Perceptual characteristic and compression research in 3D audio technology

Ruimin Hu, Shi Dong, Heng Wang, Maosheng Zhang, Song Wang, Dengshi Li

National Engineering Research Center for Multimedia Software

School of Computer Science, Wuhan University, Wuhan, 430072, China

hrm1964@163.com, edisonds@gmail.com, wh825554@163.com, eterou@163.com,

wangsongf117@163.com, reallds@126.com

Abstract. The 3D audio coding forms a competitive research area due to the standardization of both international standards (i.e. MPEG) and localized standards (i.e. Audio and Video Coding Standard workgroup of China, AVS). Perception of 3D audio is a key issue for standardization and remains a challenging problem. Besides current solutions adopted from traditional audio engineering, we are working for an original 3D audio solution for compression. This paper represents our initial results about 3D audio perception include directional measurement of Just Noticeable Difference (JND) and Perceptual Entropy (PE). We also represent the possible applications of these results in our future researches.

Keywords: 3D audio, perceptual audio processing, audio compression

1 Introduction

With the current trend of 3D movies and the popularization of 3DTV, 3D audio and video technology has become a research topic in multimedia technology. To provide the audience with a more immersive and integrated audio-visual experience, audio must work collaboratively with 3D video to provide three dimensional sound effects. However, existing 3DTV and 3D movie systems usually adopt conventional stereo audio and surround sound technology, which only provides very limited sound localization ability and envelopment in horizontal plane. Although there is not a generally acknowledged definition for 3D audio, it is widely accepted that 3D audio must have the following characteristics; localization of sound image in arbitrary direction in 3D space, realizing the distance perception of sound and giving a improved feeling of audio scene. Nowadays two types of technology are able to satisfy the requirement of 3D audio, one is based on physical principles and aims at reconstructing the original sound field, the other is based on principle of human perception and aims at giving the listener a virtual sound image. Wave Field Synthesis (WFS), Ambisonics and 22.2 multichannel systems are three typical 3D audio systems following those principles.

This paper is arranged as follows. In section 2 an introduction to the three 3D audio systems is presented and the existing problems are discussed, where we conclude the complexity of the 3D systems and efficiency of the signal compression will be two problems for the popularization of 3D audio. In section 3 we present our related work in 3D audio technology, including hearing mechanism and signal

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

compression research. More specifically, we investigate the JND of the direction perception cues for human in horizon plane. This is useful in simplification the 3D audio recording and playback systems, and removing the redundant perceptual information in 3D audio signals. In section 4 the development trends of 3D audio and our future work are discussed.

2 Brief view of typical 3D audio systems

2.1 Wave Field Synthesis (WFS)

a. The Principle of Wave Field Synthesis

The concept of WFS was introduced by Berkhout in 1988[1], its physical theory can date back to Huygens principle which suggests that a wave which propagates from a given wave front can be considered as emitted either by the original sound source or by a secondary source distribution along the wave front [2]. To reconstruct the primary sound field, the distribution of secondary source can replace primary source. The concept was later developed by Kirchhoff and Rayleigh, and the Kirchhoff-Helmholtz integral they proposed can be interpreted as follows: if appropriately secondary sources are driven by the values of the sound pressure and the directional pressure gradient caused by the virtual source on the boundary of a closed area, then the wave field within the region is equivalent to the original wave field[3]. By adding a degree of freedom to the secondary source distribution, Kirchhoff-Helmholtz generalized Huygens principle.

b. Realization of WFS

According to the above theory, WFS reproduces the primary sound field in time and space by making using of small and individually driven loudspeakers array, and can recover the spatial image precisely in the half space of receiving end from loudspeaker arrays[4].

But there is some limit for WFS in application. WFS needs a continuous, closed surface and a large number of idealized loudspeakers, but in practice there is only a discontinuous loudspeaker array. According to spatial nyquist sampling Theorem, if the interval between loudspeakers is less than half the wave length of a sound wave, aliasing will not occur[5].

So according to spatial nyquist sampling Theorem, WFS can be realized by limited and discrete loudspeakers within a certain frequency range. For example, limited line loudspeaker with even intervals can reconstruct sound field in 2D horizontal plane[6]. In the recording stage, the listening area is surrounded by a microphone array. The microphone array consists of pressure and velocity microphones, which record the primary sound field of external sound sources. In the reconstruction stage, the microphones will be replaced by the loudspeakers. Each loudspeaker is driven by signal recorded by the corresponding microphone. The geometric shape of the microphone array and loudspeaker are the same[7].

2.2 Ambisonics

a. The principle of Ambisonics

Ambisonics emerged in the 1970's and the main contributor is Gerzon [8] The principles of Ambisonics are as follows. A certain wave (sound field) can be expand on a sphere in sphere coordinate system by spherical harmonic functions. At the opposite end, superposition of spherical harmonic functions can rebuild a wave (sound field). There are $n=2m+1$ spherical harmonic functions at every order m of Ambisonics, a 3D system of M order consists of all spherical harmonic functions at every order m ($0 \leqslant m \leqslant M$), total channel number N satisfies $N=(M+1)^2$.

b. Two simple format of Ambisonics

The first format of Ambisonics proposed by Gerzon is B format, which displays an omnidirectional sound field by four channels: W, X, Y, Z [9]. Traditional monophony and stereophony can be seen as the subsystems of Ambisonics[10]. Sound location in horizontal plane is realized using three channels W, X, Y, and the fourth channel Z is used for reconstructing height information. Channel W is a pressure signal, and X, Y, Z are directional signal. B-format is used in studio and professional application.

The second format of Ambisonics is UHJ system which can convert directional sound into two or more channels and solve the incompatibility problem of four channels Ambisonics with monophony, stereophony [11][12]. The coding scheme provided by UHJ can be used in broadcasting, digital audio recording[13].

c. Playback technology of Ambisonics

According to the principle of Ambisonics, the decomposition of a sound field requires the expansion of infinite order spherical harmonic functions. But in practical application, limited order truncation of spherical harmonic functions expansion is necessary. B-format is one order expansion. Ambisonics was expanded to high order in the 1990's, the sweet point was enlarged to an area. High order Ambisonics promotes sound location with the price of more channels and loudspeakers. We can get better reconstruction quality using higher order Ambisonics. The encoding process of Ambisonics is to preserve the result of spherical harmonic functions multiplying the signal picked up by microphones. The decoding process is to calculate a group of loudspeaker signals according to the rebuilt sound field that must be equal to the primary sound field at listening point. This can be done by solving the inverse matrix which consists of spherical harmonic functions that are associated with locations of loudspeakers.

2.3 22.2 multichannel sound systems

a. Fundamentals of multichannel sound systems

The research of spatial hearing and sound source localization indicates that there are slight time and level differences between two ears when spatial sound signals arrive at the ears. For the estimation of direction and distance of sound source, the difference between the two ears signals is most relevant. Actually these differences, called binaural cues, are Interaural Time Difference (ITD) and Interaural Level

Difference (ILD). ILD and ITD indicate the level difference and time difference between left and right ears respectively [14].

b. Stereo, 5.1 surround sound and 22.2 multichannel system

The binaural localization theory is utilized in stereo system, i.e. time and level differences between signals from two loudspeakers are utilized in sound reproduction in order to reconstruct the spatial perception of the audience.

Traditional stereo cannot provide the sense of encirclement and immersion because the perception of the sound environment mainly relies on the lateral reflected sound. Surround sound, which constitutes an extension of stereophony, provides full spatial immersion by using reverberation and reflection. The most typical multichannel surround systems are the Dolby surround system, DTS Digital Surround.

Since loudspeakers in Dolby 5.1 are arranged in the same horizontal plane, the reproduction sound image cannot be extended to three dimensions. In 2009, Dolby laboratory presented ProLogic IIz, which extended Dolby 7.1 with height channels (7.1+2). By reproducing early and late reflections and reverberation, ProLogic IIz provide a much wider range of spatial sound effects such as spatial depth and spatial impression [15]. The ProLogic IIz configuration is showed in Figure 1. Audyssey Dynamic Surround Expansion (DSX) is a scalable technology that expands auditory perception by adding height channels, which is in a similar way to Dolby 9.1.

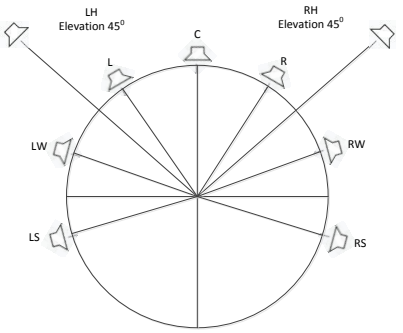

Fig. 1. Dolby IIz configuration

NHK laboratory developed the 22.2 multichannel prototype system in 2003. The system consists of three layers of loudspeakers and overcome the lack of height perception with 3D immersion and sound image localization. K. Hiyama and Keiichi Kubota evaluated the minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field respectively[16]. The results showed that if the interval between adjacent loudspeakers is 45° in both horizontal and vertical plane, there is sufficient horizontal sound envelopment and a good sense of spatial impression. Therefore, the 22.2 multichannel system consists of loudspeakers with a middle layer of ten channels, an upper layer of nine channels, and a lower layer of three regular channels and two Low Frequency Effects (LFE) channels. Figure 2 shows detailed arrangement of loudspeakers[17]. The vertical loudspeaker interval of the 22.2 multichannel is around 45°, which can induce the vertical spatial uniformity [18]. The 22.2 multichannel system reproduces sound images in all three dimensional directions around a listener and stable sound localization over the entire screen area. Subjective evaluations shows that subjects have better impressions using

Ultrahigh-Definition TV (UHDTV) contents with 22.2 multichannel sound system than with Dolby5.1 system[19].

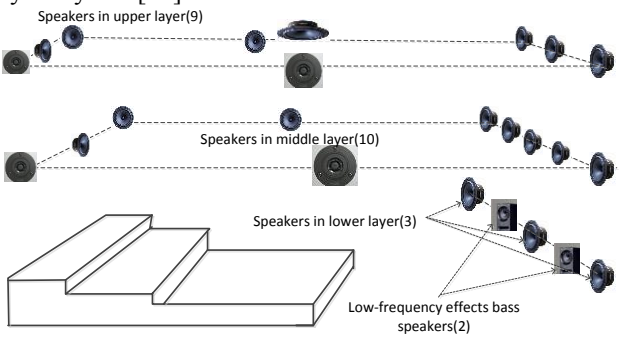

Fig. 2. 22.2 multichannel system layout

2.4 Problems of existing 3D audio systems

Not need to know the loudspeaker layout at the encoding stage is the main advantage of Ambisonics, at decoding stage the loudspeaker signal can be counted according to the loudspeaker layout. The encoding format is an effective reconstruction of 3D sound field, allowing for direct dealing with the three dimensional space characteristics of the sound field such as rotation and mirroring. But along with the increase of order, more precise direction information is carried by spherical harmonic functions, which provides a more accurate location. But data quantity increases rapidly, which requires higher CPU processing power. In addition, the hypothesis that the location of the listener is known may lead to a limit listening area.

The character of WFS is that Kirchhoff-Helmholtz integral can ensure the rebuilt sound field synthesized by secondary sources is the same as the primary source, preserving time and space characteristics of primary source. So listeners can receive and locate the sound source as if it were a real listening space, and walk in the listening area at will while sound image remains unchanged. But WFS needs more loudspeakers and has a higher requirement for site and equipment which is expensive.

The research on compression of Ambisonics and WFS is limited, although recently some progress[20][21][22] has been made. But the compression efficiency cannot meet the requirement of real-time broadcasting and transmitting.

The 22.2 multichannel system, which is based on conventional surround systems plus high and low channels to produce three dimensional sound images, can be easily downmixed for 5.1 system reproduction. It is likely to become a popular 3D system since terminals can be set up with little cost using simplified configuration (10.1 and 8.1 channels), especially when the 5.1 system has already been installed. In 2011, ITU (Report BS.2159-2) pointed out that the 22.2 multichannel system has some problems to be solved: The method to localize more efficiently by using the upper and lower layers and how to reproduce three dimensional sound image movements. In addition, although it is not difficult to downmix 22.2 channel signals to 5.1 channel signals, the 3D spatial audio effects are discarded. Hence, producing three dimensional effects in

home entertainment environments with limited loudspeakers is a problem. Furthermore, without compression, the data rate of 22.2 system can reach 28Mbps and the size of an one-hour audio file is about 100Gb. As a result, it is not possible for the current storage device and transmission channel to adapt to this enormous data. The application and development of 22.2 multichannel systems are constrained by the technology of compression.

3. Hearing mechanism and compression research in 3D audio

3.1 The research of hearing mechanism

From mono, stereo, surround sound, and then to the 3D audio, the main line of development in audio systems is to extend the range of the sound image. Audiences are able to locate the sound which is any position around them in order to bring them a better sense of encirclement and immersion. The positioning of spatial orientation for sound sources is an important content of 3D audio, while the study of perceptual characteristics is an important research field of 3D audio. For example, the arrangement position of the 24 speaker in 22.2-channel system is based on the test and analysis of the angle resolution of sound in horizontal and vertical plane by human ear. In addition, the perceptual research of spatial orientation parameters for sound source is also important for the efficient encoding of the multi-channel audio signal. Therefore, the perceptual characteristics of sound source localization parameters in the 3D sound field are an important way to solve the problems of 3D audio systems.

The perceptual sensitivity of the sound source in the horizontal plane is significantly better than that of the vertical plane or distance by the human auditory system. In the horizontal plane, the positioning of the sound source is dependent on the two binaural cues: ITD and ILD. The human ear can perceive a change in sound image orientation only when the difference of binaural cues reaches a certain threshold value. This threshold value is known as Just Noticeable Difference (JND).The influencing factors of JND for binaural cues are various, including frequency and orientation of the sound source. A wide range of measurements and analysis of these factors has been performed.

Hershkowitz in 1969 [23] and Mossop in 1998 [24] have been researching the influence of sound source position on the perceptual threshold JND of ITD and ILD. The results show that the greater the difference of left and right channels in intensity and time, the larger the JND value of the human perception. This shows that the human ear is less sensitive when the sound source is closer to the left and right sides.

Millers in 1960 measured JNDs of ILD on the midline with pure tones and there were 5 Normal-Hearing (NH) subjects took part in the experiment[25]. The result is as follows: JNDs were around 1dB for 1000Hz, around 0.5dB for frequencies higher than 1000Hz and somewhat smaller than 1dB for frequencies lower than1000Hz. The test data showed worse sensitivity of ILD at 1000Hz than at either higher or lower frequencies. Larisa in 2011 has been researching the influence of the frequency of the

signal on the JND of ITD. The results showed that the perceptual threshold of ITD has a strong dependence on the frequency[26].

The measurement data of JND for binaural cues were fragmented and the conclusions were generally described qualitatively for perceptual threshold of binaural cues. It is difficult to perform mathematical analysis and model accurately and cannot fully reveal the principal of the perceptual threshold of binaural cues. So the JND measurement of binaural cues in all-round, full-band and the mathematical analysis are important issues to reveal the perceptual characteristics of binaural cues. In order to solve the above problem, we have undertaken the research of perceptual characteristics for binaural cues:

In order to study the impact of the frequency and direction on binaural cues JND, our team measured full band JND of binaural cues and analyzed its statistics and distribution characteristics.

a. *Subjects*. 12 NH subjects participated in this study, 7 males and 5 females, all subjects were aged between 19 and 25 years.

b. *Stimuli*. The method in this article used a two-alternative-forced-choice paradigm to measure the JND. Both reference and test signals were 250 ms in duration including 10 ms raised-cosine onset and offset ramps. They were randomly combined into stimulus and separated by 500 ms duration. The stimuli were create by personal computer and presented to the subjects over headphones (Sennheiser HDA 215) at a level of 70 dB SPL. In order to exclude other factors influence on this experiment, the environment of the entire testing process should be consistent and the intensity of test sound must remain around 70 dB SPL. Meanwhile the ITD should be zero in the whole experiment in order to remove the effect on the result caused by other binaural cues and the sum of energy of left and right channels should remain unchanged.

The reference values of ILD in these experiments were 0, 1, 3, 5, 8 and 12 dB, which respond to 6 azimuths(about 0~60 °) in the horizontal plane from midline to the direction of the left ear.

The whole frequency domain was divided into 20 sub-bands, each frequency sub-band satisfied the same perceptual characteristics of human ear.

The stimuli are pure tones whose frequencies are the center frequencies of sub-bands, these frequencies are 75, 150, 225, 300, 450, 600, 750, 900, 1200, 1500, 1800, 2100, 2400, 2700, 3300, 4200, 5400, 6900, 10500, 15500Hz.

c. *Method*. Discrimination thresholds were estimated with an adaptive procedure. During any given trial, subjects would listen to two stimuli by activating a button on a computer screen by mouse-click, with a free number of repeats but the order of two stimulus were changed. The subjects should indicate which one was lateralized to the left relatively by means of an appropriate radio button response in 1.5 s.

An adaptive, 1-up-3-down method was also used in this article. The difference of ILD in dB was increased in every one wrong or decreased in every three consecutive correct judgments. The difference between reference and test signals in first trials was the initial variable which was much larger than the target JND, it was changed by an given step according to previous test results.

The step was changed adaptively, it was adjusted by 50% for the first two reversals, 30% for the next two reversals, then linear changed in a small step size for

the next three reversals, until the final step size reach the expected accuracy for the last three reversals. In a transformed-up-down experiment, the stimulus variable and its direction of change depends on the subjects responses. The direction alternates back and forth between “down” and “up”. Every transform between “down” and “up” was defined as a reversal.

Because of heavy workload of these experiments, adaptive test software was designed to simplify the experiments and the process of data collection and analysis. The software automatically generated test sequences and played one after another. According to the listener’s choice, the software changed ILD values of test stimulus properly, and saved the results to excel sheet until listener hardly distinguished the orientation differences between two sequences. And the value of ILD at this time was the JND value.

d. *Results*. After a subjective listening test for half a year, we got 120 groups(six azimuths and twenty frequencies) of data, each group containing 12 JNDs corresponding to 12 subjects. For every group, we select the data that has the confidence degree of 75% to be JND in that condition. Some JND curves in different reference of ILD were plotted in figure 3:

•The curves vary with the reference ILD, the larger the reference ILD, the higher the corresponding curve. The JND is the most sensitive in the central plane for human perception, and the least sensitive at lateral.

•Human ear is most sensitive to the middle frequency bands except 1000 Hz and less sensitive to the high frequency bands and low frequency bands.

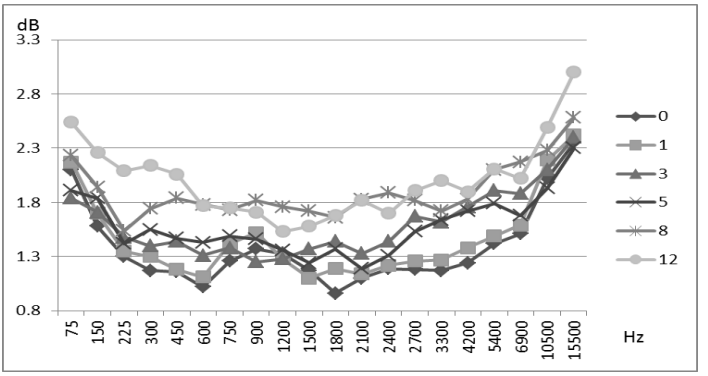

Fig. 3. JND curve of ILD

A binaural perceptual model is established and used in quantisation of ILD. It solves the problem of the perceptual redundancy removal of spatial parameters. Experimental results show that this method can reduce the bitrate by about 15% compared with parametric stereo, while maintaining the subjective sound quality.

3.2 Perceptual information measurement for multichannel audio signal

Multimedia contents abound with subjective irrelevancy—objective information we cannot sense. For audio signals, this means lossless to the extent that the distortion after decompression is imperceptible to normal human ears (usually called transparent coding). The bitrate can be much lower than for true lossless coding. Perceptual audio coding [27] by removing the irrelevancy greatly reduces communication bandwidth or storage space. Psychoacoustics provides a quantitative theory on this irrelevancy: the limits of auditory perception, such as the audible frequency range (20–20000 Hz), the Absolute Threshold of Hearing (ATH), and masking effect[28]. In state-of-the-art perceptual audio coders, such as MPEG-2/4 Advanced Audio Coding (AAC), 64 kbps is enough for transparent coding[29]. The Shannon entropy cannot measure the perceptible information or give the bitrate bound in this case.

For perceptual audio coding technology, determining the lower limit bitrate for transparent audio coding is an important question. Perceptual Entropy (PE) gives an answer to this question[30], which shows that a large amount of audio with CD quality can be compressed with 2.1 bit per sample. PE indicates the least number of bits for quantising mono audio channel without perceptual distortion. This is widely used in the design of quantisers and fast bit allocation algorithm.

Nevertheless, PE has significant limitations when measuring perceptual information. This limitation primarily comes from the underlying monaural hearing model. Humans have two ears to receive sound waves in a 3D space: not only is the time and frequency information perceived— needing just individual ears—but also spatial information or localization information—needing both ears for spatial sampling. Due to the unawareness of binaural hearing, PE of multichannel audio signals is simplified to the supposition of PE of individual channels. This is significantly larger than real quantity of information received because multichannel audio signals usually correlate.

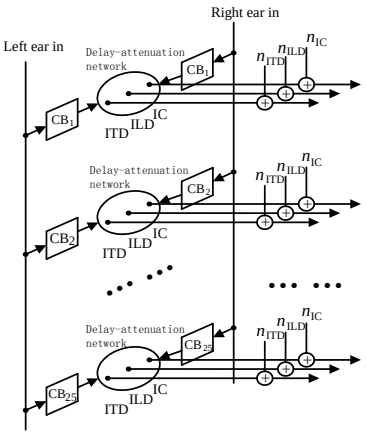

Fig. 4. Binaural Cue Physiological PerceptionModel (BCPPM).

Following the concept of PE, we establish a Binaural Cue Physiological Processing Model (BCPPM, figure 4). Based on MCPPM, we using EBR filter to simulate the human cochlea filter effect, and the JND of binaural cues to estimate the absolute threshold of spatial cues.

a. *SPE Definition*. From the information theory viewpoint, we see BCPPM as a double-in-multiple-out system (Figure 4). The double-in is the left ear entrance sound and the right ear entrance sound. The multiple-out consists of 75 effective ITDs, ILDs, and ICs (25 CBs, each with a tuple of ITD, ILD, and IC). Like in computing PE, we view each path that leads to an output as a lossy subchannel. Then there are 75 such subchannels. Unlike PE, what a subchannel conveys is not a subband spectrum but one of ITD, ILD, and IC of the subband corresponding to the sub-channel. In each sub-channel, there are intrinsic channel noises (resolution of spatial hearing), and among sub-channels, there are interchannel interferences (interaction of binaural cues). Then there is an effective noise for each sub-channel. Under this setting, each sub-channel will have a channel capacity. We denote SPE(c), SPE(t), and SPE(l) for the capacity of IC, ITD, and ILD sub-channels respectively. Then SPE is defined as the overall capacity of these sub-channels, or the sum of capacities of all the sub-channels:

$$SPE = \sum_{all\ subbands} SPE(c) + SPE(t) + SPE(l) \tag{1}$$

To derive SPE(c), SPE(t), and SPE(l), we need probability models for IC, ITD, and ILD. Although the binaural cues are continuous, the effective noise quantizes them into discrete values. Let [L • P], [T • P], and [C • P] denote the discrete ILD, ITD, and IC source probability spaces:

$$\begin{aligned}
[\mathbf{L}\bullet P]&:\begin{cases} \mathbf{L}: & l_1, & l_2, & \cdots, & l_i, & \cdots, & l_N \\ P(\mathbf{L}): & P(l_1), & P(l_2), & \cdots, & P(l_i), & \cdots, & P(l_N) \end{cases} \\
[\mathbf{T}\bullet P]&:\begin{cases} \mathbf{T}: & t_1, & t_2, & \cdots, & t_i, & \cdots, & t_N \\ P(\mathbf{T}): & P(t_1), & P(t_2), & \cdots, & P(t_i), & \cdots, & P(t_N) \end{cases} \\
[\mathbf{C}\bullet P]&:\begin{cases} \mathbf{C}: & c_1, & c_2, & \cdots, & c_i, & \cdots, & c_N \\ P(\mathbf{C}): & P(c_1), & P(c_2), & \cdots, & P(c_i), & \cdots, & P(c_N) \end{cases}
\end{aligned} \tag{2}$$

where l_i, t_i, and c_i are the ith discrete values of ILD, ITD, and IC, respectively, and $P(l_i)$, $P(t_i)$, and $P(c_i)$ the corresponding probabilities. Then we have

$$\begin{aligned}
SPE(l) &= -\sum_{i=1}^{N_L} p(l_i)\log_2 p(l_i) \\
SPE(t) &= -\sum_{i=1}^{N_T} p(t_i)\log_2 p(t_i) \\
SPE(c) &= -\sum_{i=1}^{N_C} p(c_i)\log_2 p(c_i)
\end{aligned} \tag{3}$$

b. *CB Filterbank*. We use the same method as that in PE to implement the CB filterbank. Audio signals are first transformed to the frequency domain by DFT of 2048 points with 50% overlap between adjacent transform blocks. Then a DFT spectrum is partitioned into 25 CBs.

c. *Binaural Cues Computation*. ILD, ITD, IC are computed in the DFT domain as described in [31].

d. *Effective Spatial Perception Data* . The resolutions or quantization steps of the binaural cues can be determined by JND experiments. Denote by $\Delta\tau$, $\Delta\lambda$, and $\Delta\eta$ the

resolutions of ITD, ILD, and IC, respectively. Generally, they are signal dependent and frequency dependent. For simplicity, we use constant values: $\Delta\tau = 0.02$ ms, $\Delta\lambda =$ 1dB, and $\Delta\eta = 0.1$.

We ignore the effect of IC on ILD and only consider the effect of IC on ITD for SPE computation. Lower IC leads to lower resolution of ITD. This is equivalent to higher JND of ITD. Then the effective JND on subband b, denoted as $\Delta\tau'$ (b), can be formulated as the following:

$$\Delta\tau'(b) = \frac{\Delta\tau(b)}{\mathrm{IC}(b)} \tag{4}$$

Then we have the following effective perception data q_{ILD}(b), q_{ITD}(b), and q_{IC}(b) of ILD, ITD, and IC, respectively by quantization, where $\lfloor \cdot \rfloor$ represents the round down function:

$$\begin{aligned} q_{\mathrm{ILD}}(b) &= 2\left\lfloor \left| \frac{\mathrm{ILD}(b)}{\Delta\lambda(b)} \right| \right\rfloor \\ q_{\mathrm{ITD}}(b) &= 2\left\lfloor \left| \frac{\mathrm{ITD}(b)}{\Delta\tau(b)/\mathrm{IC}(b)} \right| \right\rfloor \\ q_{\mathrm{IC}}(b) &= \left\lfloor \frac{1\text{-}\mathrm{IC}(b)}{\Delta\eta(b)} \right\rfloor \end{aligned} \tag{5}$$

Suppose that $q_{\mathrm{ILD}}(b)$, $q_{\mathrm{ITD}}(b)$, and $q_{\mathrm{IC}}(b)$ are uniformly distributed by (3), the SPE are expressed as

$$\begin{aligned} SPE = \frac{1}{N}\sum_{b=1}^{25} & \alpha \log_2\left(\mathrm{int}\left(\frac{1-\mathrm{IC}(b)}{\Delta\eta(b)} \right) + 1 \right) + \alpha \log_2\left(2\,\mathrm{int}\left(\left| \frac{\mathrm{ITD}(b)}{\Delta\tau(b)/\mathrm{IC}(b)} \right| \right) + 1 \right) \\ & + \alpha \log_2\left(2\,\mathrm{int}\left(\left| \frac{\mathrm{ILD}(b)}{\Delta\lambda(b)} \right| \right) + 1 \right) \end{aligned} \tag{6}$$

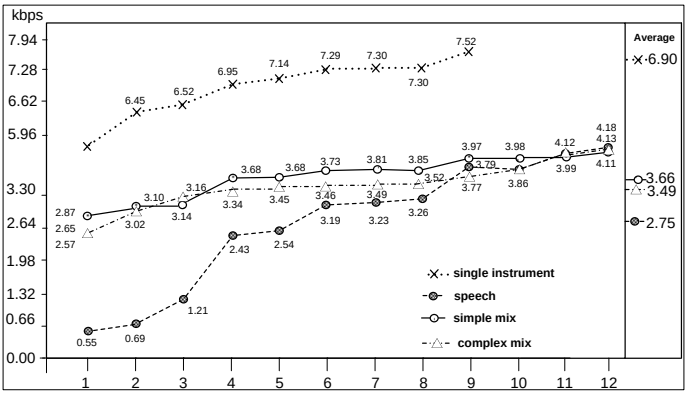

Fig. 5. Perceptual spatial information of stereo sequences sampled at 44.1 kHz.

d. *Results*. Figure 5 shows the SPE of four different stereo signals from MPEG test sequences. The experiment suggests that SPE of speech signal is very low. This is because the human voice is often recorded with fixed position without change. So coding this kind of stereo audio signals requires a low bit rate. The average SPE for

speech signals is 2.75kbps, for simple mixed audio is 3.66kbps, for complex mixed audio is 3.49kbps and for a single instrument is 6.90kbps. In other words, to achieve transparent stereo effect, audio signals required more than 7kbps, which is close to the bitrate 7.7kbps of PS. So the proposed SPE can reflect the amount of perceptual spatial information that is ignored by PE. Experiments on stereo signals of different types have confirmed that SPE is compatible with the spatial parameter bitrate of PS.

Using PE to evaluate the perceptual information, only interchannel redundancy and irrelevancy are exploited; the overall PE is simply the sum of PE of the left and right channels. Using SPE based on BCPPM, interchannel redundancy and irrelevancy are also exploited; the overall perceptual information is about one normal audio channel plus some spatial parameters, which has significantly lower bitrate. For the above reason, PE gives much higher bitrate bound than SPE. PE is compatible with the traditional perceptual coding schemes, such as MP3 and AAC, in which channels are basically processed individually (except the mid/side stereo and the intensity stereo). So PE gives meaningful bitrate bound for them. But in Spatial Audio Coding (SAC), multichannel audio signals are processed as one or two core channels plus spatial parameters. SPE is necessary in this case and generally gives much lower bitrate bound (~1/2). This agrees to the sharp bitrate reduction of SAC.

4. Tendency of 3D audio technology and our future work

4.1 Hearing mechanism research on 3D audio

The spatial orientation cues of sound include three aspects: azimuth angle, elevation angle and distance. There are many acoustic factors to perceive the distance of a sound source, such as the source of the sound (sound pressure level and spectrum), the transmission environment (reflected sound, high-frequency losses and environmental noise) as well as listening factors. So the current research focuses on the expression and extraction of distance cues. Hence, the perceptual characteristic of the 3D spatial orientation is an important research direction for 3D audio technology.

Our future work will focus on the perceptual characteristics of 3D spatial orientation. The main work will include: design experiments to obtain perceptual threshold of 3D spatial position, mathematical analysis to establish representation model of perceptual sensitivity in 3D spatial orientation, get the perceptual distortion of sound image in the different offset of spatial orientation, obtain the equivalent distortion curve of azimuth angle and elevation angle in 3D spatial orientation, and to establish a position distortion model of 3D spatial position. Through the above research, we expect to establish the basic theory of perceptual mechanism for 3D audio systems and provide theoretical support for 3D audio collection, processing, reconstruction, playback and evaluation.

4.2 High efficiency compression for 3D audio signal

Existing 3D audio compression technology has exploited the perceptual redundancy within each individual channel. From the same sound field and same sound source, 3D audio signals of different channels intrinsically exhibit strong correlation. Parametric coding is able to extract the cues of sound image direction, width and scene information to reduce the interchannel redundancy, and achieve high compression efficiency using fewer channels with side information. Parametric coding for 3D audio is able to fulfill the compression requirement of transmission and storage while keep 3D effect meantime, so it is a strong direction in 3D audio compression research.

Since the compression is highly efficient, the reconstructed 3D effect strongly depends on the cues that described corresponding spatial information. The existing 3D audio parameter coding quantises those cues uniformly and reconstruction error in every direction is the same. However, according to human perceptual characteristic in 3D space, the JND to sound direction exists and varies widely in all directions. If reconstruction error for direction cues exceed corresponding threshold, perceptible 3D effect distortion is produced. So how to utilize human perceptual characteristics in 3D space for 3D audio parametric coding will be included in our future work. Our goal is to develop the 3D spatial perception information measurement and establish a computational model of 3D audio orientation perception for effective representation of 3D audio parameterization

4.3 The evaluation of 3D audio quality

Along with the developments of the 3D audio technology, research institutions such as NHK [32] and Deutsche Telekom Laboratories[33], are carrying out the subjective evaluation of the 3D audio system. Because the subjective evaluation is based on the human who is the main body directly involved in the evaluation, the result is more explicit and reasonable in spite of spending a lot of time and manpower during the period of the assessments. So, more and more scholars[34][35][36] are trying to establish the objective evaluation model for the 3D audio system, hoping to look for an objective evaluation model based on the human perception of the audio quality to assess the effects of a 3D sound field. The performance of the proposed model is comparable with the subjective evaluation method.

However, the current methods used to establish an objective evaluation model do not introduce the spectral cues related to the elevation perception of sound events, the envelopment or immersion in diffuse sounds, or the proximity and distance of sound events as the acoustic characteristic parameters. Research of the objective evaluation methods of the 3D audio is occuring on to investigate the spectral cues of the elevation, envelopment and distance perception of the 3D sound field.

In the study of the objective evaluation method of the 3D audio quality, we draw up an objective evaluation model, based on the acoustic characteristic parameters of a 3D audio signal, to predict the perceptual acoustic attributes of the 3D sound field. Including the Basic Audio Quality (BAQ), the Timbral Fidelity (TF), the 3D Frontal Spatial Fidelity (3DFSF) and the 3D Surround Spatial Fidelity (3DSSF). The study includes establishing the acoustic characteristic parameter set related to the 3D

perceptual sound field, obtaining a predictable mapping of the perceptual acoustic attributes and the acoustic characteristic parameters of a 3D audio quality, and building up an objective evaluation model of the 3D perceptual sound field by fitting the performances of the subjective evaluation and objective evaluation. Because the main aim of this study is to express the spectral cues related to the elevation perception of a 3D sound field, we should try to analyze the duplex spectral effects of the pinna to further improve the technology of the 3D audio objective evaluation.

5. Conclusion

The complexity and large capacity limit the promotion and application of 3D audio. To solve these problems, the National Natural Science Foundation of China, Tsinghua University, Wuhan University and other colleges organized the Second International Symposium of 3D video and audio. In the 3D audio workshop, basic theory and research on the recording, compression and reconstruction for 3D audio was emphasized. We also hope to promote the research work to become part of the next generation standard for the audio and video coding (AVS2) of China. This paper gives a brief introduction on current 3D audio systems. At the same time, our research work on the hearing mechanism and compression coding are presented. Finally our future work is introduced, which includes the research of perception characteristic, compression coding and the quality evaluation.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.60832002, No.61102127), Major national science and technology special projects (2010ZX03004-003-03), Nature Science Foundation of Hubei Province (2010CDB08602, 2011CDB451), Wuhan ChenGuang Science and Technology Plan (201150431104), and the Fundamental Research Funds for the Central Universities.

References

1. Berkhout, A.J.: A holographic approach to acoustic control. Journal of the Audio Engineering Society. 36, 977-995 (1988)
2. Berkhout, A., De Vries, D., Vogel, P.: Acoustic control by wave field synthesis. J. Acoust. Soc. Am. 93, 2764-2778 (1993)
3. R. Rabenstein, S.S., P. Steffen: Wave field synthesis techniques for spatial sound reproduction. In: Topics in Acoustic Echo and Noise Control, pp. 517-545. Springer, Berlin, Heidelberg (2006)
4. De Vries, D.: Wave Field Synthesis: History, State-of-the-Art and Future (Invited Paper). In: Universal Communication, 2008. ISUC '08. Second International Symposium on, pp. 31-35.

(2008)
5. De Bruijn, W.:Application of wave field synthesis in videoconferencing. Delft University of Technology (2004)
6. Vogel, P.:Application of Wave Field Synthesis in Room Acoustics. Delft University of Technology (1993)
7. Daniel, J.M., Sebastien; Nicol, Rozenn: Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging. In: Audio Engineering Society Convention 114. Amsterdam, The Netherlands (2003)
8. Gerzon, M.A.: Ambisonics: Part two: Studio techniques. Studio Sound. (1975)
9. Malham, D.G.: Spatial hearing mechanisms and sound reproduction. University of York. (1998)
10. Furness, R.K.: Ambisonics-an overview. In: 8th International Conference: The Sound of Audio, pp. 181-189. (1990)
11. Keating, D.: The generation of virtual acoustic environments for blind people. In: Proc.1st Euro. Conf: Disability, Virtual Reality & Assoc. Tech., pp. 201-207. Maidenhead, UK (1996)
12. Elen, R.: Whatever happened to Ambisonics? AudioMedia Magazine, November. (1991)
13. Gerzon, M.A.: Ambisonics in multichannel broadcasting and video. J. Audio Eng. Soc. 33, 859-871 (1985)
14. Strutt, J.W.: On our perception of sound direction. Philosophical Magazine. 13, 214-232 (1907)
15. Theile, G., Wittek, H.: Principles in Surround Recordings with Height. In: Audio Engineering Society Convention 130. (2011)
16. Hiyama, K., Komiyama, S., Hamasaki, K.: The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. Audio Engineering Society Convention 113. (2002)
17. Ando, A.: Home Reproduction of 22.2 Multichannel Sound. In: 5th International Universal Communication Symposium. (2011)
18. Oode, S., Sawaya, I., Ando, A., Hamasaki, K., Ozawa, K.: Vertical Loudspeaker Arrangement for Reproducing Spatially Uniform Sound. In: Audio Engineering Society Convention 131. (2011)
19. Hamasaki, K., Nishiguchi, T., Okumura, R., Nakayama, Y., Ando, A.: A 22.2 multichannel sound system for ultrahigh-definition TV (UHDTV). Smpte Motion Imaging Journal. 117, 40-49 (2008)
20. Cheng, B., Ritz, C., Burnett, I.: A Spatial Squeezing approach to Ambisonic audio compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 369-372. IEEE, (2008)
21. Hellerud, E., Solvang, A., Svensson, U.P.: Spatial redundancy in Higher Order Ambisonics and its use for lowdelay lossless compression. In: Acoustics, Speech and Signal Processing, 2009(ICASSP 2009). IEEE International Conference on, pp. 269-272. (2009)
22. Pinto, F., Vetterli, M.: Space-Time-Frequency Processing of Acoustic Wave Fields: Theory, Algorithms, and Applications. Signal Processing, IEEE Transactions on. 58, 4608-4620 (2010)
23. Hershkowitz, R., Durlach, N.: Interaural Time and Amplitude JNDs for a 500 - Hz Tone. The Journal of the Acoustical Society of America. 46, 1464-1465 (1969)
24. Mossop, J.E., Culling, J.F.: Lateralization of large interaural delays. The Journal of the

Acoustical Society of America. 104, 1574-1579 (1998)
25. Mills, A.W.: Lateralization of High‐Frequency Tones. The Journal of the Acoustical Society of America. 32, 132-134 (1960)
26. Dunai, L., Hartmann, W.M.: Frequency dependence of the interaural time difference thresholds in human listeners. The Journal of the Acoustical Society of America. 129, 2485-2485 (2011)
27. Painter, T., Spanias, A.: Perceptual coding of digital audio. Proceedings of the IEEE. 88, 451-515 (2000)
28. Moore, B.C.J.: Masking in the Human Auditory System. In: Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction. Audio Engineering Society, New York, USA (1996)
29. Bosi, M., Goldberg, R.E.: Introduction to digital audio coding and standards. Kluwer Academic Publishers, Boston, Mass, USA (2003)
30. Johnston, J.D.: Transform coding of audio signals using perceptual noise criteria. Selected Areas in Communications, IEEE Journal on. 6, 314-323 (1988)
31. C. Faller, F. Baumgarte. :Binaural cue coding—part II: schemes and applications. IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 520–531 (2003)
32. Hamasaki, K.H., Koichiro; Nishiguchi, Toshiyuki; Okumura, Reiko: Effectiveness of Height Information for Reproducing the Presence and Reality in Multichannel Audio System. In: Audio Engineering Society Convention 120. Paris, France (2006)
33. Geier, M., Wierstorf, H., Ahrens, J., Wechsung, I., Raake, A., Spors, S.: Perceptual evaluation of focused sources in wave field synthesis. In: AES 128th Convention, pp. 22-25. (2010)
34. George, S.:Objective models for predicting selected multichannel audio quality attributes. Department of Music and Sound Recording, University of Surrey (2009)
35. Epain, N., Guillon, P., Kan, A., Kosobrodov, R., Sun, D., Jin, C., Van Schaik, A.: Objective evaluation of a three-dimensional sound field reproduction system. In: Proceedings of 20th International Congress on Acoustics. Sydney, Australia (2010)
36. Song, W., Ellermeier, W., Hald, J.: Psychoacoustic evaluation of multichannel reproduced sounds using binaural synthesis and spherical beamforming. The Journal of the Acoustical Society of America. 130, 2063-2075 (2011)