selected as one quantization value $C_0 = angle(\max(JND\_A_n))$ and set as the start point. JND_A_n is the JND value of azimuth A_n, $n \in \{0°,0.1°,0.2°,..,180°\}$. In general, $C_0 = 90°$. There are two different procedures from C_0 to get all other quantization values. One procedure begins from C_0 to get the quantization values between C_0 and 180°, named backward procedure. Another procedure begins from C_0 to get the quantization values between 0° and C_0, named forward procedure. In the two procedures, all quantization values between C_0 and 180° and the interval region of each quantization value are obtained.

In the backward procedure, quantization values of azimuths are obtained from C_0 to 180°. For quantization value C_i (C_i=C_0 at the beginning), one of the endpoints I_i ($i \geq 0$) of quantization interval region is calculated as

$$I_i = C_i + JND\_C_i. \tag{10}$$

If $C_i + JND\_C_i > 180°$, then set $I_i = 180°$ and end up the procedure. Afterwards, the quantization value C_{i+1} next to C_i is obtained by

$$C_{i+1} = I_i + JND\_I_i. \tag{11}$$

If $I_i + JND\_I_i > 180°$, then set $C_{i+1} = 180°$ and end up the procedure. In the same way, the other endpoint of quantization interval region for C_{i+1} is calculated as

$$I_{i+1} = C_{i+1} + JND\_C_{i+1}. \tag{12}$$

Then, the interval region of C_{i+1} is $[I_i, I_{i+1})$. All the azimuths in this interval region will be quantized as quantization value C_{i+1}. The backward procedure ends up when I_i or C_i reach 180°.

In the forward procedure, quantization values of azimuths are obtained from C_0 to 0°. For quantization value C'_i ($C'_i = C_0$ at the beginning), the other endpoint I'_i ($i \geq 0$) of quantization interval region can be obtained to meet

$$I'_i = C'_i - JND\_I'_i. \tag{13}$$

If $C'_i - JND\_I'_i < 0°$, then set $I'_i = 0°$ and end up the procedure. Afterwards, the quantization value C'_{i+1} next to C'_i can be obtained to meet

$$C'_{i+1} = I'_i - JND\_C'_{i+1}. \tag{14}$$

If $I'_i - JND\_C'_{i+1} < 0°$, then set $C'_{i+1} = 0°$ and end up the procedure. The forward procedure ends up when I'_i or C'_i reach 0°.

Together with all C_i and C'_i in the above two procedures, we can get the whole quantization codebook for azimuths from 0° to 180°. For the approximate bilateral symmetry, the obtained codebook from one side region (from 0° to 180°) can be duplicated for the other side region (from 0° to −180°) and finally get the azimuthal codebook of the full circle of 360°.

Such procedure can be illustrated in Fig. 6, for example. Generally, $C_0 = 90°$ corresponding to the side of human head. The right endpoint of quantized interval for C_0 is calculated as $I_0 = C_0 + JND\_C_0$. The next quantization value C_1 is calculated as $C_1 = I_0 + JND\_I_0$. Similarly, I_1 is calculated and the quantized interval of C_1 is $[I_0, I_1)$.

4.1.3 Azimuth distribution in quantization codebook

In the proposed JND-based spatial parameters quantization method (JSPQ), the quantization codebooks of spatial parameters are relevant to the direction information of virtual sound source. The quantization codewords densities are dependent on azimuthal JNDs. Small quantization step sizes are assigned for azimuths with small JNDs. Big quantization step sizes are assigned for azimuths with big JNDs. This is the main difference between proposed JSPQ and previous quantization methods of spatial parameters.

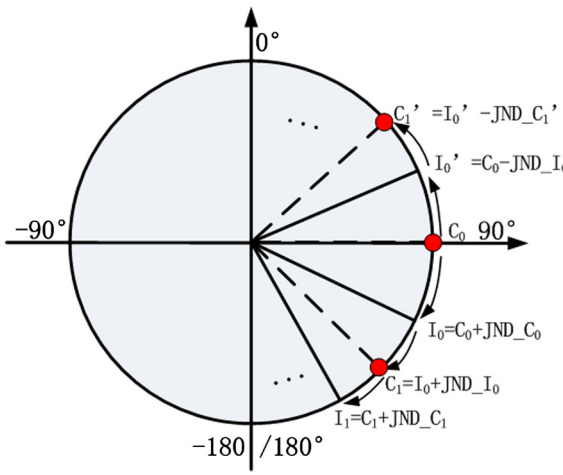

Fig. 6 Demonstration of quantization codebook generation of azimuth. All quantization codewords of azimuths are obtained with JND, trying to control the quantization error of each azimuth in accordance with the JND of the quantized azimuth

Since spatial parameters are to represent the direction information of virtual sound source, the direction information are expressed as azimuths on a circle as illustrated in Fig. 7. With the same symbols as in Fig. 3, each red point on the circle represents one quantization codeword of azimuth, and every two adjacent solid lines form the quantization interval region for the inside quantization codeword. The azimuths on the circle in the quantization interval region will be quantized as the inside quantization codeword.

With essential differences with previous methods in Fig. 3, the illustration of JSPQ in Fig. 7 shows that relative denser quantization codewords are assigned to the front region, fewer quantization codewords are for rear region, and the distribution of quantization codewords for side region are the most rare.

4.1.4 Coding of quantization index

After azimuths are quantized as quantization codewords according to the quantization codebook, the quantization indices of quantization values will be further coded. Owing to the short-time stationary of most sound signals, the sound location information in the same frequency bin in two adjacent time frames will not change rapidly. The difference between the virtual sound azimuths in each two adjacent time frames will fluctuate in a tiny range. Thus, differential Huffman coding will benefit the quantization indices compression. Both natural binary coding and differential Huffman coding are used in JSPQ to code the quantization indices of azimuths. The quantization index of azimuth of each time-frequency bin for the first time frame, *index_q*, is coded by natural binary coding with 6 bits. From the second to the last time frame, the difference of quantization index in each two adjacent time frames is calculated:

$$diff\_index(tf,k) = index\_q(tf,k) - index\_q(tf-1,k), \tag{15}$$

in which tf is the index of time frame, $tf > 1$, k is the index of frequency bins, $k = 0,\ldots 23$, and $index\_q(tf,k)$ is the quantization index for azimuth of frequency bin k in time frame tf. Then, $diff\_index(tf,k)$ is coded with Huffman coding.

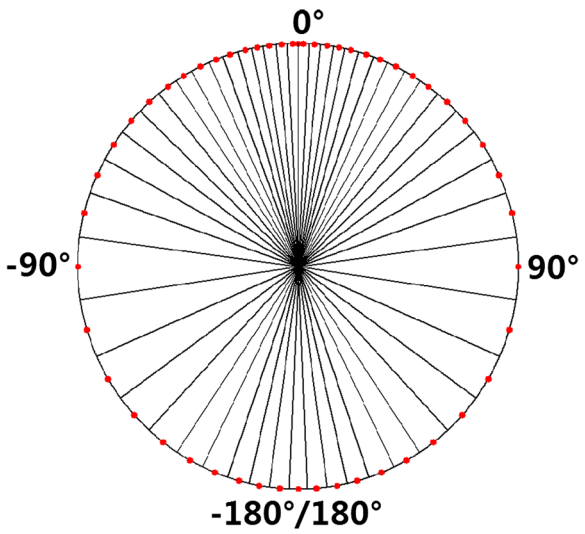

Fig. 7 Quantization codewords distribution of azimuths in proposed JSPQ. Azimuths in the front region have the smallest quantization step sizes. Azimuths in side regions have the biggest quantization step sizes. Azimuths in the rear region have the medium quantization steps

4.2 Quantization of spatial parameter ICLD

In the case of two loudspeakers, both azimuth and ICLD can be extracted from two channel signals as spatial parameters to represent direction information of virtual sound images. Since ICLD can be converted to azimuth, the quantization of ICLD can be easily achieved based on the quantization codebook of azimuth as the process procedure illustrated in Fig. 8.

In the encoder, ICLD can be calculated with two loudspeaker signals by Eq. (1). In each frequency subband, azimuth from Eq. (8) can also be expressed with the signals and azimuths of two loudspeakers as:

$$\theta_0 = \arctan\left(\frac{\sqrt{E_1(b)}\sin\theta_1 + \sqrt{E_2(b)}\sin\theta_2}{\sqrt{E_1(b)}\cos\theta_1 + \sqrt{E_2(b)}\cos\theta_2}\right). \tag{16}$$

With Eqs. (1) and (16), the azimuth of virtual sound θ_0 and ICLD can be converted to each other:

$$\theta_0 = \arctan\left(\frac{10^{\frac{ICLD(b)}{20}}\sin\theta_1 + \sin\theta_2}{10^{\frac{ICLD(b)}{20}}\cos\theta_1 + \cos\theta_2}\right) \tag{17}$$

$$ICLD(b) = 10\log_{10}\left(\frac{\sin\theta_2 - \tan\theta_0\cos\theta_2}{\tan\theta_0\cos\theta_1 - \sin\theta_1}\right)^2. \tag{18}$$

Azimuth θ_0 is quantized according to the quantization codebook of azimuths to get the quantization index. Then, natural binary coding or differential Huffman coding is used to code the quantization index. In the decoder, the azimuth is inversely quantized and converted to retrieve ICLD. Further details are not discussed in this paper (for details, see [28]).

5 Experiments

Objective and subjective experiments were conducted to verify the effectiveness and performance of proposed JND-based spatial parameters quantization method (JSPQ). Reference methods include quantization methods of spatial parameters in MPEG surround [15], Choi's [19], SLQP [22], and Cheng's methods [20].

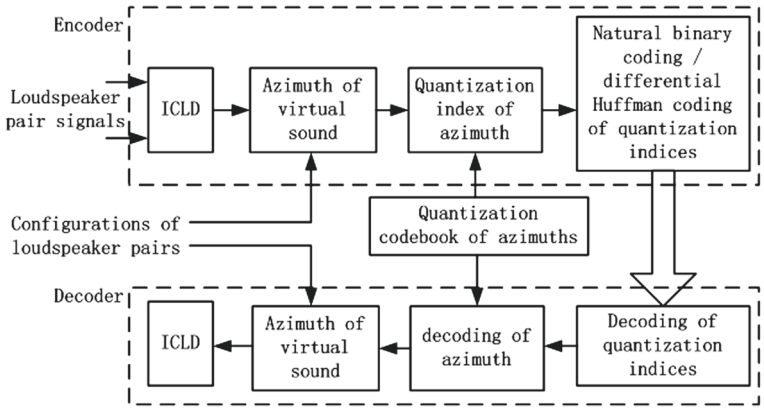

Fig. 8 Framework of ICLD quantization

There are three objective experiments conducted to compare the quantization codebook sizes, spatial distortions, and coding bit rates of spatial parameters in different methods. Subjective experiments aimed to evaluate the spatial quality of different quantization methods for spatial parameters, in which listeners need to rank the perceived spatial quality of audio excerpts encoded and decoded with different quantization methods against the original unprocessed audio excerpts. The specific experimental methods and audio materials are described in detail in the following subsections.

5.1 Comparison of quantization codebook sizes

The codebook sizes of spatial parameters directly determine the bit rates required for transmission of spatial parameters. In the current experiment, the codebook sizes required for quantization of spatial direction information in a circle of 360° at horizontal plane were compared between proposed JSPQ and reference quantization methods.

In MPEG surround and Choi's quantization methods, the main spatial parameters used to represent the spatial direction information are ICLDs. In Cheng's method and SLQP, the spatial parameters used to represent the spatial direction information are azimuths. Thus, reference quantization codebooks include quantization codebooks of ICLDs in MPEG surround and Choi's method, as well as quantization codebooks of azimuths in Cheng's method and SLQP.

Since ICLD is computed from channel pair signals, four channel signals (L, R, Ls, Rs) in standard 5.1 loudspeaker system are pairwise used to extract ICLD in this experiment, as illustrated in Fig. 9. ICLD values are extracted, respectively, with four channel pair signals (L&R, L&Ls, Ls&Rs, R&Rs) to represent spatial direction information in the whole circle of 360° at horizontal plane. The quantization codebooks of ICLD for these four channel pair signals are compared among different methods.

The extraction of azimuth from channel signals does not depend on channel pair signals. The number of channel signals used to extract azimuth of virtual sound can be two or more. In this experiment, the same channel pair signals are used for the extraction of azimuth as for ICLD. The quantization codebooks of azimuths for channel pair signals are compared among different methods.

To represent spatial direction information in a circle of 360° at horizontal plane, the total number of quantization codewords (codebook sizes) for spatial parameters ICLD and azimuth in different methods are illustrated in Fig. 3. Meanwhile, the respective codebook sizes in different regions (front, side, and rear) are also illustrated. As for MPEG surround, SLQP, and Cheng's method, each of them offers two kinds of quantization codebooks: one is coarse quantization with low precision and another is fine quantization with high precision. The quantization

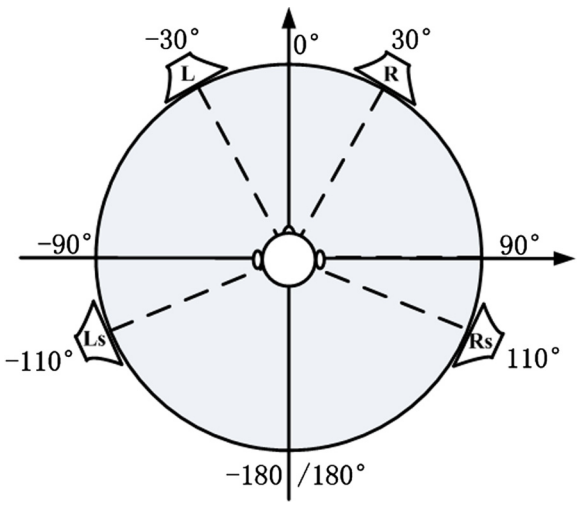

Fig. 9 Configuration of loudspeakers *L*, *R*, *Ls*, and *Rs* in the 5.1 loudspeaker system

codebooks with low and high precision are distinguished with L and H, respectively, in Fig. 10.

The quantization method of spatial parameters in MPEG surround comes from that in E-AAC+ for stereo signals. Thus, the quantization codebooks for spatial parameter ICLD of arbitrary channel pairs are the same, no matter where the loudspeaker pairs are, and no matter how big the intersection angle between two loudspeakers is. Each channel pair signal needs 15 and 30 ICLD values to quantize spatial direction information between two loudspeakers for coarse quantization and fine quantization, respectively. Given four loudspeakers located in a circle, the total numbers of ICLD values to quantize spatial direction information in circle are 60 and 120 for coarse quantization and fine quantization, respectively.

In SLQP, uniform quantization codebooks are used for spatial parameters azimuths. There are two quantization step sizes 3° and 2° for coarse quantization and fine quantization, respectively. Thus, the quantization errors of azimuths will not exceed 1.5° and 1°, which are much smaller than JND values of most azimuths. Significant perceptual redundancies in the quantization for azimuths surely exist, which leads to inefficient compressions of spatial parameters.

In Cheng's two quantization codebooks, the total numbers of azimuths are 32 and 64 for coarse and fine quantization, respectively. The quantization step sizes are nonuniform for azimuths in different directions. It is well known that the JND values of azimuths at the rear are smaller than that at side regions. But in Cheng's two quantization codebooks, contrary to what one might suppose, the total number of quantization azimuths at the rear region is much less than that at side regions. There are only three or seven quantization values for azimuths in the range of 140 degrees from −110° to 110°. These coarse quantization for spatial direction information will definitely result in obvious spatial perceptual distortions.

In the quantization codebook of proposed JSPQ for azimuths, small step sizes are set for front and rear regions, and big step sizes are set for side region. Fewer quantization values are set for azimuths at side regions compared with most of the other methods. The codebook size of JSPQ is smaller than most of other methods, which decreases by 13.3 and 56.7 % compared with two codebooks of MPEG surround and decreases by 56.7 and 71.1 % compared with SLQP's two codebooks.

5.2 Comparison of quantized azimuthal errors

The quantization error of directional information will directly influence the perceptual spatial quality. In this experiment, the absolute quantized errors of azimuths in a half circle of 360° were calculated with different quantization methods. The azimuths from 0° to 180° in horizontal plane (0° corresponds to the front) were chosen with intervals of 0.1° in the calculations. There are 1801 azimuths in total. The absolute quantized error of azimuth is calculated by

$$E_q = \left| A - C_q \right| , \tag{19}$$

in which C_q is the quantized value of azimuth A, $A \in \{0°, 0.1°, 0.2°, \ldots, 180°\}$, $q \in \{0, 1, 2, \ldots, 180\}$.

The local maximums of quantized azimuth error with different methods are computed and compared with azimuthal JND data in Fig. 11. The data of JND curve

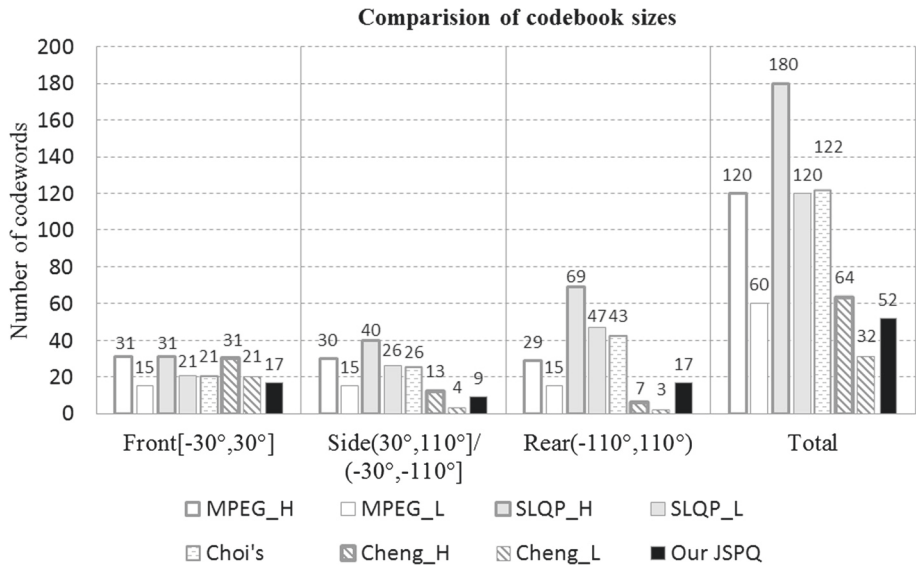

Fig. 10 Codebook sizes comparison of spatial parameters in different methods. The total numbers of codewords for spatial parameters in different regions (*front*, *side*, and *rear regions*), and in a full circle in horizontal plane are shown, respectively

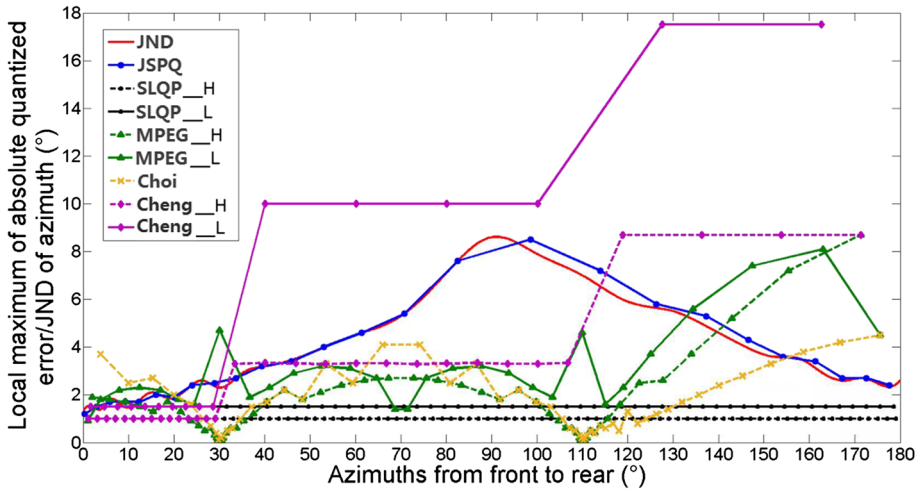

Fig. 11 Quantization errors of virtual sound azimuths in different methods compared with azimuthal JND data. Smaller quantization error than JND means that obvious perceptual redundancy exists; in contrast, bigger quantization error than JND means that sensible spatial distortion exists. Quantization errors of azimuths coincided with azimuthal JND curve means nearly perceptual lossless quantization

come from [32] and are used as references to evaluate the perceptual distortion of azimuthal quantization. Quantized error that is bigger than JND will induce perceptual distortion, and smaller quantized error than JND means perceptual redundancy exists.

Owing to the compact quantization step sizes in SLQP, the quantization errors are the smallest. The quantization errors are much smaller than JND, and it means that obvious perceptual redundancies exist and will lead to inefficient compressions of spatial parameters.

The biggest quantization errors were caused by Cheng's quantization method with lower precision. Most of the quantization errors of azimuths are much bigger than JND. Especially in the rear region, the biggest azimuthal error even exceeds 17°. It will lead to significant spatial distortions with Cheng's methods.

As for other reference methods, most quantization errors in front region are smaller than JND. However, in side regions, most errors are significantly smaller than JND, and errors in rear region are significantly larger than JND.

The proposed JSPQ has small quantization errors in front and rear regions and bigger quantization errors in side regions. The error curve of JSPQ is almost in accordance with the JND curve. Compared with reference methods, the improvement of spatial quality with JSPQ is mainly manifest in the rear region.

5.3 Comparison of bit rates

This experiment aims to compare the coding bit rates of spatial parameters with different quantization methods. The theoretical bit rates calculated according to the codebook sizes of different methods in Fig. 10 and the actual bit rates required to code spatial parameters from standard 5.1-channel signals are both presented.

According to the codebook sizes of different methods in Fig. 10, the theoretical bit rates required to code spatial parameters are calculated and shown in Table 1, which are calculated with simple binary coding without entropy coding methods such as differential Huffman coding. Based on four specific loudspeaker pairs in the front, side (left and right), and rear regions, the theoretical coding bit rates for spatial parameters between several specific channel pairs with different methods are illustrated in Table 1, as well as the bit rates for spatial parameters in a full circle in a horizontal plane with 24 frequency bands and 20 ms time frame.

As for MPEG, the uniform codebook is designed for all channel pairs; thus, the bit rates of MPEG in Table 1 for different channel pairs are the same. With the biggest codebook size, SLQP has the higgest bit rate with nearly 9 kbps for spatial parameters in a full circle in a horizontal plane. Although Cheng_L has the lowest bit rate for the full circle, the bit rates for side and rear channel pairs are so low that obvious spatial distortion will be caused by quantization, as the quantization error presented previously in Fig. 11. Except Cheng_L, the proposed JSPQ has the lowest theoretical bit rate.

Except theoretical bit rates, actual bit rates of spatial parameters for standard 5.1-channel signals are also calculated and presented. Four standard 5.1-channel signals (shown in Table 2) from MPEG and AVS (the advanced audio and video coding standard working group of China) were employed for comparison. The comparison quantization methods include the following: MPEG surround, Choi's method [19], Cheng's method [20], and proposed

Table 1 According to the codebook sizes of different methods in Fig. 10, the theoretical bit rates (kbps) required to code spatial parameters in a full circle in a horizontal plane with 24 frequency bands and 20 ms time frame

Methods	MPEG_H	MPEG_L	SLQP_H	SLQP_L	Choi's	Cheng_H	Cheng_L	Our JSPQ
L&R pair	5.89	4.69	5.95	5.27	5.27	5.95	5.27	4.90
L&Ls/R&Rs	5.89	4.69	6.39	5.64	5.64	4.44	2.40	3.80
Ls&Rs	5.89	4.69	7.33	6.67	6.51	3.37	1.90	4.90
Full circle	8.29	7.09	8.99	8.29	8.32	7.20	6.00	6.84

JSPQ. Since SLQP is mainly applied to multichannel signals with 3D loudspeakers configurations, the alternative method of Cheng's was employed in this experiment. And since Cheng's low resolution quantization method introduces obvious spatial distortion for side and rear regions and has non-comparability with other methods, it is not presented in the experiment.

To encode and decode the 5.1-channel signals, the 5.1-to-mono MPEG surround codec with OTT mode was used in the comparison. Spatial parameter ICLDs are hierarchically extracted and quantized from the six channels signals in MPEG surround codec. In order to focus on the spatial quality and not to be affected by the coding quality of downmix signals, all the downmix signals in comparison were specially coded with 96 kbps AAC codec for high fidelity of basic audio quality. The primary concern of following experiments is the compression performance and spatial quality of spatial parameters. In this comparison experiment, the core quantization of spatial parameters ICLDs were replaced with different quantization methods. For fair comparison, the quantization indices of spatial parameters have been coded with the same inter-frame difference Huffman coding in all methods to further improve compression performance. The total coding bit rates of spatial parameters ICLDs for six channel signals were averaged through time frames for each method and presented in Table 3.

Since the virtual sound distributions in different test audio signals differ greatly, the coding performances of different methods for different test audio signals are quite different as we can see in Table 3. The highest bit rate is provided by MPEG_H which reached 12.21 kbps. The lowest bit rate for all four test signals among different methods is provided by the proposed JSPQ, which is the only one that gets an average bit rate less than 8 kbps. And JSPQ is also the only one of which the highest bit rate for all four test items did not exceed 10 kbps. Compared with MPEG_H, MPEG_L, Choi's, and Cheng's_H, the average bit rate of JSPQ has decreased by 12.77, 9.44, 9.99, and 5.17 %, respectively.

5.4 Subjective evaluation

Subjective experiments were performed to evaluate the spatial quality of proposed quantization method for spatial parameters. Listeners need to rank the perceived spatial quality of audio excerpts encoded and decoded with different quantization methods against the original unprocessed audio excerpts. Spatial parameters ICLD and azimuth are to represent the spatial direction information of virtual sound image between loudspeakers. Thus, the spatial direction and location of virtual sound image are mainly considered in the subjective experiment to evaluate the spatial perceptual distortion caused by quantization of spatial parameters with different methods.

In some experiments on sound source localization, laser pointer-based methods [33–35] are often used to locate the exact location of sound source perceived and pointed by listeners. Since we mainly focus on the difference of perceived virtual sound location between processed and unprocessed multichannel audio excerpts, the listener just needs to tell how about the differences, obvious or not. Considering it is easier for listener to point out if there are differences between perceived locations than exactly pointing to the perceived locations by hands, the standard subjective evaluation scheme MUSHRA of ITU-R BS.1534-1 [36, 37] was employed in this experiment, instead of laser pointer-based methods.

There were seven audio signals used in each MUSHRA test: original audio signal from Table 2 as a reference; original audio signal as a hidden reference; spatial quality degraded original audio signal as a hidden anchor; and test signals processed from original audio signals with

Table 2 Standard 5.1-channel test audio signals

Item names	Source	Content descriptions	Duration(s)
pcm_mps_44khz_HQ.wav	MPEG	Complex electronic music	10.84
AVS_LL6001_48_24_6.wav	AVS	Complex electronic music	8.98
AVS_LL6003_48_24_6.wav	AVS	Complex electronic music	11.00
AVS_LL6004_48_24_6.wav	AVS	Complex electronic music	10.00

Table 3 Coding bit rates (kbps) for spatial parameters with standard 5.1-channel test signals. The average bit rate of JSPQ decreased by 12.77, 9.44, 9.99, and 5.47 % compared with MPEG_H, MPEG_L, Choi's , and Cheng_H, respectively

Item names	MPEG_H	MPEG_L	Choi's	Cheng_H	Our JSPQ
pcm_mps_44khz_HQ.wav	12.21	11.68	11.48	10.87	9.92
AVS_LL6001_48_24_6.wav	9.17	8.98	8.92	8.57	8.14
AVS_LL6003_48_24_6.wav	5.88	5.68	5.92	5.69	5.68
AVS_LL6004_48_24_6.wav	9.16	8.74	8.97	8.47	8.02
Average bit rates	9.10	8.77	8.82	8.40	7.94
Bit rates reduction	12.77 %	9.44 %	9.99 %	5.47 %	—

different spatial parameter quantization methods (four reference methods and proposed JSPQ).

When original audio signals are rendered with six loudspeakers in 5.1 audio system, it may reproduce virtual sound in arbitrary directions in a full circle. Since the intention of our evaluation is about spatial quality, the hidden anchor is a processed signal with degraded spatial quality, instead of 3.5 kHz low pass filtered signal in traditional evaluations for basic audio quality. Downmix of original 5.1-channel signals is a mono signal without spatial information. The channel signals of anchor are the same signal derived from the downmix signal, which can ensure that the virtual sound reproduced with anchor signal will always stay in the middle of active loudspeakers. Since the virtual sound of original signal may be located in arbitrary location among loudspeakers, the spatial quality of anchor signal is severely degraded compared with the original signal. Energy compensation is used for the anchor signal to ensure that the energies of anchor signal and original signal are equal. Thus, the basic audio qualities of anchor signal and original signal are the same, but the spatial information of them are totally different.

To eliminate the effect of basic audio quality on subjective evaluation of spatial quality, the quantization of downmix signals were nearly perceptual lossless for test signals in four reference methods and proposed JSPQ. The basic audio qualities of test signals are almost comparable with the original signal. Owing to the quantization of spatial parameters, the spatial qualities of test signals are different with that of the original signals. The spatial quality about spatial location of the virtual sound is the key concern in our subjective evaluation.

There were a total of 15 male and female graduate students chosen as test listeners in the tests, whose research areas are audio signal processing. They were aged between 22 and 35 and were trained before attending the listening tests. They were trained to mainly focus on evaluating the virtual sound images among different test audio signals. With the original audio signal as a reference, listeners should give scores to evaluate the spatial quality about accuracy of virtual sound location reproduced with different test audio signals coded and processed by different methods. The scoring criteria for subjective evaluation is shown in Table 4. Listeners need to compare the reproduced virtual sound images by processed multichannel audio signals by different methods with that of the original reference signals. The more similar the reproduced virtual sound image is with the original virtual sound image, the higher score should be given to the corresponding method. The scores were filtered to exclude some extreme scores before averaged to get the final results with standard deviations.

Average scores are calculated as:

$$\bar{u}_{jk} = \frac{1}{N}\sum_{i=1}^{N} u_{ijk}, \tag{20}$$

in which u_i is the score of listener i for audio sequence k with method j, and N is the total number of contributing listeners. The confidence interval with confidence coefficient 95 % is:

$$\left[\bar{u}_{jk} - \delta_{jk}, \bar{u}_{jk} + \delta_{jk}\right], \tag{21}$$

in which $\delta_{jk} = t_{0.05}\frac{S_{jk}}{\sqrt{N}}$, $S_{jk} = \sqrt{\sum_{i=1}^{N}\frac{(\bar{u}_{jk}-u_{ijk})^2}{N-1}}$.

The mean scores with 95 % confidence interval of MUSHRA for spatial audio quality of different quantization methods are illustrated in Fig. 12.

The feedback from testers revealed that when virtual sounds were reproduced in the front region, there were little differences among the spatial qualities of virtual sounds

Table 4 Scoring criteria for subjective evaluation of MUSHRA

Score interval	Spatial quality about accuracy of reproduced virtual sound
100–90	Precise
90–80	Tiny bias
80–60	Small deviation
60–40	Big distortion
40–20	Serious distortion
20–0	Error

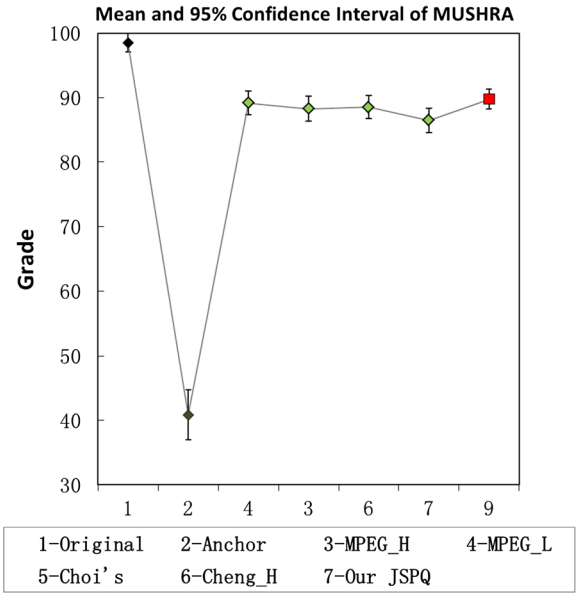

Fig. 12 Subjective evaluation MUSHRA scores for original, anchor references and different quantization methods. The average scores of most reference methods (3, 4, 5) and proposed JSPQ (7) are nearly at equivalent levels

by different quantization methods. This is probably due to that for the front region from −30° to 30°, the quantization codebooks of spatial parameters in most methods are fine enough to avoid obvious spatial perceptual distortion.

In addition, it was found that when virtual sounds were reproduced in the side regions, most reference methods did not introduce obvious spatial perceptual distortion for their enough fine quantization codebooks. Although the quantization precision of JSPQ in the side region was lower than in the front region, since the quantization errors were controlled to below perceptual JND, obvious spatial perceptual distortions were not perceived in the subjective listening tests.

As for the virtual sounds reproduced in the rear region in subjective listing tests, listeners reflected that the perceptual distortions of Cheng's methods were the most serious. This is probably attributed to the big quantization error of 35° for azimuths behind the listeners in Cheng's methods.

JSPQ assigned quantization values for spatial parameters in all directions according to the azimuthal perceptual JND. Thus, although with the lowest coding bit rate for spatial parameters, insignificant spatial distortion existed for JSPQ in all regions. As illustrated in Fig. 12, JSPQ has a comparative score with most of reference methods.

5.5 Summary of experiments

There are four respects of coding performance are analyzed by above objective and subjective experiments: quantization codebook sizes, quantization errors and coding bit rates of spatial parameters, as well as subjective MUSHRA evaluations of coding performance.

In the respect of quantization codebook sizes, it requires only 52 quantization codewords in JSPQ for the the quantization of azimuths in a full circle in horizontal plane. The quantization codebook size of JSPQ is obviously smaller than most of reference methods, for example, it has decreased by 13.3 and 56.7 % than two quantization codebooks of MPEG surround.

Meanwhile, the quantization errors of spatial parameters in JSPQ are more consistent with the JNDs of azimuths in a full circle than all of the reference methods. It means that the advantage of JSPQ is to balance between quantization error and quantization redundancy.

In respect of coding bit rates for standard test signals, the average bit rate of JSPQ to quantize spatial parameters in a full circle is lower than most of the reference methods.

In addition, subjective MUSHRA results testified that the quantization of JSPQ did not introduce obvious spatial perceptual distortion. The average MUSHRA score of JSPQ is almost comparable to the reference methods such as MPEG surround.

To sum up, a good trade-off between bit rates of spatial parameters and spatial quality is obtained with JSPQ.

5.6 Limitation and future work

According to the above objective and subjective experiments, the overall performance of JSPQ was better than the reference methods. Nevertheless, there were still some limitations in the proposed JSPQ and conducted experiments.

Although the proposed JSPQ also applies the spatial parameters quantization in different horizontal planes with different elevation angles, the implementation process and related experiments were not discussed in this paper. In 3D audio, the virtual sounds distribute widely in 3D sound field. In 3D audio, it will contain a large amount of spatial information, which is much more than in 2D horizontal plane. Thus, the application of JSPQ will achieve more significant gain in 3D audio multichannel signals coding and will be implemented in future.

In this paper, we only focused on the most important and generally used spatial parameters ICLDs and azimuths. Besides, other stereo parameters such as ICTD are also mainly related to the direction perception of virtual sound. ICTDs can be mapped to azimuth of virtual sound too. Thus, the JND characteristics of directional perception of human auditory system can be also used in the perceptual quantization of ICTDs, the perceptual redundancy removals of these spatial parameters can be the emphases of next research work.

6 Conclusions

In spatial audio coding, the spatial information of virtual sounds generated by loudspeakers are extracted as spatial parameters, as well as downmix signals are obtained from loudspeaker signals. In the case of multichannel loudspeaker systems, the virtual sounds may be widely distributed in 3D space. It contains a large amount of spatial information to achieve vivid spatial sound effects. Thus, the accurate representation of virtual sound and the efficient compression of spatial parameters are the key to the perfect reconstruction of spatial sound effects.

Human auditory system has the perceptual limitations of just noticeable difference (JND) characteristics for spatial sound location estimation. If the spatial distortion caused by quantization errors of spatial parameters can be limited below, the correlated JND threshold, then the quantization can be regarded as nearly perceptual lossless.

However, the quantization step sizes of spatial parameters in current SAC methods are not consistent well with the JND characteristics of different azimuths in a full circle. It will result in either spatial perceptual distortions of virtual sound images, or inefficient compressions of spatial parameters.

Therefore, in this paper, the characteristic of azimuthal JND was effectively used to design the quantization of spatial parameters in multichannel signals. In the proposed quantization codebook of JSPQ, the quantization step sizes of azimuths are assigned according to the JND values of azimuths in a full circle. Different azimuths from front region to rear region have different quantization step sizes.

Objective experiments and subjective evaluations confirmed the coding performance of proposed JSPQ compared with reference quantization methods of spatial parameters in respect of codebook sizes, quantization errors, coding bit rates, and spatial qualities.

The quantization codebook size of JSPQ was 13.3 and 56.7 % lower than two quantization codebooks of MPEG surround. Average bit rate reduction on spatial parameters for standard 5.1-channel signals reached up to approximately 9 and 13 % compared with MPEG Surround, while preserving comparable subjective spatial quality.

For future work, we plan to implement the framework on top of state-of-the-art 3D audio processing platforms. In particular, we are interested in advanced technologies for perceptual sensitivity analysis of complex multivariate spatial parameters [38, 39] as well as promising computing schemes for optimal 3D audio reproduction in sophisticated applications in practice [40–42].

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work has been supported by the National Nature Science Foundation of China (61231015, 61471271), National High Technology Research, and Development Program of China (863 Program, 2015AA016306).

Author details

[1]State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China. [2]National Engineering Research Center for Multimedia Software, the Key Laboratory of Multimedia and Network Communication Engineering, Computer School, Wuhan University, Wuhan, China. [3]Collaborative Innovation Center of Geospatial Technology, Wuhan, China.

Received: 20 January 2016 Accepted: 5 May 2016
Published online: 21 May 2016

References

1. A Ando, Conversion of multichannel sound signal maintaining physical properties of sound in reproduced sound field. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1467–1475 (2011)
2. G Sergi, Knocking at the door of cinematic artifice: Dolby Atmos, challenges and opportunities. New Soundtrack. **50**(10), 107–121 (2013)
3. DB Ward, TD Abhayapala, Reproduction of a plane-wave sound field using an array of loudspeakers. IEEE Trans. Speech Audio Process. **9**(6), 697–707 (2001)
4. S Spors, R Rabenstein, J Ahrens, in *Proceedings of the 124th Convention of the Audio Engineering Society (AES)*. The theory of wave field synthesis revisited (Audio Engineering Society, Amsterdam, The Netherlands, 2008)
5. S Disch, C Ertel, C Faller, et al, in *Proceedings of the 117th Convention of the Audio Engineering Society (AES)*. Spatial audio coding: next-generation efficient and compatible coding of multi-channel audio (Audio Engineering Society, San Francisco, CA, USA, 2004)
6. F Rumsey, Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. J. Audio Eng. Soc. **50**(9), 651–666 (2002)
7. M Bosi, RE Goldberg, *Introduction to digital audio coding and standards*. (Springer Science & Business Media, Dordrecht, Netherlands, 2012)
8. RM Hershkowitz, NI Durlach, Interaural time and amplitude jnds for a 500-Hz Tone. J. Acoust. Soc. Am. **46**(6), 1464–1465 (1969)
9. JC Blauert, *Spatial Hearing: The psychophysics of human sound localisation*. (MIT Press, Cambridge, MA, US, 1997)
10. AW Mills, On the minimum audible angle. J. Acoust. Soc. Am. **30**(4), 237–246 (1958)
11. J Breebaart, S van de Par, A Kohlrausch Schuijers, Parametric coding of stereo audio. EURASIP J. Appl. Signal Process. **9**, 1305–1322 (2005)
12. JW Strutt, *The theory of sound*. (Dover Publications, Mineola, NY, US, 1877)
13. F Baumgarte, C Faller, Binaural cue coding—part I. Psychoacoustic fundamentals and design principles. IEEE Trans. Speech Audio Process. **11**(6), 509–519 (2003)
14. C Faller, F Baumgarte, Binaural cue coding—part II: schemes and applications. IEEE Trans. Speech Audio Process. **11**(6), 520–531 (2003)
15. J Breebaart, C Faller, *Spatial audio processing: MPEG surround and other applications*. (Wiley, Hoboken, NJ, US, 2008)
16. ISO/IEC Standard 14496-3, Technical description of parametric coding for high quality audio (2005)
17. 3GPP TS 26.401, General audio codec audio processing functions; enhanced aacPlus general audio codec; General Description (2004)
18. T Wu, R Hu, L Gao, et al, *Analysis and comparison of inter-channel level difference and interaural level difference, multimedia modeling*. (Springer International Publishing, Miami, 2016), pp. 586–595
19. SJ Choi, Y Jung, HJ Kim, H Oh, in *Paper presented in Proceedings of the 120th Convention of the Audio Engineering Society (AES)*. New CLD quantisation method for spatial audio coding, (Paris, France, 2006)
20. B Cheng, C Ritz Burnett, Psychoacoustic-based quantisation of spatial audio cues. Electron. Lett. **44**(18), 1098–1099 (2008)
21. I Elfitri, B Guenel, AM Kondoz, Multichannel audio coding based on analysis by synthesis. Proc. IEEE. **99**(4), 657–670 (2011)
22. B Cheng, C Ritz, I Burnett, X Zheng, A general compression approach to multi-channel three-dimensional audio. IEEE Trans. Audio Speech Lang. Process. **21**(8), 1676–1688 (2013)
23. J Engdegard, B Resch, C Falch, et al, in *Proceedings of the 124th Convention of the Audio Engineering Society (AES'08)*. Spatial audio object coding

(SAOC)-the upcoming MPEG standard on parametric object based audio coding (Audio Engineering Society, Amsterdam, The Netherlands, 2008)
24. ISO/IEC JTC1/SC29/WG11 International Standard ISO/MPEG 23008-3, 3D Audio, Geneva (2015)
25. DR Perrott, Role of signal onset in sound localization. J. Acoust. Soc. Am. **45**(2), 436–445 (1969)
26. DR Perrott, K Saberi, Minimum audible angle thresholds for sources varying in both elevation and azimuth. J. Acoust. Soc. Am. **87**(4), 1728–1731 (1990)
27. DW Grantham, BWY Hornsby, EA Erpenbeck, Auditory spatial resolution in horizontal, vertical, and diagonal planes. J. Acoust. Soc. Am. **114**(2), 1009–1022 (2003)
28. L Gao, R Hu, Y Yang, X Wang, W Tu, T Wu, *Azimuthal perceptual resolution model based adaptive 3D spatial parameter coding, multimedia modeling*. (Springer International Publishing, Sydney, 2015)
29. V Pulkki, M Karjalainen, Multichannel audio rendering using amplitude panning [DSP applications]. Signal Process. Mag. IEEE. **25**(3), 118–122 (2008)
30. J Daniel, S Moreau, R Nicol, in *Proceedings of the 114th Convention of the Audio Engineering Society (AES)*. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging (Audio Engineering Society, Amsterdam, The Netherlands, 2003)
31. S Ke, X Wang, L Gao, T Wu, Y Yang, in *16th Pacific-Rim Conference on Multimedia (PCM), Gwangju, Korea*. Physical properties of sound field based estimation of phantom source in 3D, (2015)
32. H Wang, C Zhang, R Hu, W Tu, X Wang, *The perceptual characteristics of 3D orientation, multimedia modeling*. (Springer International Publishing, Dublin, 2014)
33. P Majdak, R Baumgartner, B Laback, Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. Front. Psychol. **5**(319), 1–10 (2014)
34. R Baumgartner, P Majdak, Modeling localization of amplitude-panned virtual sources in sagittal Planes. J. Audio Eng. Soc. **63**(7/8), 562–569 (2015)
35. P Majdak, MJ Goupell, B Laback, 3-D localization of virtual sound sources: effects of visual environment, pointing method, and training. Atten. Percept. Psychophys. **72**(2), 454–469 (2010)
36. ITU-R Rec. BS.1534-2, *Method for the subjective assessment of intermediate quality level of coding systems(MUSHRA)*. (International Telecommunications Union, Geneva, 2014)
37. M Schoeffler, F Stoter, B Edler, J Herry, in *1534 (MUSHRA),1st web audio conference*. Towards the next generation of web-based experiments: a case study assessing basic audio quality following the itu-r recommendation bs (Audio Engineering Society, Paris, France, 2015)
38. D Chen, X Li, D Cui, L Wang, D Lu, Global synchronization measurement of multivariate neural signals with massively parallel nonlinear interdependence analysis. IEEE Trans. Neural Syst. Rehabil. Eng. **22**(1), 33–43 (2014)
39. D Chen, X Li, D Cui, L Wang, S Khan, J Wang, K Zeng, C Cai, Fast and scalable multi-way analysis of massive neural data. Comput. IEEE Trans. **64**(3), 707–719 (2015)
40. D Li, R Hu, X Wang, S Yang, W tu, in *16th Pacific-Rim Conference on Multimedia (PCM)*. Multichannel simplification based on deviation of loudspeaker positions (Audio Engineering Society, Gwangju, Korea, 2015), pp. 544–553
41. D Li, R Hu, X Wang, W Tu, S Yang, in *15th Pacific-Rim Conference on Multimedia (PCM)*. Automatic multichannel simplification with low impacts on sound pressure at ears (Audio Engineering Society, Kuching, Malaysia, 2014)
42. D Chen, L Wang, A Zomaya, M Dou, J Chen, Z Deng, S Hariri, Parallel simulation of complex evacuation scenarios with adaptive agent models. Parallel Distributed Syst. IEEE Trans. **26**(3), 847–857 (2015)