EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Expanded three-channel mid/side coding for three-dimensional multichannel audio systems

Shi Dong, Ruimin Hu*, Xiaochen Wang, Yuhong Yang and Weiping Tu

Abstract

Three-dimensional (3D) audio technologies are booming with the success of 3D video technology. The surge in audio channels makes its huge data unacceptable for transmitting bandwidth and storage media, and the signal compression algorithm for 3D audio systems becomes an important task. This paper investigates the conventional mid/side (M/S) coding method and discusses the signal correlation property of three-dimensional multichannel systems. Then based on the channel triple, a three-channel dependent M/S coding (3D-M/S) method is proposed to reduce interchannel redundancy and corresponding transform matrices are presented. Furthermore, a framework is proposed to enable 3D-M/S compress any number of audio channels. Finally, the masking threshold of the perceptual audio core codec is modified, which guarantees the final coding noise to meet the perceptual threshold constraint of the original channel signals. Objective and subjective tests with panning signals indicate an increase in coding efficiency compared to Independent channel coding and a moderate complexity increase compared to a PCA method.

Introduction

Recently, 3D audio has attracted more attention and developed fast following the booming market of 3D movie. Many 3D audio technologies are now introduced into audio-involved applications to replace the surround sound system to provide superior sound localization and an immersive feeling. Wave field synthesis (WFS), Ambisonics and vector-based amplitude panning (VBAP) are the three most well-developed technologies. WFS generally follows Huygens principle to reconstruct the original sound field [1]. Research institutions such as IDMT of Fraunhofer and IRCAM in France have an intensive study in WFS, and attempt to bring WFS into theater and live transmission of concert. Ambisonics utilizes spherical harmonic functions to recording sound field and driving loudspeakers, its loudspeakers have rigorous configuration and give a good sound field reconstruction in the center [2]. VBAP follows the tangent law in a three-dimensional space using three adjacent loudspeakers to form a sound vector. For its simplicity, VBAP is the most common algorithm in 3D signal panning [3]. A 3D system like 22.2 multichannel system proposed by NHK in Japan utilizes VBAP to generate 3D sound image [4]. The 22.2 multichannel system is also included in the developing MPEG-H standard for rendering 3D audio scene.

There is a clear trend that 3D audio technology will become mature gradually and replace stereo and surround sound [5]. However, a main and common feature of 3D audio technologies is the great number of sound channels. For instance, WFS system always contains dozens and even hundreds of audio channels. The 22.2 system has three layers and 24 audio channels. Although the Ambisonics system can have flexible order and channel number, it usually uses dozens of channels because fewer channels will cause quality deterioration. Comparing with a two-channel stereo and a 5.1 surround sound, the increasing of audio channel causes a dramatical 3D audio data increase. A report from Fraunhofer shows 37 Mbps is needed for live transmission of WFS [6]. For the 22.2 multichannel system, uncompressed data also reaches 28 Mbps [7]. Currently, storage media and transmission bandwidth can hardly afford those huge data size. So the compression of 3D multichannel audio signals becomes an important subject.

The well-known Spatial Audio Coding (SAC) models the signals as virtual sound sources in the frequency domain, extracts the interchannel level difference (ICLD)

*Correspondence: hurm1964@gmail.com
National Engineering Research Center for Multimedia Software, Computer School, Wuhan University, Wuhan, China

and interchannel time difference (ICTD) and interchannel coherence (IC) to represent the direction and width of virtual sound source and downmixes the multichannels to reduce redundancy [8-11]. The idea of using downmixed sources with spatial parameters was later developed into Spatial Audio Object Coding (SAOC) for efficiently coding the multiple input spatial audio objects with interactive and personalized rendering ability [12]. Recently, some other investigations have been published to increase the compression efficiency for multichannel 3D audio signals. In 2007, Goodwin and Jot proposed a PCA-based multichannel compression framework for parametric coding [13], which can enhance specific audio scenarios and provide robust spatial audio coding. In 2008, Cheng et al. proposed the Spatially Squeezed Surround Audio Coding (S^3AC) for parametrically compressing the Ambisonics signal [14]. In 2009, Hellerud used an inter-channel prediction-based coding method to remove the redundancy between Ambisonics channels [15], which has low algorithm delay but high computational complexity. Tzagkarakis used a sinusoidal model and linear prediction to parameterize the separate spot microphone channels, then downmixed the residual signals. This coding scheme is more suitable for multichannel signals with weak correlation, and such scenarios require Independent channel decoding [16]. In 2010, Pinto et al. utilized a space/time-frequency transform to decompose the WFS signals into plane waves and evanescent waves. By discarding the evanescent waves and perceptually coding the plane wave signals, coding gain is obtained. Coding efficiency increases along with the number of audio channels, because the transform decomposition accuracy depends on the spatial resolution which is the number of WFS channels [17,18]. In 2013, Cheng further proposed a Spatial Localization Quantization Point (SLQP) codec using localization cues to compress the 3D audio signals [19,20]. Since SLQP extracts the spatial cues and downmixes the channels, it achieved high compression ratio for SLQP signals and other 3D audio systems.

In order to increase the coding efficiency at high bitrates, some non-parametric coding schemes were developed. Yang proposed a scalable multichannel codec, using the Karhunen-Loeve Transform (KLT) to remove the interchannel redundancy to realize scalable multichannel audio coding [21]. Mid/side (M/S) coding was introduced by J.D. Johnston [22] and adopted by many audio codec such as MPEG2-Layer III and MPEG4-AAC. In 2003, Liu et al. proposed a bit allocation method for M/S coding based on allocation entropy, which increases the objective quality by allocating more bits to high energy channel in M/S coding [23]. In 2008, Derrien et al. proposed an error model for M/S coding. The error model enables tuning of the quantizer used for channels M and S at the encoder with respect to the distortion of L and R at the decoder side, which increased the coding efficiency of M/S without much complexity [24]. Since M/S coding works as the simplest interchannel prediction, Krueger generalized it using linear prediction instead of M/S transformation and residual signal instead of difference signal [25]. In 2012, Schafer further developed Krueger's method, the multichannel case, which has low algorithmic delay [26]. Recently, M/S coding was combined with parametric stereo coding at low bitrates in the MPEG-USAC standard [27] by predicting the residual channel using spatial cue-based parameters, which aimed to bridge the stereo quality gap between low bitrates and high bitrates [28]. M/S coding also works alone at high bitrates utilizing a novel complex prediction to achieve better performance [29].

The above model-based codec and parametric codec can offer a considerable compression ratio. However, those methods need to know the direction of the real audio source to do objective-oriented coding, or estimated a virtual source direction to do downmixing and parametric coding. In practice, such as live recording, it is very difficult to obtain the real audio source direction. Downmixing and parametric coding will cause interchannel interference such as 'tone leakage' artifacts when channel signals differ greatly [30]. Furthermore, the computational complexity of an audio codec should be acceptable while maintaining enough coding efficiency, and parametric coding can only achieve a performance gain at low bitrates. This paper focuses on the situation that only the multichannel signals of audio sources are recorded, instead of their directions. And we consider high-quality/high-bitrate application and focus on the non-parametric coding method. Section 'M/S coding in 3D space' describes the conventional M/S coding process and presents a three-channel Dependent M/S coding (3D-M/S) method. The main idea is to expand M/S coding to three-dimensional audio by designing a new transform matrix, which remove the redundancy of three channels in 3D space rather than just two channels in the horizontal plane. Section '3D-M/S psychoacoustic model' discusses the psychoacoustic model for transformed 3D-M/S signals. Section 'Framework for general channel configuration' specifies a new framework enables 3D-M/S to be applied to a more general channel configuration. Section 'Experiment' gives a comparison of 3D-M/S coding with PCA coding and Independent channel coding to justify the performance of compression ratio and computational complexity. Section 'Conclusion' summarizes and concludes this paper.

M/S coding in 3D space

Conventional M/S coding

M/S coding is based on the fact that most stereo channels are strongly correlated. By simply transforming the stereo channel pair into the M/S domain, core codec encodes a summation channel and a difference channel instead of the original channels. The difference channel has much lower energy than the original channel, so more frequency bins can be quantized to similar and smaller quantized values, which leads to the entropy of the resulting quantised time-frequency samples is lower and hence lossless coding using Huffman coding achieves a higher compression rate. To illustrate how M/S coding works, a generalized sine stereo model is used. Here, a stereo pair is denoted as a vector $\mathbf{V_0} = (C_L, C_R)$ where

$$\begin{aligned} C_L &= S\sin\theta \\ C_R &= S\cos\theta \end{aligned} \tag{1}$$

S is the virtual audio source, θ is the stereo panning angle and $\theta \in \left[0, \frac{\pi}{2}\right]$. The M/S coding can be denoted as two transform matrices $\mathbf{M_0}$ and $\mathbf{M_1}$, the summation vector of $\mathbf{M_1}$ is denoted as $\mathbf{V_1} = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$

$$\mathbf{M_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{M_1} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \tag{2}$$

And the mid/side channels are obtained by multiplying input signals with $\mathbf{M_1}$ as $(C_M, C_S) = \mathbf{M_1 V_0}$, where

$$\begin{aligned} C_M &= \frac{\sqrt{2}}{2}C_L + \frac{\sqrt{2}}{2}C_R = S\sin\left(\theta + \frac{\pi}{2}\right) \\ C_S &= \frac{\sqrt{2}}{2}C_L - \frac{\sqrt{2}}{2}C_R = S\sin\left(\theta - \frac{\pi}{2}\right) \end{aligned} \tag{3}$$

In practice, subband energy or masking threshold will be used instead of C_L, C_R in $\mathbf{V_0}$. Only when two channels is sufficiently correlated, for example energy difference is less than a threshold Thr = 2 dB as shown in (4), will the M/S mode be used to avoid being frequently transformed and recalculating the masking threshold [22].

$$\left| 10\log_{10}\left(\frac{C_L}{C_R}\right)^2 \right| \leq \mathrm{Thr} \tag{4}$$

To discuss the switching condition more conveniently, the M/S switching condition will be expressed using the distance between vectors $\mathbf{V_0}$ and $\mathbf{V_1}$. Given $\frac{C_L}{C_R} = \tan\theta$, (4) can be denoted as

$$\frac{1}{\sqrt{1 + 10^{\frac{\mathrm{Thr}}{10}}}} \leq \cos\theta \leq \frac{1}{\sqrt{1 + 10^{-\frac{\mathrm{Thr}}{10}}}} \tag{5}$$

The left side of Figure 1 shows the M/S switching condition (5) in vector space, and Figure 2 illustrates a stereo signal and corresponding transformed sum and difference signals. Equation (3) and the two figures denote that when input signal vector $\mathbf{V_0}$ is close to $\mathbf{V_1}$, where $\cos\theta \approx \frac{1}{\sqrt{2}}$ and $\theta \approx \frac{\pi}{4}$, the switching condition (4) will be satisfied. The difference signal has less amplitude than the original signals, and M/S coding will be used. Since θ is the angle of $\mathbf{V_0}$, switching condition (4) can be represented by the inner product between signal vector $\mathbf{V_0}$ and summation vector $\mathbf{V_1}$

$$\frac{\langle \mathbf{V_0}, \mathbf{V_1} \rangle}{|\mathbf{V_0}|\,|\mathbf{V_1}|} \geq \mathrm{Thr}_v \tag{6}$$

This is an equivalent expression to the energy condition. It indicates that only when the input signal vector is close enough to the summation vector of a M/S transform matrix, this matrix will be used. This idea will be helpful when later discussing the 3D-M/S coding where more than one transform matrix exists. Here, Thr_v is

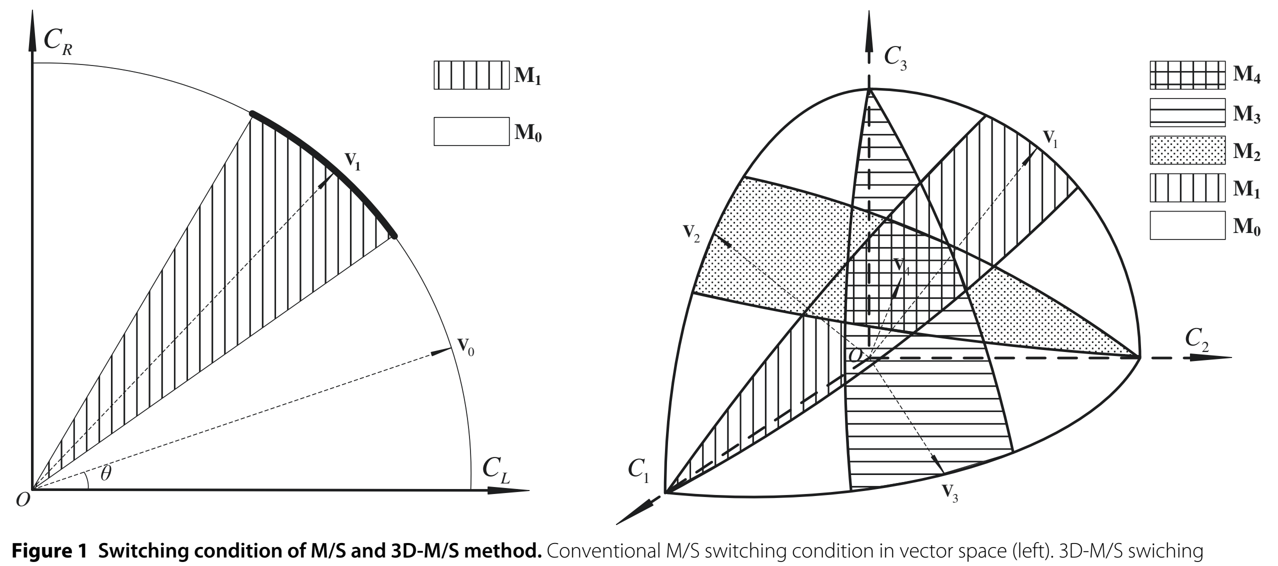

Figure 1 Switching condition of M/S and 3D-M/S method. Conventional M/S switching condition in vector space (left). 3D-M/S swiching condition in vector space (right).

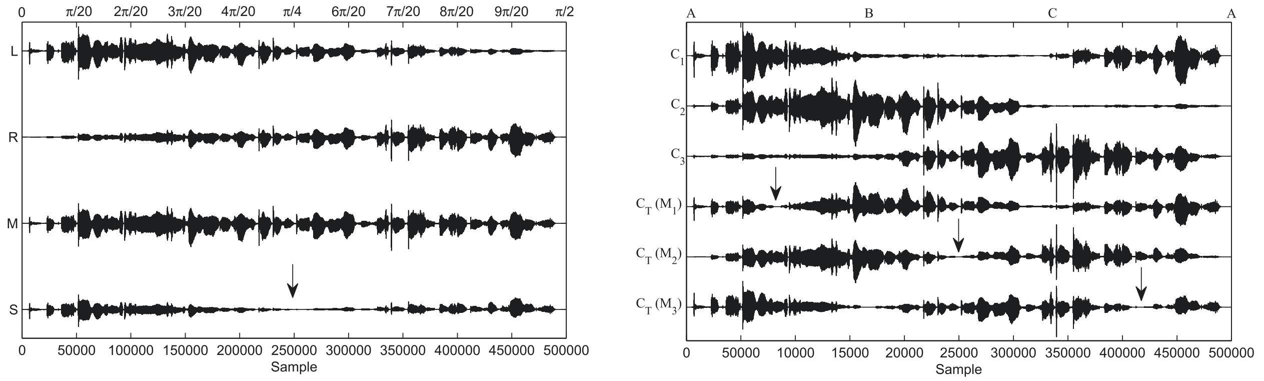

Figure 2 Two-dimensional stereo and three-dimensional VBAP signals and their transformed signals. The left and right channels of a stereo signal and its summation (M) and difference (S) signals (left). The sound source moves from the left loudspeaker to the right by the panning angle θ. When sound source is close to one loudspeaker and θ tends to be 0 and $\frac{\pi}{2}$, the original signal has less dynamic range. When sound source is close to the center and θ tends to be $\frac{\pi}{4}$, the transformed signal has less dynamic range. The three channels of a VBAP signal and its 3D-M/S signals (right). The sound source moves from A to B, B to C and C to A shown in Figure 3. When sound source is close to the center of two channels, transformed signal C_T has less dynamic range with corresponding matrix $\mathbf{M_1}$, $\mathbf{M_2}$, and $\mathbf{M_3}$.

the corresponding switching threshold of Thr in vectorial distance and

$$\mathrm{Thr}_v = \frac{\sqrt{2}}{2}\left(\frac{1}{\sqrt{1+10^{\frac{\mathrm{Thr}}{10}}}} + \frac{1}{\sqrt{1+10^{-\frac{\mathrm{Thr}}{10}}}}\right) \qquad (7)$$

M/S coding in three-dimensional space

For diffuse audio sources like ambient sound, the multichannel signals are weakly correlated. For directional audio sources, the signals are highly correlated which make it possible to reduce the interchannel redundancy. To maintain the stability of the sound image, in a stereo and surround audio system, the virtual source is always being panned or recorded by the two most adjacent channels. So two adjacent channels always have the maximum similarity, and M/S coding and parametric coding are performed based on a two-channel unit. But for Ambisonics, VBAP and 22.2 multichannel systems, sound channels are spherically configured in 3D space as shown in Figure 3. In a VBAP system, three adjacent channels form a directional sound image and have the maximum correlation. In other

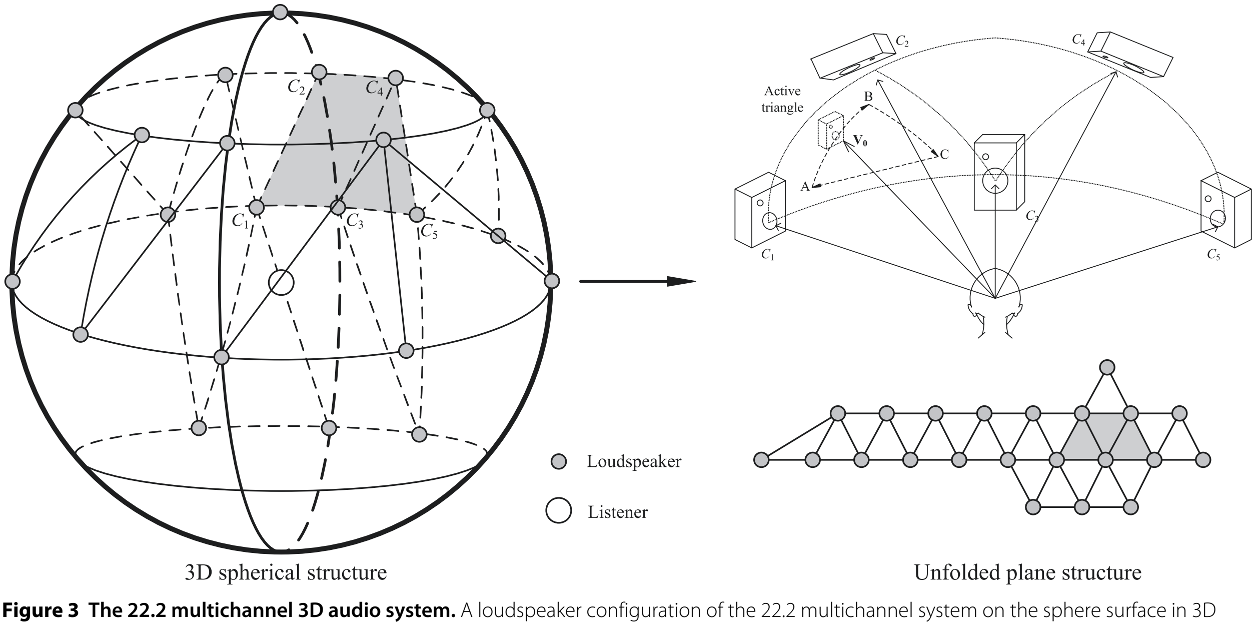

Figure 3 The 22.2 multichannel 3D audio system. A loudspeaker configuration of the 22.2 multichannel system on the sphere surface in 3D space and its plane triangle structure.

3D systems such as Ambisonics, three channels cover a basic area of 3D space and adjacent three channels also have the most similar signals [31]. If the multichannel signals are still grouped and compressed on the channel pair basis, like Ando et al. recent work for coding 22.2 multichannel signals with the AAC codec [32], there will be more than one adjacent channel for each channel. It will require extra computation to select the channel pair but may still lead to a possible to mismatch. For example, in Figure 3, channel C_1, C_2 and C_3 formed a virtual source. If C_1 and C_2 were grouped for parametric or M/S coding, C_3 would be grouped with another less correlated channels, which decreases the coding performance. And if we dynamically group the signals, we have to use complex correlation analysis algorithm to analyze its six adjacent channels. Because the channel pair grouping is based on frequency subbands, it will not only increase the codec complexity dramatically but will also be unable to reduce the overall redundancy that exist in more than two channels. In brief, the conventional channel pair unit should be redesigned for 3D audio systems.

Considering the basic unit for any 3D surface is a triangle, a spherical multichannel configuration can be easily unfolded to a plane triangle structure as shown in Figure 3. The above 3D systems also utilize three or more channels to produce 3D audio effects, which leads to interchannel signal redundancy existing in or more than three channels. Hence, channel triples should be the basic unit to remove interchannel redundancy rather than channel pair in conventional coding schemes. More specifically, in a VBAP system, the input signal $\mathbf{V_0} = (C_1, C_2, C_3)$ is calculated following the tangent model in 3D space as

$$\begin{aligned} C_1 &= S \sin\theta \cos\varphi \\ C_2 &= S \sin\theta \sin\varphi \\ C_3 &= S \cos\theta \end{aligned} \tag{8}$$

where $\theta, \varphi \in \left[0, \frac{\pi}{2}\right]$, which determine the gain factor of the three channels.

There are infinite possible situations, where a virtual source can be located. But for the VBAP model, those possibilities can be reduced to three basic situations. First, the virtual source is located near the position of one channel. This situation corresponds to the source panned mainly using one channel, or one channel forms a virtual source with another two channels which are out of current three loudspeakers. This situation is similar to stereo audio with only one active channel, so no transform is performed and $\mathbf{M_0}$ will be used. Second, the virtual source is located between two channels. This situation corresponds to the source panned mainly using two channels, or two channels form a virtual source with one other channel out of the current three channels. This situation is similar to conventional stereo audio, and M/S coding can be applied. However, the M/S transform matrix must be modified to adapt to the three channels condition which is expressed in Equation 9. Third, the source is panned using all the three channels. This is a new situation that stereo audio never contains. To remove the interchannel redundancy, a new transform matrix $\mathbf{M_4}$ is designed following the rule of conventional M/S coding. The first vector is the summation of three channels, and the rest vectors are orthogonal with the first vector. To guarantee the conservation of energy after transformation, unit vectors are used. This matrix realizes the sum-difference processing for 3D channel, and guarantees that when three channel signals are nearly the same, two channels primarily contain the difference signal.

$$\begin{aligned} &\mathbf{M_0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{M_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \\ &\mathbf{M_2} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix} \mathbf{M_3} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &\mathbf{M_4} = \begin{bmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{3} \end{bmatrix} \end{aligned} \tag{9}$$

An example is shown in Figure 2. It can be observed that when the source is close to the center of two or three channels, a corresponding matrix can produce difference signals with lower dynamic range compared to the original channel signals. Under a certain masking threshold, far less bits are required for quantizing the difference signals which brings the coding gain.

Then we discuss the switching condition for 3D-M/S as shown on right side of Figure 1, and the transformed mid/second/third channels are shown in Table 1. Firstly, if all three channel signals are almost the same which means the input vector is close to spherical triangle center $\mathbf{V_4}$ ($\cot\theta \approx \frac{\sqrt{2}}{2}$ and $\varphi \approx \frac{\pi}{4}$), matrix $\mathbf{M_4}$ will be chosen to give two difference channels. Secondly, if there are only two channels satisfying the conventional M/S switching condition and the projection of input vector is close to $\mathbf{V_1}$ ($\cot\theta \approx \sin\varphi$) or $\mathbf{V_2}$ ($\cot\theta \approx \cos\varphi$) or $\mathbf{V_3}$ ($\varphi \approx \frac{\pi}{4}$), 3D-M/S will select the matrix having the nearest distance from the input vector. The distance is measured by vector distance following expression (6). Compared with conventional switching condition, it can be seen that conventional M/S coding works on the two-dimensional space and has two switching areas. 3D-M/S switching condition is an expansion of M/S coding, where its input vector works in three-dimensional space and has five switching

Table 1 Transformed channel signals with five matrices

	C_M	C_S	C_T
M_0	C_1	C_2	C_3
M_1	C_1	$\frac{\sqrt{2}}{2}S\sin\theta(\cot\theta+\sin\varphi)$	$\frac{\sqrt{2}}{2}S\sin\theta(\cot\theta-\sin\varphi)$
M_2	$\frac{\sqrt{2}}{2}S\sin\theta(\cot\theta+\cos\varphi)$	C_2	$\frac{\sqrt{2}}{2}S\sin\theta(\cot\theta-\cos\varphi)$
M_3	$S\sin\theta\sin(\varphi+\frac{\pi}{4})$	$S\sin\theta\sin(\varphi-\frac{\pi}{4})$	C_3
M_4	$\frac{\sqrt{6}}{3}S\sin\theta(\cot\theta+\sin(\varphi+\frac{\pi}{4}))$	$S\sin\theta\sin(\varphi-\frac{\pi}{4})$	$\frac{\sqrt{6}}{3}S\sin\theta(\frac{\sqrt{2}}{2}\sin(\varphi+\frac{\pi}{4})-\cot\theta)$

The expression of transformed signals using five matrices with parameters θ, φ.

areas. Following the vector distance switching condition, the switching rule of 3D-M/S can be denoted as

$$\text{mode} = \begin{cases} M_4, \text{ if} & \frac{\langle \mathbf{V_0}, \mathbf{V_4}\rangle}{|\mathbf{V_0}||\mathbf{V_4}|} \geq \text{Thr}_v \\ M_i, \text{ else if} & \frac{\langle \mathbf{V_{0i}}, \mathbf{V_i}\rangle}{|\mathbf{V_{0i}}||\mathbf{V_i}|} \geq \text{Thr}_v \\ \quad\text{and} & \frac{\langle \mathbf{V_{0i}}, \mathbf{V_i}\rangle}{|\mathbf{V_{0i}}||\mathbf{V_i}|} \geq \frac{\langle \mathbf{V_{0j}}, \mathbf{V_j}\rangle}{|\mathbf{V_{0j}}||\mathbf{V_j}|}, \quad \forall j \neq i \\ M_0, \text{ else} & \end{cases} \tag{10}$$

where $i, j \in \{1, 2, 3\}$, $\mathbf{V_{01}} = (0, C_2, YC_3)$, $\mathbf{V_{02}} = (C_1, 0, C_3)$, $\mathbf{V_{03}} = (C_1, C_2, 0)$ are the two channel projections of input vector $\mathbf{V_0}$. $\mathbf{V_1} = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$, $\mathbf{V_2} = \left(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$, $\mathbf{V_3} = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0\right)$, $\mathbf{V_4} = \left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right)$ are the summation vectors of each transform matrix.

3D-M/S psychoacoustic model

The perceptual threshold is the maximum amount of noise for an audio signal that will not degrade its subjective quality. 3D-M/S coding transforms the original signals into summation and difference signals and then sends these to the core codec. The perceptual thresholds calculated by the core codec are for the transformed signals, which can not guarantee the noise in original signals is imperceptible. So after transformation, the perceptual threshold calculation in the core encoder needs a redesign. An example is the conventional M/S stereo coding, where the masking threshold model for the main channel and side channel is revised to avoid perceptible noise in reconstructed signals [23]. To derive the masking threshold for 3D-M/S signals, we reference the deducing method in [23] and expand it to channel triple case. Consider the transformed signals $(C_M, C_S, C_T) = \mathbf{M_4}\mathbf{V_0}^T$ as

$$\begin{aligned} C_M &= \tfrac{C_1+C_2+C_3}{\sqrt{3}} \\ C_S &= \tfrac{C_1-C_2}{\sqrt{2}} \\ C_T &= \tfrac{C_1+C_2-2C_3}{\sqrt{6}} \end{aligned} \tag{11}$$

where C_M is the first signal after transformation and C_S and C_T are the second and third signals, respectively. After core codec quantization, independent noise is introduced into the three signals which is denoted as N_M, N_S and N_T. So at the decoder side:

$$\begin{aligned} C_M + N_M &= \tfrac{\hat{C}_1+\hat{C}_2+\hat{C}_3}{\sqrt{3}} \\ C_S + N_S &= \tfrac{\hat{C}_1-\hat{C}_2}{\sqrt{2}} \\ C_T + N_T &= \tfrac{\hat{C}_1+\hat{C}_2-2\hat{C}_3}{\sqrt{6}} \end{aligned} \tag{12}$$

where $\hat{C}_1$, $\hat{C}_2$ and $\hat{C}_3$ are the reconstructed signals. Compared with the original signals shown in (11), the noise energy for the original signals can be obtained

$$\begin{aligned} \sigma^2_{C_1-\hat{C}_1} &= \tfrac{1}{3}N_M^2 + \tfrac{1}{2}N_S^2 + \tfrac{1}{6}N_T^2 \leq T_{C_1} \\ \sigma^2_{C_2-\hat{C}_2} &= \tfrac{1}{3}N_M^2 + \tfrac{1}{2}N_S^2 + \tfrac{1}{6}N_T^2 \leq T_{C_2} \\ \sigma^2_{C_3-\hat{C}_3} &= \tfrac{1}{3}N_M^2 + \tfrac{2}{3}N_T^2 \leq T_{C_3} \end{aligned} \tag{13}$$

where $\sigma^2_{C_1-\hat{C}_1}$, $\sigma^2_{C_2-\hat{C}_2}$ and $\sigma^2_{C_3-\hat{C}_3}$ are the expected noise energy for original signals. T_{C_1}, T_{C_2} and T_{C_3} are the masking thresholds calculated for the original signals. The sufficient conditions that guarantee the expected noise energy will not exceed masking threshold are $N_M^2, N_S^2, N_T^2 \leq \min(T_{C_1}, T_{C_2}, T_{C_3})$. This means the noise energy for any transformed signal must be less than the minimum threshold of their original signals. So the masking thresholds for the transformed signals can be derived from the masking threshold of the original signals as

$$T_{C_M} = T_{C_S} = T_{C_T} \leq \min(T_{C_1}, T_{C_2}, T_{C_3}) \tag{14}$$

The same results can be deduced for other matrices.

Framework for general channel configuration

3D-M/S only works for three channels, actually all 3D audio systems contain dozens of channels. A general channel configuration is shown in Figure 4, where channel C corresponds to a loudspeaker. Here, a framework is proposed based on 3D-M/S coding as shown in Figure 5. Because all spatially placed loudspeakers can be decomposed into basic triangle units, this structure will enable 3D-M/S coding to work for arbitrary channel configurations. The framework processes the audio channels triangle by triangle until all channels are coded. C_M is the summation channel and C_S and C_T are the second and third channels, respectively. Every 3D-M/S unit shares two channels with the previous unit and only one new

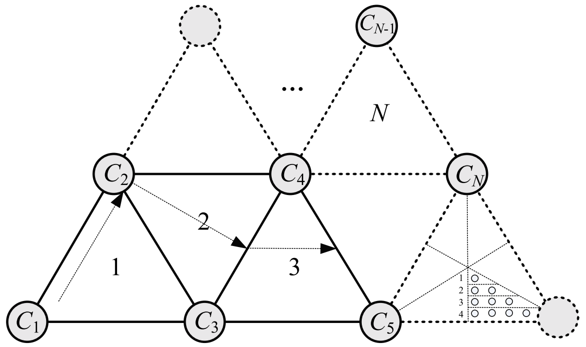

Figure 4 The general plane structure of 3D multichannel systems. A general plane structure for 3D multichannel audio systems and experimental sound source setting.

channel is added in. So, it only needs to compress the channel which contains the signal of the new channel. For all matrices M_0, M_1, M_2, M_3 and M_4, C_T is the third channel after 3D-M/S transform. Because every unit outputs only one that contains a new input channel, the whole coding framework keeps the number of channels exactly the same as original input signals. And because the output channel contains either the difference signal or original signal, coding gain can be obtained. The original signals can be obtained by multiplying 3D-M/S inverse transform matrix subband by subband at the decoder side. This framework is also suitable for other methods. For example, replacing the 3D-M/S with PCA, the codec can achieve better interchannel redundancy removing performance.

Experiment

The experiment used five channels (C_1, C_2, C_3, C_4, C_5) in spherical 22.2 multichannel configuration as shown in Figure 4. Considering that PCA is the best decorrelation transform theoretically and Independent channel coding is widely used for 22.2 multichannel compression, the experiment compared the proposed 3D-M/S method with PCA and Independent channel coding in bitrate, complexity and objective quality. Three MPEG test sequences (es01 voice signal, sc03 symphony music signal, si02 castanets transient signal, mono 48-kHz sampling) were used as the moving virtual sources following the VBAP rule, four sequences (si03, si01, sc01, es02) were used as the discrete fixed-position virtual sources. The virtual sources and respective azimuth and altitude panning angle are generated on a per-frame basis. Here, only point virtual sources were used to test the best performance of three methods, as subband signals can be regarded as point sources in subband coding when bandwidths are small enough. Signals with decorrelated elements are beyond the scope of VBAP model and will decrease the coding performance, for its difference signals retains high energy which depends on the correlation and the energy of the decorrelated elements. Uncorrelated signals with independent audio content is tested in the end.

Three basic virtual sound movements were used to cover some basic possible virtual source locations. The three movements in a triangle are movement 1 from point to point (virtual source es01), movement 2 from point to edge (virtual source sc03) and movement 3 from edge to edge (virtual source si02). Three movements are as shown

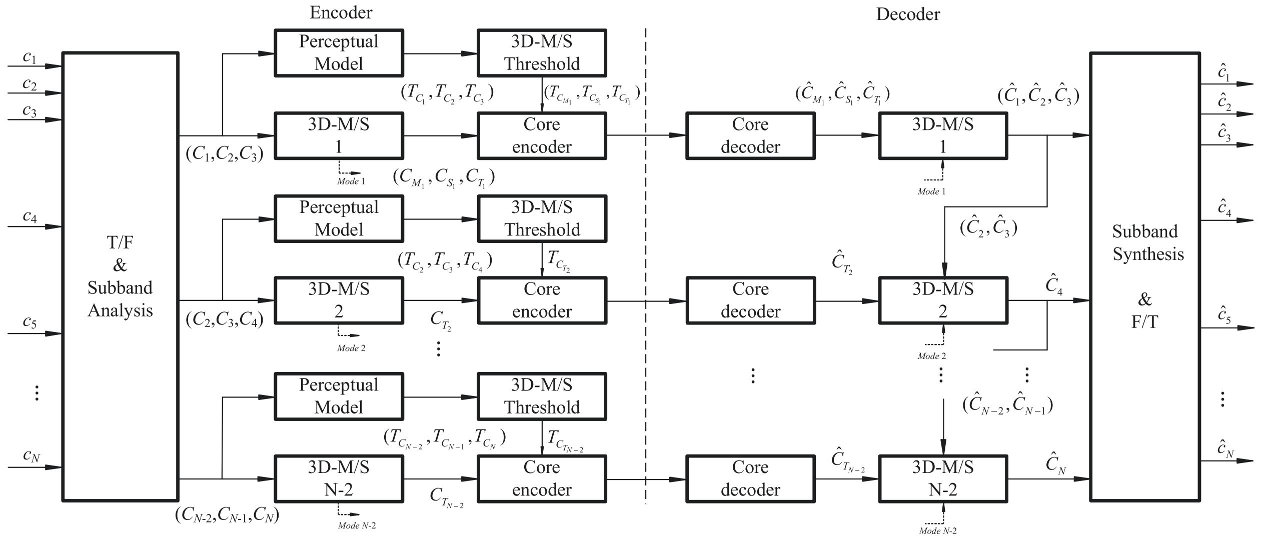

Figure 5 Three-channel unit-based coding framework. The framework using 3D-M/S coding for *N* channels.

Table 2 Discrete virtual source setting and objective results

	θ	φ				Average
		$\frac{7\pi}{32}$	$\frac{5\pi}{32}$	$\frac{3\pi}{32}$	$\frac{\pi}{32}$	
Ind	$\frac{11\pi}{36}$	20.76	–	–	–	20.76
3D-M/S	$\frac{11\pi}{36}$	21.45	–	–	–	21.45
PCA	$\frac{11\pi}{36}$	21.95	–	–	–	21.95
Ind	$\frac{13\pi}{36}$	15.37	21.14	–	–	18.26
3D-M/S	$\frac{13\pi}{36}$	16.07	21.88	–	–	18.98
PCA	$\frac{13\pi}{36}$	16.54	22.74	–	–	19.64
Ind	$\frac{15\pi}{36}$	23.48	23.28	22.08	–	22.95
3D-M/S	$\frac{15\pi}{36}$	23.77	22.93	22.54	–	23.08
PCA	$\frac{15\pi}{36}$	23.86	23.78	23.49	–	23.71
Ind	$\frac{17\pi}{36}$	21.82	19.06	18.93	19.36	19.79
3D-M/S	$\frac{17\pi}{36}$	21.88	18.52	19.02	19.51	19.73
PCA	$\frac{17\pi}{36}$	21.92	21.36	19.18	20.57	20.76

SNR (dB) of three methods for 10 discrete virtual source positions at about 64 kbps.

in Figure 4. Different sources were used for different panning patterns, because the experiments were designed to compare three methods on the same condition and change the condition simultaneously to see how three methods work on different kinds of sources and locations, instead of comparing the performance of one method on different virtual source locations with the same source. Considering the symmetry in the triangle, 10 discrete virtual source positions in $\frac{1}{6}$ triangle as shown in Table 2 and Figure 4 are used which equally divided the $\frac{1}{6}$ triangle: si03 was panned to $\left(\theta = \frac{11\pi}{36}, \varphi = \frac{7\pi}{32}\right)$; si01 was divided into two sequences and panned to $\left(\theta = \frac{13\pi}{36}, \varphi = \frac{7\pi}{32}\right)$, $\left(\theta = \frac{13\pi}{36}, \varphi = \frac{5\pi}{32}\right)$; sc01 was divided into three sequences and panned to $\left(\theta = \frac{15\pi}{36}, \varphi = \frac{7\pi}{32}\right)$, $\left(\theta = \frac{15\pi}{36}, \varphi = \frac{5\pi}{32}\right)$ and $\left(\theta = \frac{15\pi}{36}, \varphi = \frac{3\pi}{32}\right)$; es02 was divided into four sequences and panned to $\left(\theta = \frac{17\pi}{36}, \varphi = \frac{7\pi}{32}\right)$, $\left(\theta = \frac{17\pi}{36}, \varphi = \frac{5\pi}{32}\right)$, $\left(\theta = \frac{17\pi}{36}, \varphi = \frac{3\pi}{32}\right)$ and $\left(\theta = \frac{17\pi}{36}, \varphi = \frac{\pi}{32}\right)$.

The 3D-M/S and PCA was used in each subband in the frequency domain. The three encoders were realized based on FAAC-1.28, and decoders were based on FAAD2-2.7. AAC-LC was used as the core codec and only the long window was enabled for simplification. To avoid the influence of dynamic bandwidth setting of the FAAC, the experiment fixed the bandwidth at 12 kHz with 35 subbands.

Independent channel coding: Audio signals were sent into the core codec and compressed directly.

3D-M/S: The vector was calculated using the subband energy of three channels from AAC psychoacoustic module with no extra energy computation. Then 3D-M/S matrix switching was performed and 3 bits were used per mode parameter. The transformed signals were sent into the core codec, and the masking threshold was modified accordingly.

PCA: The eigenvectors were calculated for each subband. Subband signals were transformed using eigenvector matrix and then sent into core codec. The covariance matrix was quantized and transmitted to the decoder following a previous KLT-based multichannel audio coding scheme [21], with 4 bits per non-redundant element.

Objective evaluation

Complexity was measured by the running time of each method on PC (CPU: Intel Core2 Duo P8600 2.53GHz, RAM: 8GB). The main application scenario of the proposed method is medium and high bitrate, and the sound pressure at the center listening point is the most

Table 3 Bitrate setup and overall SNR

	Quality	Bitrate/channel (kbps)								Average
	(SNR)	C_1	C_2	C_3	C_4	C_5	P_1	P_2	P_3	
Ind	16.08	61.6	60.8	60.3	58.7	58.7	–	–	–	61.2
3D-M/S	18.21		229.5		21.7	57.9	4.9	4.9	4.9	64.7
PCA	18.87		133.2		25.2	42.8	39.3	39.3	39.3	63.8

Overall SNR (dB) and bitrate of three methods. For PCA and 3D-M/S, C_1, C_2 and C_3 share a total bitrate, and C_4 and C_5 are the C_T channels. P_1, P_2 and P_3 are the parameter bitrates for triangles 1, 2 and 3.

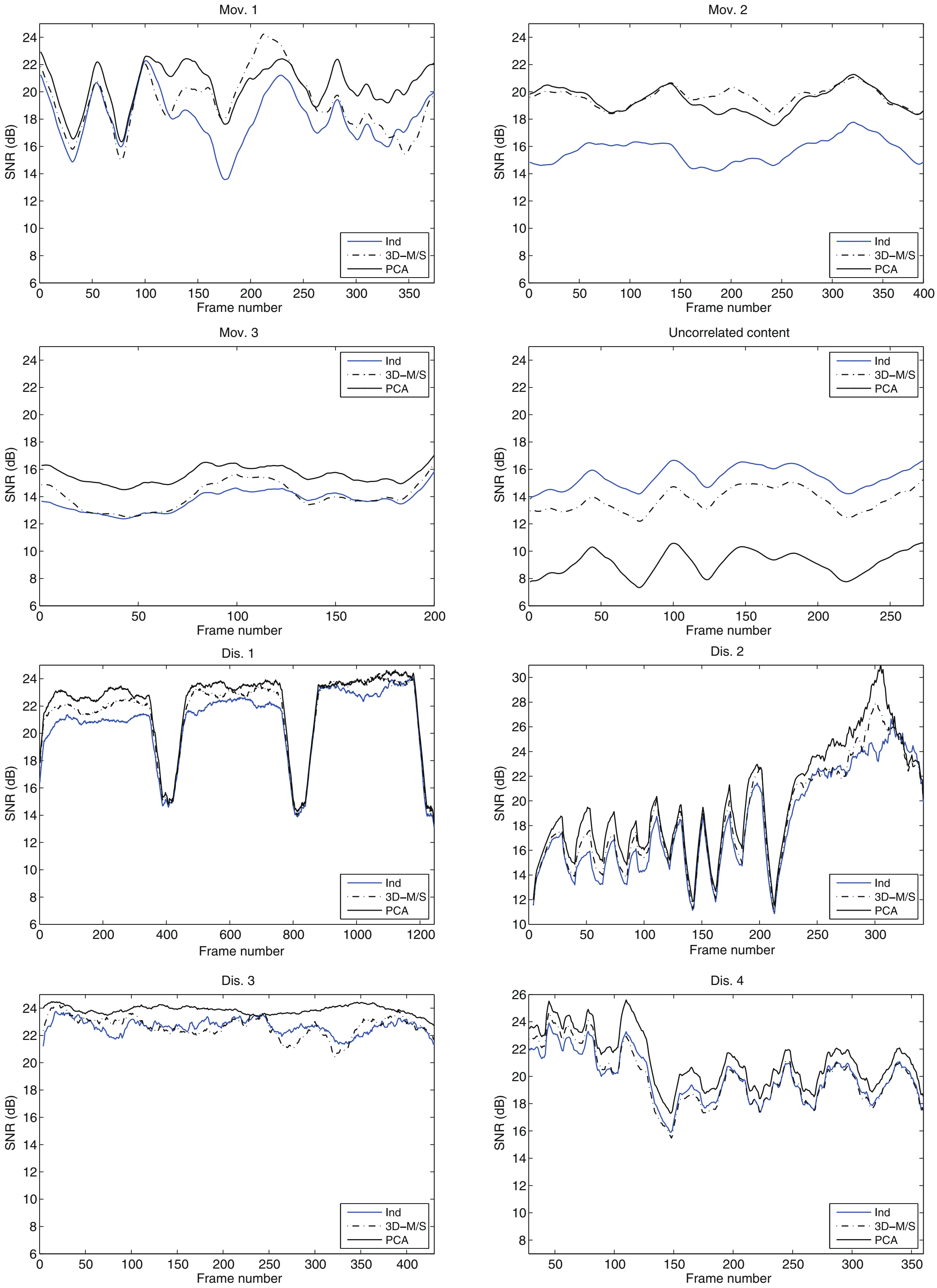

Figure 6 Objective evaluation results. Average SNR (12 kHz bandwidth) per frame for Independent channel coding, 3D-M/S and PCA methods in movements 1, 2 and 3, uncorrelated content and discrete positions 1, 2, 3 and 4.

important in multichannel audio and the VBAP panning method. Here we compared the sound pressure SNR (original and reconstructed signal sound pressure in the center listening point) to measure the objective quality following the sound pressure equation in [33]

$$p(\omega) = G \sum_{j=1}^{5} \frac{e^{-ikr_j}}{r_j} C_j(\omega), \tag{15}$$

where $p(\omega)$ is the sound pressure, G is a proportionality coefficient, $k = \frac{\omega}{c}$ is the wave number and c is the wave speed. r is the distance from the loudspeaker to the listening point, and in the spherical 22.2 multichannel, all loudspeakers have the same distance. And the SNR is calculated by

$$\mathrm{SNR} = \frac{\int_0^{2\pi} \left|p(\omega)\right|^2 d\omega}{\int_0^{2\pi} \left|p(\omega) - \hat{p}(\omega)\right|^2 d\omega} \tag{16}$$

Firstly, the bitrates of all three methods have been set to be nearly identical to compare the performance shown in Table 3. The bitrates were adopted by modifying the masking threshold in core codec and iterating two nested loops until rate constrain satisfied. Figure 6 shows the SNR curves for each method, where each SNR curve was smoothed to omit the details. The smoothing method is a typical average filtering over 20 frames length. An overall SNR was presented in Table 3 to give an overview by averaging the SNR for all frames. The three SNR curves have a downtrend because transient signal and symphony signal have a wider bandwidth than the voice signal, which require more bits to achieve the similar SNR. When the virtual source came close to the middle of the two channels (between C_1, C_2 around the 200th frame in Mov. 1, between C_3, C_4 for all frames in Mov. 2, between C_3, C_4 around the start frames, between C_3, C_5 around the 100th frame and between C_4, C_5 around the end frames in Mov. 3), 3D-M/S gets a higher SNR than the Independent channel coding and close to PCA. Moveover, around the 200th frame in Mov. 1 and Mov. 2, where all two and three channels are nearly the same, $\mathbf{M_3}$ and $\mathbf{M_4}$ can remove redundancy to the largest extent and outperform the PCA method. This is because some transformed subband signals came below the masking threshold and more bits were reserved for summation channel. The same results can be seen in the discrete virtual sources in Table 2, where $\varphi = \frac{7\pi}{32}$ ($\varphi \approx \frac{\pi}{4}$) and ($\theta = \frac{11\pi}{36}, \varphi = \frac{7\pi}{32}$) ($\theta = \frac{13\pi}{36}, \varphi = \frac{5\pi}{32}$) ($\theta = \frac{15\pi}{36}, \varphi = \frac{3\pi}{32}$) ($\theta = \frac{17\pi}{36}, \varphi = \frac{\pi}{32}$) ($\cot\theta \approx \sin\varphi$). But when the virtual source located beyond the middle of two channels, such as ($\theta = \frac{15\pi}{36}, \varphi = \frac{5\pi}{32}$), M/S coding cannot bring coding gain. In conclusion, if the input signals are located in one of the five switching areas, the coding gain can be obtained by transforming them into summation and difference signals.

Secondly, the PCA parameter bitrate of 39.3 kbps/channel is considerably higher than 3D-M/S. If the three channels have little correlation (e.g. channels with different contents or ambient sound), the transformed signals will not save any bits and cause the decrease of coding efficiency. To test the three methods under such condition, the virtual sources of three different signals were fixed at three channels and coded all at 64 kbps. The experimental result is shown in Figure 6. We can see Independent channel coding achieves the best performance in this case; meanwhile, 3D-M/S degrades about 1 dB and PCA degrades nearly 7 dB. It is because, for PCA requirement, too many bits are used for parameters which now cannot bring any coding gain. But for 3D-M/S, parameter bits for modes are only 4.9 kbps/channel. It will not reduce the coding efficiency much for medium and high bitrate conditions, which is the main application scenario of M/S coding. Although the high bitrate for PCA can be alleviated by reducing the refresh rate of PCA parameters, but it will decrease the coding performance on VBAP signals at the same time.

Finally, Table 4 demonstrates the complexity of each method. Due to the computation for covariance matrix and eigenvector matrix, PCA increases about 30% complexity compared with the original AAC codec. For 3D-M/S, the matrix switching and signal transformation increased by about 11% complexity.

Subjective evaluation

Eight subjects who are actively working in the domain of audio compression participated in the subjective test based on ITU MUSHRA [34] method using a 3.5-kHz low-pass filtered original channels as anchor, and the test was carried out in a quite room with five-spherical channel configuration as shown in Figure 7. Subjects were required to evaluate the sound quality and sound orientation separately, and can draw the perceived sound position and movement to help their rating.

The MUSHRA test results are shown in Figure 8. For the sound quality, the subjective result generally matches

Table 4 Time complexity

	Complexity (s)		
	Encoder	Decoder	Ratio
Ind	2.382	0.223	100.0%
3D-M/S	2.604	0.306	111.7%
PCA	2.977	0.416	130.2%

Time complexity of three methods, the length of the three sequences are 27 s.

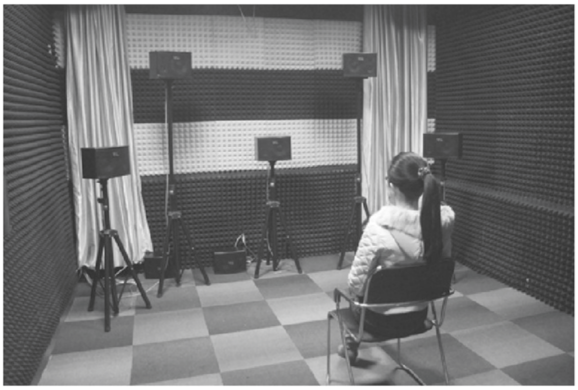

Figure 7 Subjective test setup. Subjective test environment and configuration.

the SNR result where PCA and 3D-M/S got a subjective quality improvement compared with the Independent channel coding. For the sound orientation, subjects reported that the voice and castanet signals were easier to locate. Although the 3.5-kHz anchor signal was poor in sound quality test, some subjects indicated its virtual source position was not so bad in the movement than the discrete virtual source. This may be because it is easier to distinguish the direction change for fixed sound source than the moving sound source. For 3D-M/S, subjects felt the virtual source had a sharper position in the center of two loudspeakers compared with Independent channel coding. On the whole, the subjective improvement in sound orientation exists but is not as obvious as the sound quality.

From the above results on three point sources and uncorrelated signals, it can be observed that both PCA and 3D-M/S method get about 13% SNR improvement

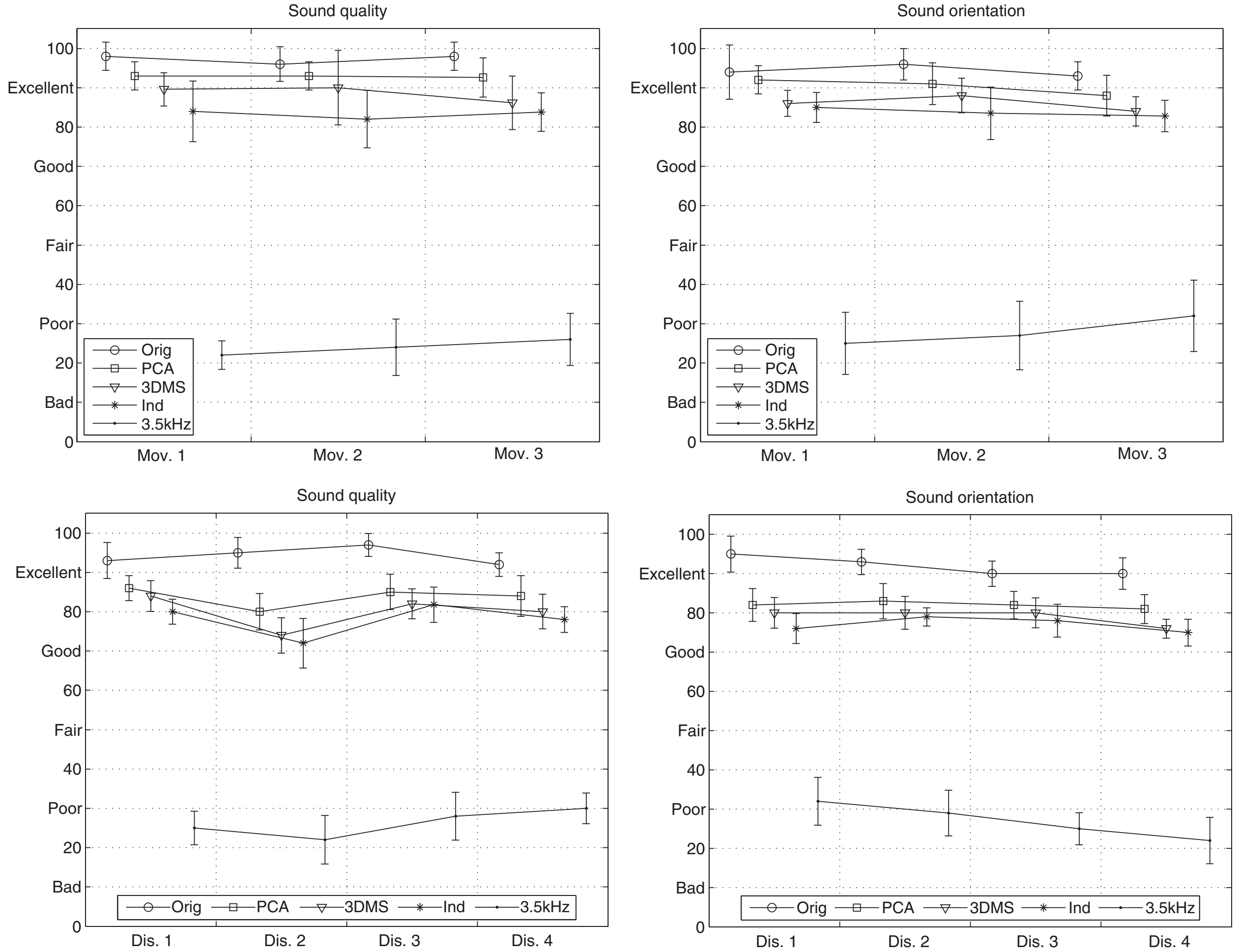

Figure 8 Subjective evaluation results. MUSHRA test results on sound quality and orientation for Independent channel coding, 3D-M/S and PCA methods in movements 1, 2 and 3 and discrete positions 1, 2, 3 and 4 with 95% confidence interval and the mean.

for each channel. But the complexity of 3D-M/S is much lower than PCA to achieve similar performance. It can be explained that the fixed matrix transform can be regarded as some special vectors in PCA. The special vectors are chosen based on the assumption that channel signals are either quite similar or quite different. This assumption may not be always true for the diversity of subband signals, but it makes a good compromise between coding efficiency and complexity.

Conclusion

This paper proposed a 3D-M/S coding method, which inherits the low complexity of conventional M/S coding. Moreover, 3D-M/S performs the sum and difference coding triple by triple, rather than couple by couple of the conventional method. This structure is more suitable for a 3D multichannel audio configuration, because adjacent three channels form a triangle and will have the maximum redundancy in spatial configured 3D audio channels. Besides, it is also convenient to unfold 3D audio multichannel structure into plane triangles. Combining the proposed framework, 3D-M/S and PCA methods can be applied to more than three channels. An experiment on VBAP signals indicates the performance of proposed method with relatively low complexity, comparing to the PCA and independent channel coding. Considering the development of 3D audio technology and its requirement for compression efficiency, a low complexity 3D audio codec will be promising and preferable for practical application.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (nos. 61231015, 61102127, 61201340, 61201169) and Natural Science Foundation of Hubei (nos. 2011CDB451, 2012FFB04205).

Received: 1 November 2013 Accepted: 12 March 2014
Published: 24 March 2014

References

1. AJ Berkhout, D de Vries, P Vogel, Acoustic control by wave field synthesis. J. Acoust. Soc. Am. **93**(5), 2764–2778 (1993)
2. MA Gerzon, Ambisonics in multichannel broadcasting and video. J. Audio Eng. Soc. **33**(11), 859–871 (1985)
3. J Cooperstock, Multimodal telepresence systems. IEEE Signal Process. Mag. **28**, 77–86 (2011)
4. A Staff, Multichannel audio systems and techniques. J. Audio Eng. Soc. **53**(4), 329–335 (2005)
5. F Rumsey, Cinema sound for the 3-D era. J. Audio Eng. Soc. **61**(5), 340–344 (2013)
6. J Nettingsmeier, Birds on the wire - WFS live transmission project report. Tech. rep., Fraunhofer 2008
7. S Sakaida, K Iguchi, N Nakajima, Y Nishida, A Ichigaya, E Nakasu, M Kurozumi, S Gohshi, The super hi-vision codec, in *IEEE International Conference on Image Processing, 2007. ICIP 2007, Volume 1*, (2007), pp. I-21–I-24
8. F Baumgarte, C Faller, Binaural cue coding-part I: psychoacoustic fundamentals and design principles. IEEE Trans. Speech Audio Process. **11**(6), 509–519 (2003)
9. C Faller, F Baumgarte, Binaural cue coding-part II: schemes and applications. IEEE Trans. Speech Audio Process. **11**(6), 520–531 (2003)
10. W Oomen, E Schuijers, B Brinker den, J Breebaart, Advances in parametric coding for high-quality audio, in *Audio Engineering Society Convention 114*, (2003)
11. J Breebaart, Par van de S, A Kohlrausch, E Schuijers, Parametric coding of stereo audio. EURASIP J. Adv. Sig. Pr. **2005**(9), 561917 (2005)
12. J Herre, S Disch, New concepts in parametric coding of spatial audio: from SAC to SAOC, in *2007 IEEE International Conference on Multimedia and Expo*, (2007), pp. 1894–1897
13. M Goodwin, J Jot, Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement, in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007, Volume 1*, (2007), pp. I-9–I-12
14. B Cheng, C Ritz, I Burnett, A spatial squeezing approach to ambisonic audio compression, in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, (2008), pp. 369–372
15. E Hellerud, A Solvang, U Svensson, Spatial redundancy in Higher Order Ambisonics and its use for lowdelay lossless compression, in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009* (2009), pp. 269–272
16. C Tzagkarakis, A Mouchtaris, P Tsakalides, A multichannel sinusoidal model applied to spot microphone signals for immersive audio. IEEE Trans. Audio Speech Lang. Process. **17**(8), 1483–1497 (2009)
17. F Pinto, M Vetterli, Wave field coding in the spacetime frequency domain, in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, (2008), pp. 365–368
18. F Pinto, M Vetterli, space-time-frequency processing of acoustic wave fields: theory, algorithms, and applications. IEEE Trans. Signal Process. **58**(9), 4608–4620 (2010)
19. B Cheng, Spatial squeezing techniques for low bit-rate multichannel audio coding. PhD thesis. University of Wollongong 2011
20. 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)
21. D Yang, H Ai, C Kyriakakis, CC Kuo, High-fidelity multichannel audio coding with Karhunen-Loeve transform. IEEE Trans. Speech Audio Process. **11**(4), 365–380 (2003)
22. J Johnston, A Ferreira, Sum-difference stereo transform coding, in *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92, Volume 2* (1992), pp. 569–572
23. CM Liu, WC Lee, YH Hsiao, M/S coding based on allocation entropy, in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, (2003)
24. O Derrien, G Richard, A new model-based algorithm for optimizing the MPEG-AAC in MS-Stereo. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1373–1382 (2008)
25. H Krueger, P Vary, A new approach for low-delay joint-stereo coding, in *2008 ITG Conference on Voice Communication (SprachKommunikation)*, (2008), pp. 1–4
26. M Schafer, P Vary, Hierarchical multi-channel audio coding based on time-domain linear prediction, in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, (2012), pp. 2148–2152
27. M Neuendorf, M Multrus, N Rettelbach, G Fuchs, J Robilliard, J Lecomte, S Wilde, S Bayer, S Disch, C Helmrich, R Lefebvre, P Gournay, B Bessette, J Lapierre, K Kjorling, H Purnhagen, L Villemoes, W Oomen, E Schuijers, K Kikuiri, T Chinen, T Norimatsu, CK Seng, E Oh, M Kim, S Quackenbush, B Grill, MPEG unified speech and audio coding-the ISO/MPEG standard for high-efficiency audio coding of all content types, in *Audio Engineering Society Convention 132* (Audio Engineering Society, 2012)
28. M Multrus, M Neuendorf, J Lecomte, G Fuchs, S Bayer, J Robilliard, F Nagel, S Wilde, D Fischer, J Hilpert, N Rettelbach, C Helmrich, S Disch, R Geiger, B Grill, ed. by A Heuberger, G Elst, and R Hanke, MPEG unified speech and audio coding - bridging the gap, in *Microelectronic Systems* (Springer Berlin Heidelberg Berlin, Heidelberg, 2011), pp. 351–362
29. C Helmrich, P Carlsson, S Disch, B Edler, J Hilpert, M Neusinger, H Purnhagen, RettelbachN, J Robilliard, L Villemoes, Efficient transform coding of two-channel audio signals by means of complex-valued stereo prediction, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2011), pp. 497–500
30. CM Liu, HW Hsu, WC Lee, Compression artifacts in perceptual audio coding. IEEE Trans. Audio Speech Lang. Process. **16**(4), 681–695 (2008)

31. F Zotter, M Frank, All-round ambisonic panning and decoding. J. Audio Eng. Soc. **60**(10), 807–820 (2012)
32. A Ando, T Sugimoto, K Irie, Coding of 22.2 multichannel audio signal by MPEG-AAC, in *IEICE Tech. Rep., Volume 113 of EA2013-46*, (2013), pp. 75–80
33. 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)
34. ITU-T, *Method for the subjective assessment of intermediate sound quality (MUSHRA)*, (2001)

doi:10.1186/1687-4722-2014-10
Cite this article as: Dong *et al.*: **Expanded three-channel mid/side coding for three-dimensional multichannel audio systems.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:10.