Distortion Analysis of Hierarchical Mixing Technique on MPEG Surround Standard

Ikhwana Elfitri, Mumuh Muharam, Muhammad Shobirin
Department of Electrical Engineering, Faculty of Engineering, Andalas University
Kampus UNAND Limau Manih, Padang, 25163, West Sumatera, Indonesia
E-mail: ikhwana@ft.unand.ac.id

***Abstract*—Distortion due to mixing processes in Moving Picture Expert Group (MPEG) Surround, an audio standard based on spatial audio coding technique, has been studied and reported in this paper. The distortion, particularly introduced due to the hierarchical down-mixing technique applied in MPS, was analytically and experimentally studied. Experiments showed the results consistent to the analysis based on the derived formulation. Tested using 5 audio materials, it is shown that approximately 4 dB additional distortion was introduced in encoding 5 audio channels compared to encoding of 2 audio channels. The results also shows that the distortion, represented as squared-error and then plotted, can be visually identified. This study is very useful for future work on improving MPS performance.**

I. Introduction

Various high quality spatial audio coding techniques [1]–[4] have been proposed in the last decade that include a Moving Picture Expert Group (MPEG) standard, called MPEG Surround (MPS) standard [5]–[9]. While MPS developed based on spatial audio coding [10]–[12] as channel-based method, object-based audio coding technique [13]–[16] has also emerged as a popular option for future audio applications. This new object-based method is promising in that it offers many new features such as a possibility for audio remixing. However, MPS still play important role as it can be employed in an object-based audio coding scheme such as the one in MPEG standard for spatial audio object coding.

Considering the importance of MPS, in this paper, a study on mixing distortion that introduced in MPS is presented. The distortion is studied analytically and experimentally. The paper is started with a brief discussion on MPS standard in Section II. Mathematical derivation of distortion introduced in hierarchical mixing technique in MPS is given in Section III. The results of experiments are provided in Section IV followed by Section V with conclusion and future work.

II. Overview of MPEG Surround

MPEG Surround (MPS) is an advance MPEG standard for encoding multichannel audio signals that works based on a principle of spatial audio coding. Instead of individually encoding every channel of multichannel audio, spatial audio benefits from only encoding the down-mixed signals which can be a mono or stereo audio signals. In order to be able to properly re-create every channel of multichannel audio at the decoder side, spatial parameters as well as residual signal are essential. In particular, the residual signal is highly important for fully reconstruction of the audio waveform.

In general, at least 4 advantages of MPS can be highlighted. First, due to the need of only transmitting the down-mix signal accompanied with spatial parameters and residual signal without sending every single channel of the multichannel audio, MPS can transmit multichannel audio efficiently. For instance 5.1 audio channels can be transmitted at a total bitrate as low as of 64 kb/s when MPS is employed in combination with HE-AAC [17]. Second, with the method of generating down-mix signals, MPS has a backward compatibility that means that it can be applied on existing mono or stereo audio transmission to introduce new multichannel audio rendering. Third, MPS standard also provide many features such as a capability to offer end users with binaural audio rendering [6] that is very useful to allow mobile terminal in experiencing of surround audio scene. Fourth, down-mixing and up-mixing processes are performed for every frequency-band that is made to be similar to that of the critical-band of human hearing system. For this purpose, efficient filter-bank is applied to decompose a full-band audio signal into a number sub-band signals where each of them having the desired frequency-band.

In details, the MPS can be simply illustrated based on its block diagram as given in Fig. 1. Spatial parameters, consists of channel level difference (CLD), interchannel coherence (ICC) and channel prediction coefficient (CPC), are first extracted at the encoder side. Then, all channels are down-mixed based on the extracted spatial parameters. Residual signals can be determined to compensate for the distortion due to the down-mixing process. The down-mix signal is subsequently encoded by a core mono or stereo encoder and then transmitted with the spatial parameters and residual signal as side information. At the decoder side, multiple audio channels can be re-generated from the decoded down-mix audio signal based on the information provided by the spatial parameters and the residual signal.

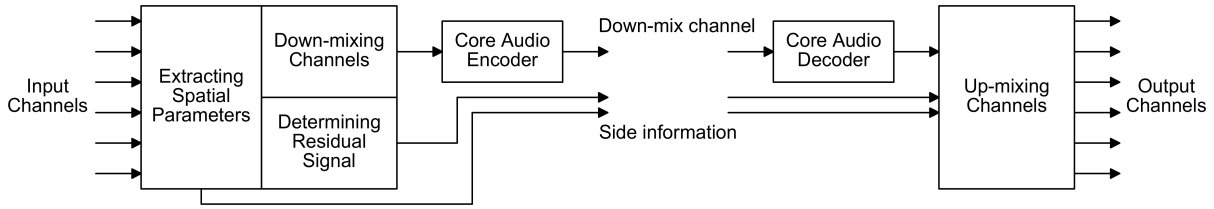

Fig. 1. Basic block diagram of MPEG Surround for a case of 6 input and output channels.

As a standard, MPS specifies two modules, called one-to-two (OTT) and two-to-three (TTT), for up-mixing audio channel at the decoder side. An OTT module can be used to up-mix a single audio channel into two channels while a TTT module can be applied to convert stereo audio channel into three channels. These modules can be employed in tandem to upmix a mono or stereo audio channels into a bigger number of audio channels. For instance, to re-create 5 audio channels from a single mono audio channel, 4 OTT modules can be used in a tree structure. In contrast, the encoder can down-mix multiple audio channels by employing a number of reverse modules: reverse-OTT (R-OTT) and reverse-TTT (R-TTT).

A. Spatial Parameter Extraction

CLD and ICC, denoted as C_{12} and I_{12}, are calculated in an R-OTT module, having two input signals $x_1(n)$ and $x_2(n)$, as follow,

$$C_{12} = \frac{\sum_n x_1(n) \cdot x_1^*(n)}{\sum_n x_2(n) \cdot x_2^*(n)} \tag{1}$$

and,

$$I_{12} = \frac{\sum_n x_1(n) \cdot x_2^*(n)}{\sqrt{\sum_n x_1(n) \cdot x_1^*(n) \sum_n x_2(n) \cdot x_2^*(n)}} \tag{2}$$

where n is the index of audio samples while the sign of $(^*)$ represents a complex conjugate operation.

B. Down-mixing Method

Down-mixing two audio signals in an R-OTT module means adding both input channels where each channel is scaled as,

$$y_{12}(n) = \frac{x_1(n)}{a_{12} + b_{12}} + \frac{x_2(n)}{a_{12} + b_{12}} \tag{3}$$

where the energy constants, a_{12} and b_{12}, are introduced to keep the energy of input and outpul channels are equal.

C. Residual Signal Generation

Residual signal $r_{12}[n]$ is determined to compensate the error due to mixing process and can be counted from the following decomposition:

$$x_1[n] = a_{12} \cdot y_{12}[n] + r_{12}[n] \tag{4a}$$
$$x_2[n] = b_{12} \cdot y_{12}[n] - r_{12}[n] \tag{4b}$$

so that a single residual signal can be created for reconstructing both $x_1[n]$ and $x_2[n]$.

D. Up-mixing process

At the decoder side, replica of both audio signals, $\hat{x}_1[n]$ and $\hat{x}_2[n]$, are recreated in an OTT module from the decoded down-mix signal, $\hat{y}[n]$, and the decoded residual signals, $\hat{r}[n]$, as follows:

$$\hat{x}_1[n] = \hat{a}_{12} \cdot \hat{y}_{12}[n] + \hat{r}_{12}[n] \tag{5a}$$
$$\hat{x}_2[n] = \hat{b}_{12} \cdot \hat{y}_{12}[n] - \hat{r}_{12}[n] \tag{5b}$$

where the estimated energy constants, $\hat{a}_{12}$ and $\hat{b}_{12}$, are calculated from the quantised values of CLD, $\hat{C}_{12}$, and the quantised values of ICC, $\hat{I}_{12}$. Further details on this calculation can be referred to [7]

III. Analysis of Hierarchical Mixing Error

For a case of encoding 5 audio channels to a mono audio channel, the mixing process i.e down-mixing and up-mixing, can be discussed as follows. Assume that $x_1(n), x_2(n), x_3(n), x_4(n),$ and $x_5(n)$ are audio signals for the first to the fifth channel, respectively. In an upper R-OTT module in layer 1 (please refer to Fig. 2), x_1 and x_2 can be taken as input channels and down-mixed to be a single channel denoted as $y_{12}(n)$ as represented in (3) while a residual signal, $r_{12}(n)$, can be calculated as given in (4). In another lower R-OTT module, the other two audio signals, x_3 and x_4, are down-mixed as $y_{34}(n)$:

$$y_{34}(n) = \frac{x_3(n)}{a_{34} + b_{34}} + \frac{x_4(n)}{a_{34} + b_{34}} \tag{6}$$

and corresponding residual signal , $r_{34}(n)$, is obtained by using similar way.

In layer 2, both resulting downmixed signals, $y_{12}(n)$ and $y_{34}(n)$, are fed to an R-OTT module to create a single audio signal, denoted as $y_{14}(n)$, by the way that:

$$y_{14}(n) = \frac{y_{12}(n)}{a_{14} + b_{14}} + \frac{y_{34}(n)}{a_{14} + b_{14}} \tag{7}$$

From this R-OTT module another residual signal, $r_{14}(n)$, can be created.

For layer 3, the down-mixed signal created in layer 2, $y_{14}(n)$, and the fifth input signals, $x_5(n)$, are down-mixed to produce a final mono down-mix signal as below,

$$y_{15}(n) = \frac{y_{14}(n)}{a_{15} + b_{15}} + \frac{x_5(n)}{a_{15} + b_{15}} \tag{8}$$

while its corresponding residual signal, $r_{15}(n)$, is also calculated.

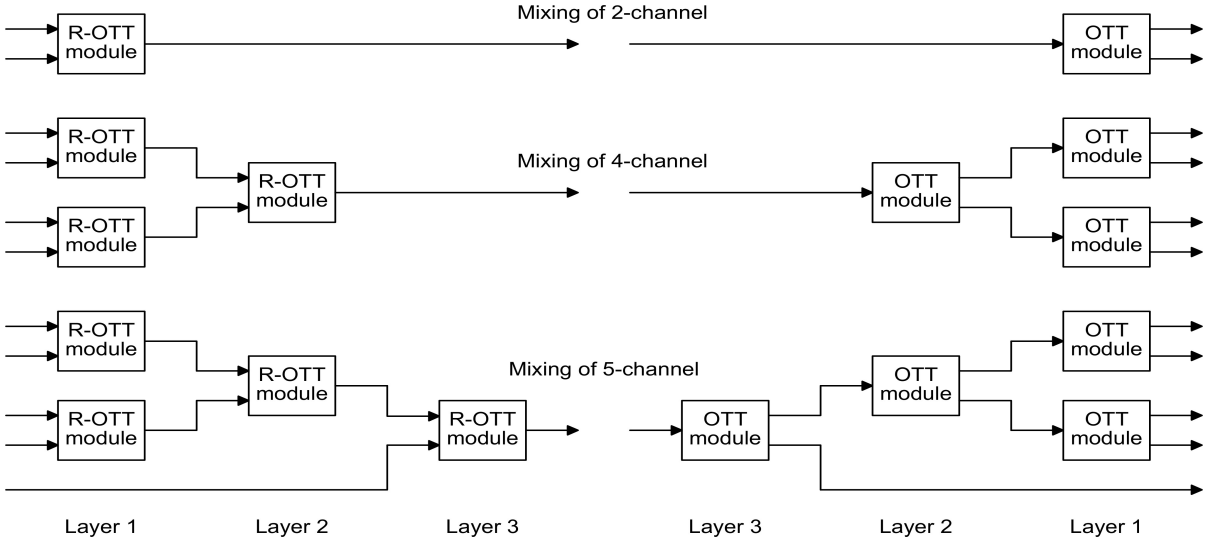

Fig. 2. Structure of the R-OTT modules for down-mixing 2, 4, and 5 audio channels in the encoder and the OTT modules for up-mixing 2, 4, and 5 channels in the decoder.

Incorporating all of (3), (6), (7), and (8), it can be derived that,

$$\begin{aligned} y_{15}(n) = \; & \frac{x_1(n)}{(a_{12}+b_{12})(a_{14}+b_{14})(a_{15}+b_{15})} + \\ & \frac{x_2(n)}{(a_{12}+b_{12})(a_{14}+b_{14})(a_{15}+b_{15})} + \\ & \frac{x_3(n)}{(a_{34}+b_{34})(a_{14}+b_{14})(a_{15}+b_{15})} + \\ & \frac{x_4(n)}{(a_{34}+b_{34})(a_{14}+b_{14})(a_{15}+b_{15})} + \\ & \frac{x_5(n)}{a_{15}+b_{15}} \end{aligned} \tag{9}$$

In the decoder all 5 audio channels can be reconstructed by employing the same number of OTT modules in a reverse way. Decoded down-mix signal, $\hat{y}_{15}(n)$, along with all decoded residual signals, $\hat{r}_{12}(n), \hat{r}_{34}(n), \hat{r}_{14}(n),$ and $\hat{r}_{15}$ are required for this up-mixing. Using (5), $\hat{y}_{14}(n)$ and $\hat{x}_5(n)$ can be obtained by an OTT module at layer 3 as,

$$\hat{y}_{14}(n) = \hat{a}_{15} \cdot \hat{y}_{15}(n) + \hat{r}_{15}(n) \tag{10a}$$
$$\hat{x}_5(n) = \hat{b}_{15} \cdot \hat{y}_{15}(n) - \hat{r}_{15}(n) \tag{10b}$$

Following the same way, $\hat{y}_{12}(n)$ and $\hat{y}_{34}(n)$ can be reproduced by an OTT module in layer 2 as below,

$$\hat{y}_{12}(n) = \hat{a}_{14} \cdot \hat{y}_{14}(n) + \hat{r}_{14}(n) \tag{11a}$$
$$\hat{y}_{34}(n) = \hat{b}_{14} \cdot \hat{y}_{14}(n) - \hat{r}_{14}(n) \tag{11b}$$

Subsequently, $\hat{x}_1(n)$ and $\hat{x}_2(n)$ can be recreated by the first module in layer 1 as,

$$\hat{x}_1(n) = \hat{a}_{12} \cdot \hat{y}_{12}(n) + \hat{r}_{12}(n) \tag{12a}$$
$$\hat{x}_2(n) = \hat{b}_{12} \cdot \hat{y}_{12}(n) - \hat{r}_{12}(n) \tag{12b}$$

while $\hat{x}_3(n)$ and $\hat{x}_4(n)$ can be reconstructed by another OTT module in layer 1 as,

$$\hat{x}_3(n) = \hat{a}_{34} \cdot \hat{y}_{34}(n) + \hat{r}_{34}(n) \tag{13a}$$
$$\hat{x}_4(n) = \hat{b}_{34} \cdot \hat{y}_{34}(n) - \hat{r}_{34}(n) \tag{13b}$$

Based on (10), (11), (12), and (13), the first channel of reproduced audio signals, $\hat{x}_1(n)$, can be simply rewritten as,

$$\begin{aligned} \hat{x}_1(n) = \; & \hat{a}_{12}\hat{a}_{14}\hat{a}_{15} \cdot \hat{y}_{15}(n) + \\ & \hat{a}_{12}\hat{a}_{14} \cdot \hat{r}_{15}(n) + \\ & \hat{a}_{12} \cdot \hat{r}_{14}(n) + \\ & \hat{r}_{12}(n) \end{aligned} \tag{14}$$

while similar way can be done to the other channels.

Using the same method the reconstructed audio signals for mixing of 4-channel and 2-channel audio can be derived. On one hand for mixing of 4-channel, the first channel of reproduced audio signals can be written as,

$$\begin{aligned} \hat{x}_1(n) = \; & \hat{a}_{12}\hat{a}_{14} \cdot \hat{y}_{14}(n) + \\ & \hat{a}_{12} \cdot \hat{r}_{14}(n) + \\ & \hat{r}_{12}(n) \end{aligned} \tag{15}$$

On the other hand for mixing of 2-channel audio, the first channel of recreated audio signals can be written as,

$$\begin{aligned} \hat{x}_1(n) = \; & \hat{a}_{12} \cdot \hat{y}_{12}(n) + \\ & \hat{r}_{12}(n) \end{aligned} \tag{16}$$

In all 3 equations of (14), (15), and (16), it can be seen that $\hat{x}_1(n)$ of mixing of 5-channel depends on more number of spatial parameters and residual signals than both mixing of 4-channel and mixing of 2-channel. Moreover, mixing of 4-channel seems to be affected by more number of spatial parameters, which showed by $\hat{a}_{12}$ and $\hat{a}_{14}$, than mixing of 2-channel. Higher number of spatial parameters and

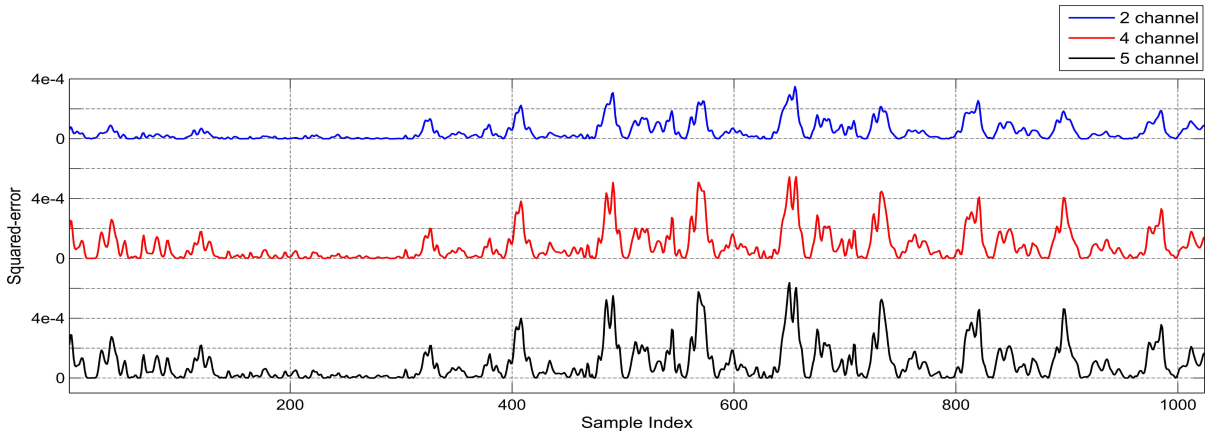

Fig. 3. Distortion, represented as squared-error, introduced in the first channel in the mixing of 2-channel, 4-channel, and 5-channel. These distortion were taken of Speeches audio excerpt for the 104^{th} frame.

residual signals in (14) indicates that larger amount of quantisation noise and encoding losses contributes in the distortion. This suggest that more number of input audio channels suffers from more amount of error introduced during the quantisation of the spatial parameters.

IV. Results

A. Experimental Setup

In order to show the effect of hierarchical mixing method, three experiments have been conducted that are mixing of 2-channel, mixing of 4-channel, and mixing of 5-channel. The schematic of the R-OTT and OTT modules simulated in Matlab are as given in Fig. 2. In the experiments, continuous down-mix and residual signals, without encoding, were transmitted while the spatial parameters were quantised as specified in MPS standard. Five audio excerpts, as discussed in the result section, were used. Signal-to-distortion ratio (SDR), defined as a comparison of energy of input audio signal to energy of error signal for a frame of 1024 samples, is used as a metric for benchmarking. The SDR measured in all frames are then averaged for the whole duration of audio signal.

B. Analysis of Signal-to-Distortion Ratio

TABLE I
Signal-to-Distortion Ratio (SDR) in decibel (dB) for 3 Different Mixing-Schemes

Audio Material	Mixing of 2-channel	Mixing of 4-channel	Mixing of 5-channel
Acoustic	19.26	19.26	19.24
Applause	35.15	27.19	26.73
Classics	27.21	24.98	22.42
Laughter	26.57	25.35	24.89
Speeches	30.24	26.95	26.61
Average	27.69	24.75	23.98

Table I provides SDR measured in the experiments. These results show that, in average, approximately 4 dB of further distortion in terms of SDR are introduced in the Mixing of 5-channel compared to the Mixing of 2-channel. However, the amount of distortion introduced in each audio excerpt tends to be signal dependence. The maximum effect of hierarchical structure on the SDR is shown in the Applause while the minimum effect is seen in Acoustic. In Fig. 3, distortion represented as squared-error, introduced on a frame of audio signal due to mixing process in the first channel of Speeches audio excerpt are plotted. The results demonstrate that the level of additional distortion introduced in both Mixing of 4-channel and Mixing of 5-channel can be easily identified compare to distortion caused in Mixing of 2-channel. The results indicate that the hierarchical mixing technique applied in MPS introduces significant distortion. The results indicate that if MPS is employed to encode larger number of audio channels the distortion could be higher which motivates for finding an approach for reducing the distortion.

V. Conclusion

This paper has presented a study on MPEG Surround mixing technique and found that the distortion introduced due to the hierarchical mixing method was significant. This has been studied analytically and it has been found that additional distortion, in terms of signal-to-distortion ratio (SDR), as many as 4 dB in average, was introduced due to hierarchical mixing method. This has suggested future work for finding a method to either reduce the distortion or compensate the error in a suitable way in order to improve the performance of MPS.

Acknowledgment

This work was funded by the Ministry of Education and Culture, Republic of Indonesia, under DIPA Universitas Andalas, contract no. 023.04.2.415061/2014. The authors thank anonymous reviewers for the constructive comments and suggestions.

REFERENCES

[1] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn, "Perceptual coding of high-quality digital audio," Proceedings of the IEEE, vol. 101 No. 9, pp. 1905–1919, 2014.

[2] J. Herre, "Audio coding: An all-round entertainment technology," in Proc. the 22^{th} Int. Conf.: virtual, synthetic, and entertainment audio, Espoo, Finland, June 2002.

[3] T. Painter and A. Spanias, "Perceptual coding of digital audio," Proc. of the IEEE, vol. 88, no. 4, pp. 451–513, April 2000.

[4] M. Bosi, "High-quality multichannel audio coding: trends and challenges," J. Audio Eng. Soc., vol. 48 No. 6, pp. 588–595, 2000.

[5] J. Hilpert and S. Disch, "The MPEG Surround audio coding standard [Standards in a nutshell]," IEEE Signal Processing Mag., vol. 26, no. 1, pp. 148–152, Jan. 2009.

[6] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjorling, J. Plogsties, and J. Koppens, "Multi-channel goes mobile: MPEG Surround binaural rendering," in Proc. the AES 29^{th} Int. Conference, Seoul, Korea, September 2006.

[7] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. V. de Par, "Background, concepts, and architecture for the recent MPEG Surround standard on multichannel audio compression," J. Audio Eng. Soc., vol. 55, pp. 331–351, 2007.

[8] J. Roden, J. Breebart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Holzer, "A study of the MPEG Surround quality versus bit-rate curve," in Proc. the 123^{th} Convention of the Audio Engineering Society, New York, USA, Oct. 2007.

[9] S. Quackenbush and J. Herre, "MPEG Surround," IEEE Trans. on Multimedia, vol. 12, no. 4, pp. 18–23, 2005.

[10] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in Proc. IEEE Int. Conf. on Multimedia and Expo, San Fransisco, CA, USA, Oct. 2007.

[11] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," in Proc. the 117^{th} Convention of the Audio Engineering Society, San Fransisco, CA, USA, Oct. 2004.

[12] J. Herre, "From joint stereo to spatial audio coding - recent progress and standardization," in Proc. of the 7^{th} Int. Conf. on Digital Audio Effects (DAFx'04), Naples, Italy, October 2004.

[13] S. Gorlow, E. A. P. Habets, and S. Marchand, "Multichannel object-based audio coding with controllable quality," in Proc. 2013 IEEE Int. Conf. Acoustics, Speech and Signal Proc., Vancouver, Canada, June 2013.

[14] J. Herre and L. Terentiv, "Parametric coding of audio objects: Technology, performance, and opportunities," in Proc. the 42^{nd} Int. Conference: Semantic Audio, Ilmenau, Germany, July 2011.

[15] J. Herre, C. Falch, D. Mahne, G. del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," in Proc. the 128^{th} Convention of the Audio Engineering Society, London, UK, May 2010.

[16] J. Engdegard et al., "Spatial audio object coding (SAOC)-The upcoming MPEG standard on parametric object based audio coding," in Proc. the 124^{th} Convention of the Audio Engineering Society, Amsterdam, The Netherlands, May 2008.

[17] J. Herre and M. Dietz, "MPEG-4 high-efficiency AAC coding," IEEE Signal Proc. Mag., vol. 25, no. 3, pp. 137–142, 2008.