AN EXPANDED MID/SIDE CODING FOR 3D AUDIO SIGNAL COMPRESSION

Shi Dong, Ruimin Hu, Xiaochen Wang, Weiping Tu, Xiang Zheng

National Engineering Research Center for Multimedia Software, Wuhan University, China

ABSTRACT

Three dimensional (3D) audio technologies are booming with the success of 3D video technology. The sharply increased audio channels make its huge data unacceptable for transmitting bandwidth and storage media. This paper investigates the conventional Mid/Side (M/S) coding method, and expands it to a Three-channel Dependent M/S coding (3D-M/S) method. 3D-M/S perform sum and difference coding based on three channels instead of conventional two channels, and corresponding transform matrixes are presented. Furthermore, a framework is proposed to enable 3D-M/S compress any number of audio channels. Experiment shows proposed method obtains 25.4% objective quality improvement comparing with independent channel coding, and only increases 11.3% complexity comparing with the 29.9% of PCA method.

Index Terms— 3D audio compression, VBAP, M/S coding

1. INTRODUCTION

3D audio has attracted more attention and developed fast recently following the booming market of 3D movie. Many 3D audio technologies are now introduced into audio involved application to replace the surround sound system for the ability to provide superior sound localization and immersive feeling. Wave Field Synthesis (WFS), Ambisonics and Vector Based Amplitude Panning (VBAP) are the three most well-developed 3D audio theories. WFS generally follows Huygens principle to reconstruct the original sound field. Research institutions such as IDMT of Fraunhofer and IRCAM in France have a long 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. VBAP follows sine law in three dimensional space using three adjacent loudspeakers to form a sound vector. 3D System like 22.2 multichannel system proposed by NHK in Japan utilizes VBAP to generate 3D sound image. The 22.2 multichannel system is also included in MPEG-H standard for rendering 3D audio scene which is now in developing.

It is clear that 3D audio technology will become mature gradually and replace stereo and surround sound. 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. 22.2 system has three layer and 24 audio channels. Although Ambisonics system can have flexible order and channel number, it usually uses dozens of channel because fewer channels will cause quality deterioration. Comparing with two channel stereo and 5.1 surround sound, the increasing of audio channel causes 3D audio data increase dramatically. A report of Fraunhofer shows 37Mbps is needed for live transmitting WFS, and for 22.2 system uncompressed data also requires 28Mbps [1]. Currently, storage media and transmission bandwidth can hardly afford those huge data size. So the compression of 3D audio signals becomes an important subject in 3D audio technologies.

Recently some valuable works have been done to increase the compression efficiency for 3D audio signals. In 2007, Goodwin proposed a PCA based multichannel compression framework for parametric coding [2], which can be applied to enhance specific audio scenarios and robust spatial audio coding. In 2008, Cheng proposed Spatially Squeezed Surround Audio Coding (S^3AC) for parametrically coding the Ambisonics signal [3]. In 2009, Hellerud used an inter-channel prediction based coding method to remove the redundancy between Ambisonics channels [4], it has low algorithm delay but high computational complexity. In 2010, Pinto 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. And coding efficiency increases along with the number of audio channels, because the transform decomposition accuracy depends on spatial resolution which is the number of WFS channels [5, 6]. In 2011, Cheng further proposed a Spatial Localization Quantization Point (SLQP) codec using localization cues to compress the VBAP signals [7]. Since SLQP extracts the spatial cues and downmixes the channels, it achieved high compression ratio.

Above model based codec and parametric codec can offer considerable compression ratio. However, in practice the computational complexity of an audio codec should be acceptable while maintain enough coding efficiency. And parametric coding can only get performance gain at low bitrate. In this paper, we consider high-quality/high-bitrate application and focus on the conventional Mid/Side (M/S) coding method. A Three-channel Dependent M/S coding (3D-M/S) method and corresponding framework are proposed to compress VBAP-like 3D audio system signals. The main idea is to expand M/S coding to three-dimensional way by designing new transform matrix, which remove the redundancy of three channels in 3D space rather than just two channels in horizon plane. And a new framework enables 3D-M/S to compress any channel configuration. The 3D-M/S retains the low complexity and high efficiency property of conventional M/S coding. A comparison of 3D-M/S coding with PCA coding and independent channel coding is performed to justify the performance of compression ratio and computational complexity.

2. M/S CODING IN 3D SPACE

2.1. Conventional M/S coding

M/S coding was proposed by J.D. Johnston [8] and adopted by many audio codec such as MPEG2-Layer III and MPEG4-AAC. It is based on the fact that most stereo channels are strongly correlated. By simply transforming the stereo channel pair into M/S domain, a summa-

This work is supported by NSFC(No.61231015, No.61201340, No.61102127, No.61201169), Hubei NSF(2011CDB451, 2012FFB04205).

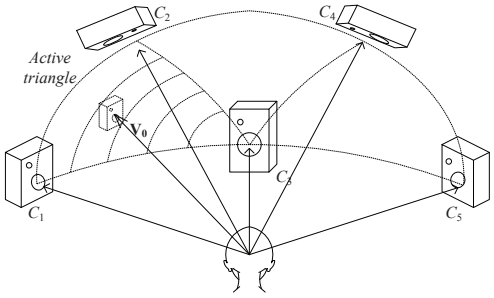

Fig. 1: A loudspeakers configuration of the VBAP system on the sphere surface in 3D space.

tion channel and a difference channel are coded instead. Since difference channel has less dynamic range than original channel, less bits are required and coding gain is obtained. To illustrate how M/S coding works, a generalized sine stereo model is used. Here 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 stereo panning angle and $\theta \in [0, \frac{\pi}{2}]$. The M/S coding can be denoted as two transform matrix M_0 and M_1, the summation vector of M_1 is denoted as $\mathbf{V_1} = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$

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

In practice, subband energy or maksing threshold will be used instead of C_L, C_R in $\mathbf{V_0}$. Only when two channels is sufficiently close enough (eg. energy difference is less than a threshold $Thr = 2$dB [8]) will the M/S mode be used to avoid being frequently transformed and recalculating the masking threshold.

$$\left| 10\log_{10}\left(\frac{C_L}{C_R}\right)^2 \right| \leq Thr \tag{3}$$

To discuss the switching condition more conveniently, the M/S switching condition will be expressed using the distance between two vectors. For $\frac{C_L}{C_R} = \tan\theta$, (3) can be denoted as

$$\frac{1}{\sqrt{1+10^{\frac{Thr}{10}}}} \leq \cos\theta \leq \frac{1}{\sqrt{1+10^{-\frac{Thr}{10}}}} \tag{4}$$

It suggests only when $\theta \approx \frac{\pi}{4}$ and $\cos\theta \approx \frac{1}{\sqrt{2}}$ will the switching condition (3) be satisfied, then M/S coding will be used. Since θ is the angle of $\mathbf{V_0}$, switching condition (3) can be represented by the inner product between signal vector $\mathbf{V_0}$ and summation vector $\mathbf{V_1}$

$$\langle \mathbf{V_0}, \mathbf{V_1} \rangle \geq Thr_v \tag{5}$$

This is an equivalent expression of the energy condition. It illustrates that only when input signal vector is close enough to the summation vector of a M/S transform matrix, will this matrix be used. This idea will be helpful when later discussing the 3D-M/S coding where more than one transform matrix exist. Here Thr_v is the corresponding switching threshold of Thr in vectorial distance and

$$Thr_v = \frac{\sqrt{2}}{2}\left(\frac{1}{\sqrt{1+10^{\frac{Thr}{10}}}} + \frac{1}{\sqrt{1+10^{-\frac{Thr}{10}}}}\right) \tag{6}$$

2.2. M/S coding in three-dimensional space

In stereo and surround audio system, to maintain the stability of sound image, the virtual source always being panned or recorded in two most adjacent channels. So two adjacent channels have the maximum similarity, and M/S coding and parametric coding is always performed based on two channel unit.

For Ambisonics and VBAP system, sound channels are spherically configured in 3D space as shown in figure 1. In VBAP system, three adjacent channels form a directional sound image and have the maximum correlation. In other 3D system like Ambisonics, three channels cover a basic area of 3D space. Three channels should be the basic unit to remove interchannel redundancy rather than two channels in conventional coding schemes. More specifically, in VBAP system the input signal $\mathbf{V_0} = (C_1, C_2, C_3)$ is calculated following sine model in 3D space as

$$\begin{aligned} C_1 &= S\sin\theta\cos\varphi \\ C_2 &= S\sin\theta\sin\varphi \\ C_3 &= S\cos\theta \end{aligned} \tag{7}$$

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 located. But for VBAP model, those possibilities can be reduced to three basic situations. First, the source is panned mainly using one channel. This situation correspond to the virtual source is located right in the position of 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 the condition that only one channel is active in stereo audio, so no transform is performed and M_0 will be used. Second, the source is panned mainly using two channels. This situation correspond to the virtual source is located between two channels, or two channels form a virtual source with one channel which is out of 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 (8). 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 M_4 is designed following the rule of conventional M/S coding. The first vector is the summation of three channels, and rest vectors are orthogonal with the fist vector. It realizes the sum-difference processing for 3D channel, and guarantees that when three channel signals are nearly the same, two channels will just remain the difference signal.

$$\begin{aligned} M_0 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad M_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \\ M_2 &= \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix} M_3 = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ M_4 &= \begin{bmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{3} \end{bmatrix} \end{aligned} \tag{8}$$

All matrixes are designed the same way as conventional M/S switching did. When there are two channels satisfy the conventional M/S switching condition, 3D-M/S coding will select corresponding transform matrix having the nearest distance with the input vector. If all

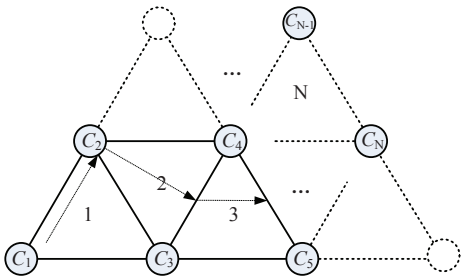

Fig. 2: A general channel configuration of 3D audio system.

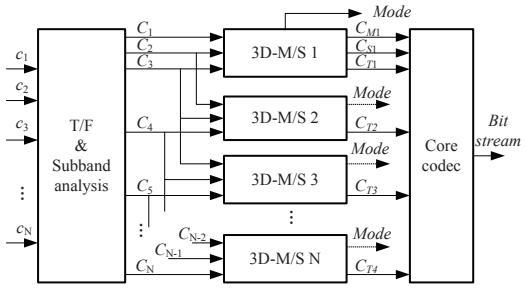

Fig. 3: The framework using 3D-M/S coding for N channels.

channel pair of three channels satisfy the conventional M/S switching condition, it suggests that all the three channel signals are nearly the same and new transform matrix will be chosen. Following the vector distance switching condition, the switching rule of 3D-M/S can be denoted as

$$\text{mode} = \begin{cases} M_4,\ if\ \langle \mathbf{V_0}, \mathbf{V_4} \rangle \geq Thr_v \\ M_i,\ else\ if\ \langle \mathbf{V_{0i}}, \mathbf{V_i} \rangle \geq Thr_v \\ \quad and\ \ \langle \mathbf{V_{0i}}, \mathbf{V_i} \rangle \geq \langle \mathbf{V_{0j}}, \mathbf{V_j} \rangle,\ \ \forall j \neq i \\ M_0,\ else \end{cases} \tag{9}$$

Where $i, j \in \{1, 2, 3\}$, $\mathbf{V_{01}} = (0, C_2, C_3)$, $\mathbf{V_{02}} = (C_1, 0, C_3)$, $\mathbf{V_{03}} = (C_1, C_2, 0)$ are the subsets of $\mathbf{V_0}$. $\mathbf{V_1} = (0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$, $\mathbf{V_2} = (\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2})$, $\mathbf{V_3} = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0)$, $\mathbf{V_4} = (\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3})$ are the summation vectors of each transform matrix.

3. FRAMEWORK FOR GENERAL CHANNEL CONFIGURATION

3D-M/S only works for three channels, actually all 3D audio system contain dozens of channels. A general channel configuration is shown in figure 2, each channel C corresponds to a loudspeaker. Here a framework is proposed based on 3D-M/S coding as shown in figure 3. Since all spatially placed loudspeakers are composed of basic triangle area, this structure will enable 3D-M/S coding works for arbitrary channel configurations.

The framework processes the audio channels triangle by triangle until all channels are coded. C_M is the summation channel, C_S and C_T is the second and third channel. Every 3D-M/S unit shares two channels with previous unit and only one new channel is added in. So it only needs to compress the channel which contains the signal of the new channel. For all matrixes 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 output only one that contains new input channel, the whole coding framework keep the number of channel 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.

4. EXPERIMENT

Considering that PCA is the best decorrelation transform theoretically, the experiment compares the proposed 3D-M/S method with PCA method and independent channel coding in bitrate, complexity and objective quality. Complexity is measured by the running time of each method on PC (CPU: Intel Core2 Duo P8600 2.53GHz, RAM: 8GB). Objective quality is measured by the SNR. A synthetic VBAP signal is used as test 3D audio signal, which panning three MPEG test sequences (es01 voice, sc03 symphony, si02 castanets, 48kHz sampling) as virtual source following VBAP rule. The experiment uses five channel configuration (C_1, C_2, C_3, C_4, C_5) for testing as shown in figure 2. This experiment is not able to cover all possible position where a virtual source located, so three basic sound source movements are used. The three movements in a triangle are point to point, point to edge and edge to edge as shown in figure 2.

The 3D-M/S and PCA is used on each subband in frequency domain. The encoders are realized based on FAAC-1.28, and decoders are based on FAAD2-2.7. AAC-LC is used as core codec and only long window is enabled for simplification. To avoid the influence of dynamic bandwidth setting of FAAC, the experiment fix the bandwidth at 12kHz with 35 subbands. For 3D-M/S, 3 bits is used for one mode parameter. The vector is calculated using the subband energy of three channels from their psychoacoustic module, and do not require extra computation. For PCA method, the eigenvectors are calculated for each subband and subband signals are transformed using eigenvector matrix. The covariance matrix is quantized and transmitted to decoder following a previous KLT based multichannel audio coding scheme [9], with 4 bits per non-redundant element.

Figure 4 shows the average SNR per frame of three active channels. Movement 1 is approximately from 1 to 400 frame, movement 2 is 400 to 800 frame, movement 3 is 800 to last frame. All SNR curves have a downtrend, because transient signal and symphony signal have more abundant frequent components, so they are more difficult to compress. When the virtual source come close to the middle of two channels (between C_1, C_2 around 200 frame, between C_4, C_5 from 400 to 800 frame), 3D-M/S get higher SNR than independent channel coding. Especially around 600 frame, where all three channels are nearly the same, M_4 can remove the redundancy to the largest extent and outperform the PCA method. It is because by removing the interchannel redundancy, more bits could be reserved for summation channel. Table 1 shows the bitrate and complexity of each method. Bitrates of three methods are set to be nearly the same to compare the performance, and ODG scores are calculated for the whole signals. It can be observed that both PCA and 3D-M/S method get about 0.66 ODG improvement. However, 3D-M/S only increase 11.3% codec complexity compared with 29.9% of the PCA method to achieve similar performance. Finally, the PCA parameters bitrate 39.3kbps/channel is considerably high than 3D-M/S method. If the three channels has little correlation (eg channels with different contents), the transformed signals will not

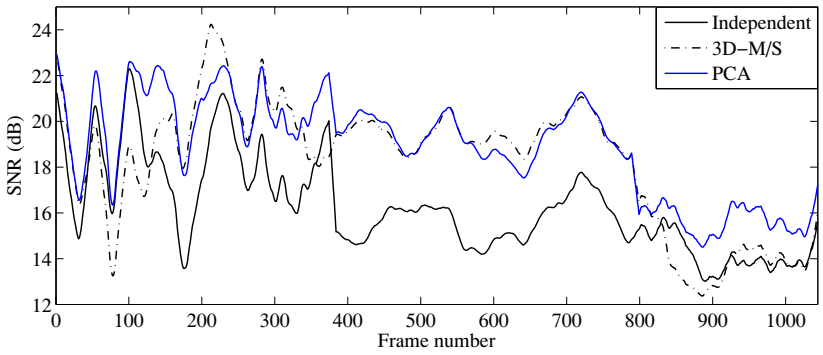

Fig. 4: Average SNR of three channels per frame of Independent channel coding, 3D-M/S and PCA method.

Table 1: ODG quality, bitrate and complexity of three methods. For PCA and 3D-M/S, C_1 C_2 C_3 share a total bitrate, and C_4 C_5 are the C_T channels. P_1, P_2 and P_3 are the parameter bitrates for triangles 1, 2 and 3.

	Quality(ODG)	Bitrate/channel(kbps)									Complexity(s)		
		C_1	C_2	C_3	C_4	C_5	P_1	P_2	P_3	Avg	Encoder	Decoder	Ratio
Ind	-2.558	61.6	60.8	60.3	58.7	58.7	–	–	–	61.2	2.382	0.223	100.0%
3D-M/S	-1.909	229.7			21.9	57.6	4.9	4.9	4.9	64.8	2.597	0.302	111.3%
PCA	-1.871	133.4			25.6	42.0	39.3	39.3	39.3	63.8	2.973	0.411	129.9%

save any bits and cause the decrease of coding efficiency. But for 3D-M/S, parameter bits for modes are only 4.9kbps/channel. It will not reduce the coding efficiency much for medium and high bitrate conditions, which is main application scenario of M/S coding.

5. RELATION TO M/S CODING AND 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 conventional method. This structure is more suitable for 3D audio channel 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 decompose the surface where audio channels located into triangles. Combining the proposed framework, 3D-M/S can encode more than three channels. Experiment on VBAP signals verify 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 audio codec. In the future, bit reserve method and masking threshold for 3D-M/S will be studied for better performance.

6. REFERENCES

[1] S. Sakaida, K. Iguchi, N. Nakajima, Y. Nishida, A. Ichigaya, E. Nakasu, M. Kurozumi, and S. Gohshi, "The super hi-vision codec," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 16 2007-oct. 19 2007, vol. 1, pp. I –21 –I –24.

[2] M.M. Goodwin and J. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, april 2007, vol. 1, pp. I–9 –I–12.

[3] Bin Cheng, C. Ritz, and I. Burnett, "A spatial squeezing approach to ambisonic audio compression," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 369 – 372.

[4] E. Hellerud, A. Solvang, and U.P. Svensson, "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*, april 2009, pp. 269 –272.

[5] F. Pinto and M. Vetterli, "Wave field coding in the spacetime frequency domain," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 365 –368.

[6] F. Pinto and M. Vetterli, "Space-time-frequency processing of acoustic wave fields: Theory, algorithms, and applications," *Signal Processing, IEEE Transactions on*, vol. 58, no. 9, pp. 4608 –4620, sept. 2010.

[7] Bin Cheng, *Spatial squeezing techniques for low bit-rate multichannel audio coding*, Ph.D. thesis, University of Wollongong, 2011.

[8] J.D. Johnston and A.J. Ferreira, "Sum-difference stereo transform coding," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, mar 1992, vol. 2, pp. 569 –572 vol.2.

[9] Dai Yang, Hongmei Ai, C. Kyriakakis, and C.-C.J. Kuo, "High-fidelity multichannel audio coding with karhunen-loeve transform," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 4, pp. 365 – 380, july 2003.