MPEG Surround Extension Technique for MPEG-H 3D Audio

Seungkwon Beack, Jongmo Sung, Jeongil Seo, and Taejin Lee

In this paper, we introduce extension tools for MPEG Surround, which were recently adopted as MPEG-H 3D Audio tools by the ISO/MPEG standardization group. MPEG-H 3D Audio is a next-generation technology for representing spatial audio in an immersive manner. However, considerably large numbers of input signals can degrade the compression performance during a low bitrate operation. The proposed extension of MPEG Surround was basically designed based on the original MPEG Surround technology, where the limitations of MPEG Surround were revised by adopting a new coding structure. The proposed MPEG-H 3D Audio technologies will play a pivotal role in dramatically improving the sound quality during a lower bitrate operation.

Keywords: MPEG Surround, Spatial audio coding, MPEG-H 3D Audio.

I. Introduction

The ISO/MPEG standardization group has recently developed new coding technologies for immersive audio under the name MPEG-H 3D Audio [1]. The main feature of immersive audio in MPEG-H 3D Audio is that the spatial sound scene is not restricted to the 2D horizontal plane but is extended to the vertical plane. The extension of the vertical plane for the representation of a real-world surround scene is basically based on a multi-layered loudspeaker setup, which consists of a middle layer for the horizontal plane similar to conventional 5.1 and 7.1 layouts, and an additional lower layer and an upper layer. These two additional layers combined with the middle layer allow a sound scene to realize a vertical representation. This increase in the number of layers results in a considerably large number of speakers from 9.1 to 22.2, and the need to deliver a multitude of sound sources in a compressed form [2], [3]. The first main requirement of MPEG-H 3D Audio is therefore to successfully compress a large number of input channel signals at a given bitrate, and the second is to flexibly represent the decoded signals according to the desired output speaker layout even when the delivered signals do not match the desired target position of the output channels. The first issue is related to how the audio coding efficiency can be achieved despite a considerably large number of input signals, and the second is how to represent the given channel signals according to the desired position and loudness of the outputs. Consequently, both issues have been resolved using the coding and rendering techniques of MPEG-H 3D Audio.

According to [4], the activities of MPEG-H 3D Audio consist of two parts: Phase 1 activity, which is a normal delivery operation mode for a realization of immersive home

Manuscript received Nov. 26, 2015; revised Mar. 16, 2016; accepted Apr. 5, 2016.

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0101-16-0295, Development of UHD Realistic Broadcasting, Digital Cinema, and Digital Signage Convergence Service Technology).

Seungkwon Beack (corresponding author, skbeack@etri.re.kr) and Jongmo Sung (jmseong @etri.re.kr) are with the 5G Giga Communication Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Jeongil Seo (seoji@etri.re.kr) and Taejin Lee (tjlee@etri.re.kr) are with the Broadcasting Media Research Laboratory, ETRI, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.16.2615.0014

theater environments, and Phase 2 activity, which is for the delivery of an immersive sound scene even under a mobile environment transmission. For this reason, the evaluated target bitrates of Phase 1 are set to relatively higher bitrates, that is, 256 kbps, 512 kbps, and 1.2 Mbps. The target bitrates of Phase 2, however, are lower than those of Phase 1, that is, 96 kbps and 128 kbps. Our proposed work focuses on the Phase 2 activity because it is more challenging to successfully compress a large number of input channels under lower bitrates.

Regarding the Phase 1 activity, to successfully encode the input channels, the Unified Speech and Audio Coding (USAC) scheme was adopted as a core coding module combined with prediction-based stereo coding tools [5], which is specifically called USAC_3D for MPEG-H 3D Audio. This scheme has demonstrated a remarkable level of performance. For Phase 2, MPEG-H 3D Audio has adopted the combined technologies of USAC and MPEG Surround. In this case, MPEG Surround first encodes multichannel signals using spatial cues such as the channel level difference (CLD) and inter-channel coherence (ICC), and outputs a downmix signal encoded using USAC _3D. In this case, the performance does not show a sufficiently promising level of quality owing to the limitations of the MPEG Surround structure, which is described in detail in the following section. Motivated by this limitation, we propose extending the coding structure of MPEG Surround and combining it with USAC_3D core coding as a parametric coding mode. The proposed structure shows a dramatically improved sound quality at the target bitrates. As a result, it was adopted as a coding tool for MPEG-H 3D Audio by the ISO/MPEG Audio group under the name MPEG Surround Extension for 3D Audio [6]. In this paper, we introduce the proposed MPEG Surround Extension technology and provide subjective results confirming its level of performance. The remainder of this paper is organized as follows: Section II provides an overview of the architecture of MPEG-H 3D Audio, with a focus on the coding structure. Section III discusses some limitations of the current MPEG Surround specifications when combined with MPEG-H 3D Audio. Section IV describes a technical perspective of the proposed structure, and finally, Sections V and VI show the evaluation results of the proposed method and provide some concluding remarks regarding this research.

II. Overview of MPEG-H 3D Audio

The architecture of MPEG-H 3D Audio is relatively complicated compared with other MPEG audio tools because it should successfully compress various types of input signals and represent the desired immersive sound scene through the rendering process [7]. The possible input types of MPEG-H 3D Audio are channel-, object-, and high-order ambisonic (HOA) based signals. To reproduce a sound scene based on object-based signals, the additional metadata need to include sound scene information such as the corresponding position and loudness of the delivered audio objects. HOA-based signals also require additional HOA parameter information for reconstructing channel-based signals corresponding to the desired output channel layouts. A more detailed structure and the functionality, including the rendering process, can be found in [1] and [7]. Apart from the rendering issue, the scope of the present study is limited to an overview of the channel compression process of MPEG-H 3D Audio. All input types can be encoded using the USAC_3D module, which is a revised version of USAC that adopts one additional channel coding tool, that is, quad channel element (QCE) coding, which was newly integrated in the USAC coding scheme. QCE can jointly encode two pairs of stereo signals concurrently. Therefore, the coding efficiency of MPEG-H 3D Audio is mainly derived from USAC_3D core encoding. Figure 1 shows the high-level architecture of the MPEG-H 3D Audio coding process, which is a simplified version in terms of the channel-based compression process. At a higher bitrate of around 256 kbps and above, USAC_3D is solely applied as a compression module; for a lower bitrate operation (for example, below 128 kbps), however, MPEG Surround is first applied to compress the multitude of input signals with a high compression ratio. Although the data rate of MPEG Surround varies depending on the encoding mode, MPEG-H 3D Audio only supports a low bitrate mode. A rate of only around 2 kbps is needed to encode two channel signals using parametric encoding mode from MPEG Surround. For instance, a 10-channel signal input into the MPEG Surround encoder is converted to around 10 kbps for the spatial parameter bitstream, and for mono or stereo downmix (DMX) signals. The DMX signals, which are outputs of MPEG Surround, are inputted into the USAC_3D encoder, and the spatial parameter bitstream is transported to a

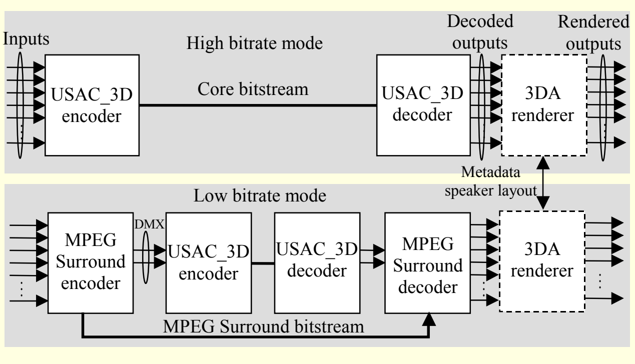


Fig. 1. High-level block diagram of MPEG-H 3D Audio encoder/decoder.

USAC bitstream as an extended container defined in the USAC bitstream payload [5].

III. Limitations of MPEG Surround for MPEG-H 3D Audio

MPEG Surround is multichannel audio coding with a high compression ratio [8]. The main coding efficiency is achieved by utilizing spatial cue parameters, such as CLD and ICC. The number of parameters is estimated according to the number of analyzed subbands, that is, from 5 to 28, and thus a set of channel pair signals can be represented through a single mono DMX and spatial parameters of a few kbps. If more bits are available for allocation during the encoding stage, residual signals can be delivered together with the spatial parameters after encoding through AAC. In our work, we mainly consider the use of spatial-cue based parametric coding mode to achieve a high compression ratio for cases in which the input audio layout has a minimum of ten channels. To briefly review the decoding process of MPEG Surround, a one-to-two (OTT) box is used as the basic encoding unit through the following matrix operation:

$$\begin{bmatrix} ch_1 \\ ch_2 \end{bmatrix} = \underbrace{\begin{bmatrix} H_{\mathrm{LL}} & H_{\mathrm{LR}} \\ H_{\mathrm{RL}} & H_{\mathrm{RR}} \end{bmatrix}}_{\mathbf{H}_{\mathrm{OTT}}} \begin{bmatrix} \mathrm{DMX} \\ \mathrm{D}(\mathrm{DMX}) \end{bmatrix}. \quad (1)$$

In (1), D(·) is an operation used to generate decorrelated signals by applying a decorrelator filter to the DMX signals [8]. A decorrelated DMX filter can be used to control the width of a spatial sound image owing to its complete non-correlation with the delivered DMX signals, allowing the portion of decorrelated signals within the synthesis output to determine the degree of diffuseness of the sound image. However, an undesirable mixing between the DMX and decorrelated signals can degrade the sound quality because decorrelated signals, artificially generated from DMX signals, cause a loss of fidelity in the original DMX. Upmix matrix $\mathbf{H}_{\mathrm{OTT}}$, whose elements are calculated from the CLD and ICC parameters for each sub-band, can successfully control the number of decorrelated signals in the DMX signals, producing output signals ch_1 and ch_2. One of the properties of $\mathbf{H}_{\mathrm{OTT}}$ is that H_{RR} and H_{LR} are complementary versions of each other. This complementary property can be used to reconstruct a DMX from ch_1+, which indicates that the artificial part of a synthesis signal can be discarded if we want to return back to a DMX signal, a process that is normally requested in the rendering part of MPEG-H 3D Audio for flexible channel-based rendering [7].

The main constraint of MPEG Surround technology is that the number of DMX signals, that is, the input to the MPEG Surround decoder, should provide backward compatibility for legacy audio coding delivery systems such that the DMX is restricted to mono or stereo signals even when the number of original input channels is considerable, such as for 22.2-channel systems. In some specific cases, the number of DMX channels as inputs of the MPEG Surround decoder can be extended. As an example, a DMX signal can be delivered as a 5.1 signal, but only for 7.1-channel output. In addition, an extension of the DMX channels from 5.1 DMX can be handled by adopting the “arbitrary tree coding mode” of MPEG Surround [9]. However, the spatial quality of a sound

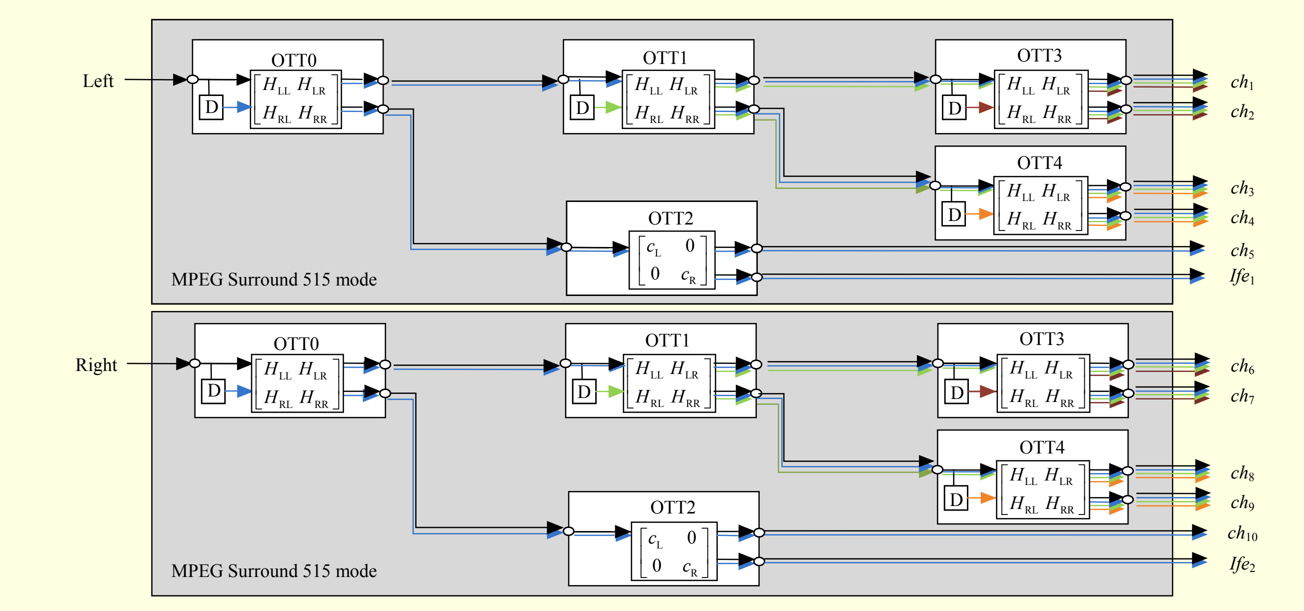


Fig. 2. Example block diagram of 515 parametric mode based MPEG Surround synthesis process for 10.2 sound generation.

scene cannot be achieved owing to the absence of a decorrelated process in arbitrary tree coding mode used for controlling the spatial width.

Figure 2 shows one of the possible configurations when using MPEG Surround 515 coding mode for a 10.2 input channel layout, which is the multichannel format adopted as the speaker layout for the next-generation broadcasting system in Korea [10]. To produce 10.2 outputs, two sets of 515 MPEG Surround decoding modes can be used to maintain backward compatibility for stereo DMX. With the exception of the *lfe* channel, one of the outputs in an example case can be composed of the following:

$$\begin{aligned} Ch_i = {} & H_{\mathrm{LL}}^{2}\left[H_{\mathrm{LL}}^{1}\left(H_{\mathrm{LL}}^{0}x_1 + H_{\mathrm{LR}}^{0}\mathrm{D}_0(x_1)\right) + H_{\mathrm{LR}}^{1}\mathrm{D}_1\left(H_{\mathrm{LL}}^{0}x_1 + H_{\mathrm{LR}}^{0}\mathrm{D}_0(x_1)\right)\right] \\ & + H_{\mathrm{LR}}^{2}\mathrm{D}_2\left[H_{\mathrm{LL}}^{1}\left(H_{\mathrm{LL}}^{0}x_1 + H_{\mathrm{LR}}^{0}\mathrm{D}_0(x_1)\right) + H_{\mathrm{LR}}^{1}\mathrm{D}_1\left(H_{\mathrm{LL}}^{0}x_1 + H_{\mathrm{LR}}^{0}\mathrm{D}_0(x_1)\right)\right], \end{aligned} \tag{2}$$

where a superscript indicates the index of an OTT box. From (2), it can be seen that one of the output signals is obtained by applying several mixing processes using a scaled version of the decorrelated signals generated from each OTT. In this case, despite a high compression gain from 515 mode, the output signals of 515 show a limited level of performance [8]. The main reason for such degradation is that artificially generated decorrelated signals are included several times in the final output signals, which can significantly lower the original fidelity of the sound. This results in a severe degradation of the spatial quality when applying a rendering process during MPEG-H 3D Audio decoding, such as in the use of a format converter, because the main process of a format converter during the rendering of MPEG-H 3D Audio is aligning the DMX phase; however, the complicated components derived from decorrelated signals make it difficult to fix the phase alignment between the output channels. This phenomenon can be clearly observed in the case of the 515 structure noted in the MPEG Surround specifications. In the case of this structure, the performance of the sound quality is limited even though the number of allocated bits is increased [8].

IV. Proposed Extension of MPEG Surround

1. Motivation

In the previous section, the design of the current architecture of MPEG Surround under the constraint of mono or stereo DMX support was described. According to the upcoming audio system specifications, however, the new activities of the MPEG audio standardization do not need to maintain backward compatibility for further realized immersive audio such as MPEG-H 3D Audio [11]. This means that the MPEG Surround coding scheme can be extended based on the number of DMX signals even when exceeding the number of stereo channels. Based on the new requirements of MPEG-H 3D Audio, we updated the MPEG Surround coding structure by adopting a new decoding channel configuration, that is, a new tree configuration. Freedom from backward compatibility can provide the opportunity to design a flexibly configured decoding tree structure to accommodate various types of input channel types while supporting a flexible number of DMX channels as outputs of the MPEG Surround encoder.

2. New Tree Configuration for MPEG Surround Extension

A newly proposed tree configuration for extending the MPEG Surround coding scheme was designed to maintain the following two considerations.

- *Minimal number of decorrelated signals per output channel*

Apart from the limitation in sound quality owing to the complicated decorrelated signals used, to successfully support the rendering process, the outputs should include a minimal number of such signals. Although the main scope of our proposal is to increase the compression ratio while maintaining a reliable level of quality, the decoded output signals should also be successfully used during the rendering process. As mentioned in the previous section, the multiple decorrelated signals used in the output channel signals can generate a difficulty in the phase alignment of the format converter, which is a rendering tool used for controlling the number of output channels according to the output playback system used.

Another problem is that object-based output signals can be arbitrary downmixed depending on the rendering information; in addition, because the multiple decorrelated signals in the output signals are also uncorrelated with each other, the uncorrelated parts of the signals cannot be discarded. They may be boosted-up after being downmixed during the rendering process, resulting in a degradation of the spatial image. In the proposed tree structure, to prevent negative artifacts after the rendering process is applied, the number of decorrelated signals per output channel is restricted to one.

- *Low complexity for matrix operation*

The synthesis process of the current MPEG Surround consists of pre-and post-upmix matrix operations. The number of dimensions of the synthesis matrix is determined based on the number of output channel signals. For example, if the number of output channels is N, the number of dimensions of the post-upmix matrix is $N \times N$. This matrix is applied to each time slot of the quadrature mirror filter-bank (QMF) domain with a length of K. Thus, in the current MPEG Surround, the required computation power of the matrix operation used to produce one output sample is $2 \times N \times N \times K$ multiplications. If the number of output channels is increased, the computational

http://dx.doi.org/10.4218/etrij.16.2615.0014

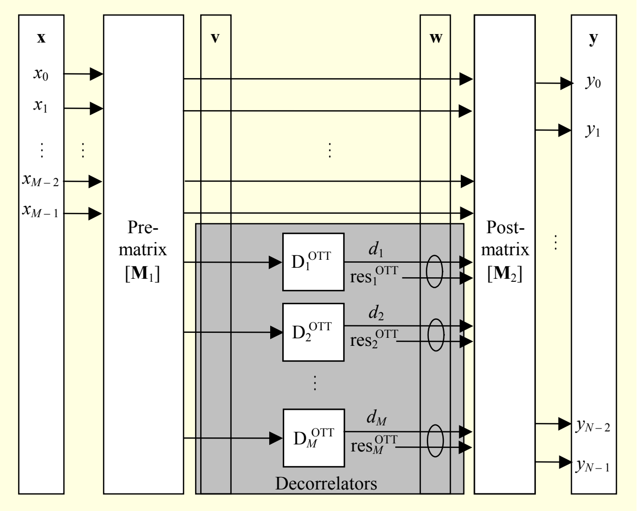


Fig. 3. Matrix view of the proposed decoder structure of MPEG Surround extension.

power of the matrix operation is considerable. It can be observed that the proposed extension provides a simple point-wise upmixing operation, thereby reducing the computational power to a minimum of 2 × N × K.

These considerations are achieved in a straightforward manner by adopting the new parallelized tree configurations in the current MPEG Surround structure. A parallelized tree configuration can be easily reconstructed if the relationship between the number of input and output channels is satisfied as

$$2M = N, \tag{3}$$

where M is the number of DMX signals, and N is the number of output channels. Figure 3 shows our tree configuration as a matrix operation. The numbers of input and output channels in the figure are satisfied by (3), and the number of OTTs is identical to the number of DMX channels. Decorrelated signals $d_i (0 \le i \le M-1)$ of M from each OTT are produced from each OTT box, and are used in a single instance to synthesize each output channel as long as the output channels do not include *lfe* channels. If one *lfe* channel is included in the output of OTT, the OTT box does not produce any decorrelated signals, and thus d_i is $0 \le i \le M-L-1$, where L is the number of *lfe* channels from the output channels. If the residual signals res$_i$ are available, d_i can be replaced by res$_i$. This process, which is based on the proposed tree configuration, can be understood more clearly by reviewing the matrix operation in the following section.

3. Matrix Operation of the Proposed Extension

The proposed extension of MPEG Surround has a parallelized tree configuration. This means that each OTT box does not need to be connected to another OTT box, and the output of each OTT box is dealt with as a DMX signal. The synthesis process is based on the matrix operation, which mainly consists of pre- and post-upmixing matrixes. The pre-upmixing matrix controls the number of decorrelators that generate a decorrelated signal. Regardless of the DMX process used, such as an artistic DMX conversion, the pre-upmixing matrix in the proposed structure can be defined as

$$\mathbf{M}_1^{n,k} = \begin{bmatrix} \mathbf{I}_M \\ ------ \\ \mathbf{I}_{M-nlfe} \end{bmatrix}, \quad 0 \le k < B, 0 \le n < L. \tag{4}$$

In (4), superscripts k and n indicate the index of the processing sub-band of B and the L time slot of QMF, respectively. In addition, $\mathbf{I}_M$ is a unity matrix; and the subscript index indicates the matrix dimension. Here, *nlfe* is the index of the number of *lfe* channels in the output; for example, *nfle* is 2 for a 22.2 system, and therefore *nlfe* should be known to the decoder side to allow this information to be defined in the bitstream syntax. The intermediate vector signal **v** in Fig. 3 can be obtained through the product of $\mathbf{M}_1$ and the expanded input DMX vector **x** as

$$\mathbf{x} = \left[x_0, \dots, x_{M-1}, x_0, \dots, x_{M-Numlfe-1}\right]^{\mathrm{T}}. \tag{5}$$

Another intermediate vector signal **w** is generated through the superposition of the input signals, decorrelated signals, and residual signals according to

$$\mathbf{w}^{n,k} = \begin{bmatrix} v_0^{n,k} \\ v_1^{n,k} \\ \cdots \\ v_{M-1}^{n,k} \\ \delta_0(k)\mathrm{D}_0\left(v_0^{n,k}\right) + \left(1-\delta_0(k)\right)v_{\mathrm{res}_0}^{n,k} \\ \delta_1(k)\mathrm{D}_1\left(v_{M_2}^{n,k}\right) + \left(1-\delta_1(k)\right)v_{\mathrm{res}_1}^{n,k} \\ \cdots \\ \delta_{M-nlfe-1}(k)\mathrm{D}_{M-nlfe-1}\left(v_{M-nlfe-1}^{n,k}\right) + \left(1-\delta_{M-nlfe-1}(k)\right)v_{\mathrm{res}_{M-nfle-1}}^{n,k} \end{bmatrix} = \begin{bmatrix} w_0^{n,k} \\ w_1^{n,k} \\ \cdots \\ w_{M-1}^{n,k} \\ w_{\mathrm{OTT}_0}^{n,k} \\ w_{\mathrm{OTT}_1}^{n,k} \\ \cdots \\ w_{\mathrm{OTT}_{M-nlfe}}^{n,k} \end{bmatrix}, \tag{6}$$

where, if the residual signal is available, the delta function is set to zero. Otherwise, decorrelated signals are used instead of residual signals. Residual signals are commonly used in a parametric synthesis with a high compression ratio. Post-upmixing matrix $\mathbf{M}_2$ is generated from the transmitted spatial information. In the proposed case, $\mathbf{M}_2$ has the following diagonalizable form:

$$\mathbf{M}_2^{l,m} = \begin{bmatrix} \mathbf{H}_{\mathrm{OTT}_0}^{n,k} & \mathbf{O}_2 & \cdots & & \mathbf{O}_2 \\ \mathbf{O}_2 & \ddots & & & \vdots \\ \vdots & & \mathbf{H}_{\mathrm{OTT}_i}^{n,k} & & \\ & & & \ddots & \mathbf{O}_2 \\ \mathbf{O}_2 & \cdots & & \mathbf{O}_2 & \mathbf{H}_{\mathrm{OTT}_{M-1}}^{n,k} \end{bmatrix}, \tag{7}$$

where $\mathbf{O}_N$ is a null matrix whose subscript index indicates the matrix dimension, and $\mathbf{H}^{n,k}_{\mathrm{OTT}_i}$ is a 2 × 2 matrix calculated from the corresponding OTT box using a spatial parameter synthesis as follows:

$$\mathbf{H}^{n,k}_{\mathrm{OTT}_i} = \begin{bmatrix} c^{n,k}_{1,i} \cos\left(\beta^{n,k}_{2,i} - \alpha^{n,k}_{2,i}\right) & c^{n,k}_{1,i} \sin\left(\beta^{n,k}_{1,i} - \alpha^{n,k}_{1,i}\right) \\ c^{n,k}_{2,i} \cos\left(\beta^{n,k}_{2,i} - \alpha^{n,k}_{2,i}\right) & c^{n,k}_{2,i} \sin\left(\beta^{n,k}_{2,i} - \alpha^{n,k}_{2,i}\right) \end{bmatrix}. \quad (8)$$

The gain parameters c_1, c_2 are calculated from the CLD parameters, and angle information α is calculated from the ICC parameter from the cosine angle of the inner product between the two OTT output channels. In addition, β is a parameter used to maintain the balance while satisfying the following:

$$c_1 \sin(\beta - \alpha) = c_2 \sin(\beta + \alpha) \,. \quad (9)$$

Based on (9), the available value of β is specifically selected using a geometric calculation [12]. From (4) and (7), it can be observed that the matrix operation during the upmixing is reducible because a diagnosable matrix operation can be implemented simply using element-wise multiplications. In addition, from (7), one decorrelated signal contributes to the generation of one output channel such that the original fidelity of the output signals is minimally distorted even after an artificially mixed rendering process.

V. Performance Evaluation

1. Subjective Evaluation

For an evaluation of the proposed MPEG Surround Extension, subjective listening tests were carried out according to the MUSHRA methodology [13]. Two bitrates, 96 kbps and 128 kbps, which are within the target bitrates of MPEG-H 3D Audio for maximally encoding 24-channel based contents, were used. Regarding the MUSHRA methodology, the evaluated systems are depicted in Table 1.

In Table 1, SYS1 is the current reference model (RM) of MPEG-H 3D Audio when the target bitrate is below 128 kbps. The same core codec, that is, USAC_3D from MPEG-H 3D Audio, is used for SYS_1 and SYS_2. Next, SYS1 and SYS2 were connected to a legacy MPEG Surround decoder and the proposed extension of MPEG Surround, respectively. Seven test items were selected for the evaluation, as shown in Table 2. All of the items are 24-channel based contents officially used in the development of MPEG-H 3D Audio. In addition, CO_01, CO_02, and CO_03 were recorded from a real-world environment, and the other items were produced by sound engineers for immersive sound scenes using real sound stems from real movie content. All of the test items were selected to evaluate the activities of MPEG-H 3D Audio [4], [14].

Four test labs in the MPEG audio group participated in this evaluation. The number of subjects from each test lab is summarized in Table 3. After a post-screening based on the rule from MUSHRA, 33 and 32 subjects were collected for an analysis of the final pooling result data for rates of 96 kbps and 128 kbps, respectively.

Figures 4 and 5 show the absolute score with a 95% confidence interval for the case of 96 kbps and 128 kbps, respectively. These two bitrates were selected as operation points for the Phase 2 activity of the MPEG-H 3D Audio development [11].

During the item-based observation, the proposed system demonstrated statistically significantly better results for five of the items, showing a non-overlapping confidence interval compared with the RM of MPEG-H 3D Audio. One of the items, CO_02, did not show any improvement because it uses a static sound scene, and thus the effectiveness of the spatial

Table 1. Description of evaluated systems under a subjective test.

System index	System description
SYS1	Initial RM of MPEG-H 3D Audio combined with common MPEG Surround scheme
SYS2	Proposed system of MPEG-H 3D Audio combined with the proposed MPEG Surround Extension
HR	Hidden reference
LP35	3.5 low-pass filtered hidden anchor

Table 2. Test items.

Test items	Remarks
CO_01	Real recorded item in a church
CO_02	Real recorded pipe organ item
CO_03	Real recorded steam train item
CO_04	Music-based sound scene item
CO_05	Aircraft sound scene item
CO_06	Car-racing sound scene item
CO_07	Shooting sound scene item

Table 3. Configuration of subjects participating in the evaluation.

Test labs	Number of subjects (post-screen before/after)	
	96 kbps	128 kbps
Site_1	10/8	9/8
Site_2	10/10	10/10
Site_3	9/9	8/8
Site_4	11/6	10/6
Total	40/33	37/32

http://dx.doi.org/10.4218/etrij.16.2615.0014

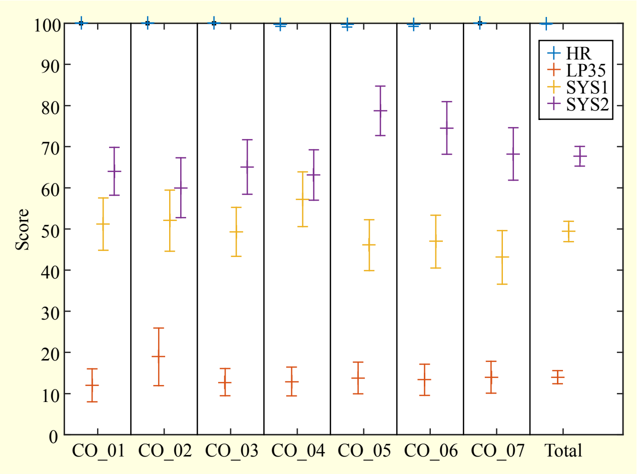


Fig. 4. Average absolute score with 95% confidence interval at 96 kbps.

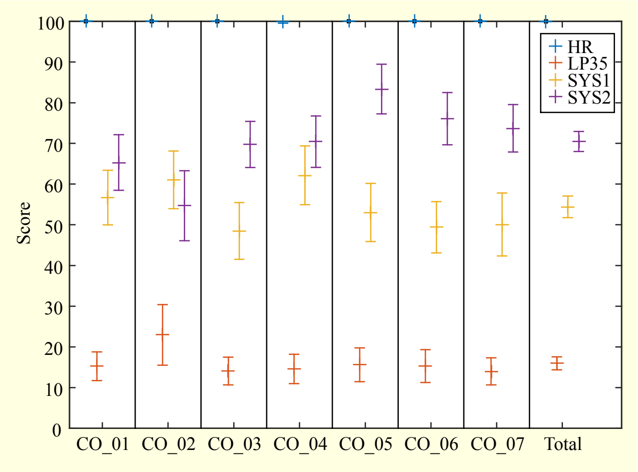


Fig. 5. Average absolute score with 95% confidence interval at 128 kbps.

audio coding gain is relatively lower than for dynamic sound scene based items such as CO_03, CO_05, and CO_06. For the overall average absolute score, the proposed system also demonstrated significantly better results statistically than the RM of MPEG-H 3D Audio for both bitrates. The most remarkable aspect is that the overall quality of SYS1 (RM) showed a "fair" level of performance with an overall confidence interval score of 50 to 60, which is normally considered a tolerable level of quality compared with an original sound. The overall quality of the proposed system can reach a "good" level of quality, however, within the range of 60 to 80.

Figures 6 and 7 show the average difference in score for both bitrates. The difference in score was measured as the difference between the score of the proposed system (SYS2) and the score of the RM (SYS1). This means that a positive difference indicates that the proposed system is statistically better than the RM system, and when the lower bound of the confidence interval shows a non-overlapping confidence interval at a value of zero, we can state that the proposed system is significantly better than the compared systems. From these figures, it can be seen that the quality of five of the items is significantly better at 96 kbps, and that the quality of six of the items is significantly better at 128 kbps, as compared with the RM of MPEG-H.

A remarkable aspect of this analysis is the 16-point improvement in the mean score of the sound quality. Based on these subjective results, MPEG-H 3D Audio adopted the proposed MPEG Surround Extension as a parametric-based multichannel coding tool, particularly for a low bitrate delivery. It can be noted that the main coding gain comes from the coding structure of the proposed extension used to cover all of the input channel signals by utilizing a parallelized form of the MPEG Surround coding tool, as described in Section IV. The

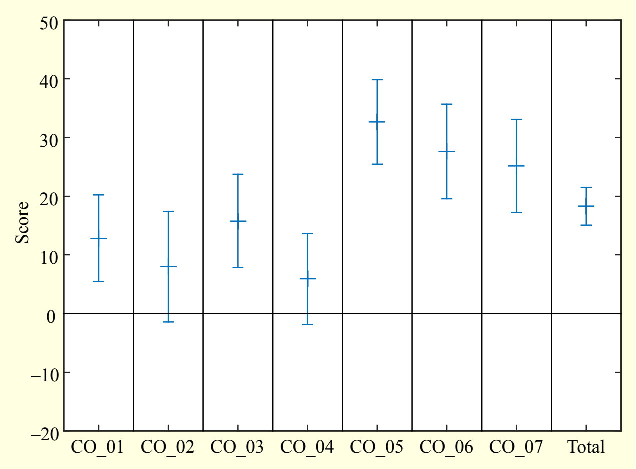


Fig. 6. Average difference score with a 95% confidence interval at 96 kbps.

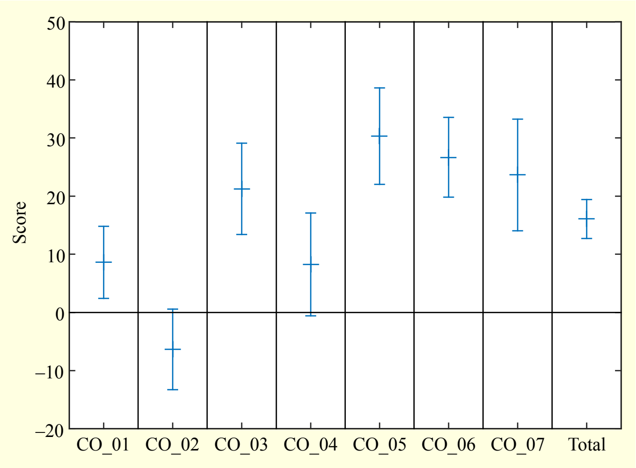


Fig. 7. Average difference score with a 95% confidence interval at 128 kbps.

initial RM cannot cover all of the input channels owing to a structural limitation because the legacy MPEG Surround used in the initial RM is normally constrained to producing 5.1 or 7.1 audio even when a greater number of input channels are used.

2. Complexity

In this section, the computational complexity of the proposed algorithm is analyzed based on the weighted million operations per second. The total computational complexity can be calculated based on the usage of the MPEG-H 3D Audio decoding components according to the decoder configuration. For instance, for decoding 22.2 channels, one of the possible decoder configurations is the direct application of normal MPEG-H 3D Audio core decoding mode, that is, using only the USAC_3D core decoding mode as a unit of channel pair elements (CPE). This means that 12 CPE operations of USAC_3D are required to decode 22.2 channels. Another configuration for 22.2 is to apply the initial RM0 configurations combined with the normal MPEG Surround tool. In this case, five CPE operations are used, and the normal MPEG Surround decoding process is applied four times to generate 22.2. The last configuration is our proposed MPEG Surround Extension tool, which uses six CPE operations and applies the OTT-based spatial decoding module ten times.

The computational complexity of each component (that is, USAC_3D, MPEG Surround, and OTT-based spatial decoding module) can be driven by referring to [9] and [15], and is summarized in Table 4. A comparison of the total complexity is also shown in Table 5. As these tables indicate, the proposed configuration has the least amount of complexity, that is, 0.69, when the total complexity of the initial RM0 case is set to 1.

Table 4. Computational complexity of the decoding components.

Complexity	USAC_3D	MPEG Surround	OTT
WMOPS	13.2 (stereo)	25 (5 channels)	3.23

Table 5. Complexity comparison.

	3DAcore only	Initial RM0	Proposed
# of CPE	11	5	6
# of MPS	N/A	4	N/A
# of OTT	N/A	N/A	10
Complexity	13.2 × 11	13.2 × 5 + 25 × 4	13.2 × 6 + 3.23 × 10
Total (WMOPS)	145.2	166	111.5

VI. Conclusion

MPEG-H 3D Audio is a newly standardized technology of the ISO/MPEG Audio Group for an improvement in immersive sound quality. A large number of audio objects and channels should be accommodated within a single bitstream, and thus MPEG-H 3D Audio requires a more efficient compression tool. An extension of MPEG Surround was proposed to enhance the coding efficiency of MPEG-H 3D Audio, particularly at a lower bitrate. As a result, a remarkable improvement in performance was achieved, and the proposed extension can be adopted as a tool for the pre-encoding system used by MPEG-H 3D Audio. As future work, we will consider extending the residual coding mode through a combination of spatial cues to support a higher bitrate.

References

[1] ISO/IEC 23008-3:2015, *MPEG-H (High Efficiency Coding and Media Delivery in Heterogeneous Environments), Part 3: 3D Audio*, 2015.

[2] B.V. Daele, “The Immersive Sound Format: Requirements and Challenges for Tools and Workflow,” *Int. Conf. Spatial Audio*, Berlin, Germany, Apr. 26–29, 2014.

[3] K. Hamasaki et al., “The 22.2 Multichannel Sounds and Its Reproduction at Home and Personal Environment,” *Int. Conf. Audio Wirelessly Netw. Personal Devices*, Pohang, Rep. of Korea, Sept. 29, 2011, pp. 3–1.

[4] ISO/IEC JTC1/SC29/WG11/N13411, *Call for Proposals for 3D Audio*, Geneva, Switzerland, Jan. 2013.

[5] ISO/IEC 23003-3:2012, *MPEG-D (MPEG Audio Technologies), Part 3: Unified Speech and Audio Coding*, 2012.

[6] ISO/IEC 23003-1:2007, *MPEG-D (MPEG Audio Technologies), Part 1: MPEG Surround, Amendment 3: MPEG Surround Extension for 3D Audio*, 2015.

[7] J. Herre et al., “MPEG-H 3D Audio: the New Standard for Coding of Immersive Spatial Audio,” *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, Aug. 2015, pp. 770–779.

[8] J. Breebaart et al., *MPEG Spatial Audio Coding/MPEG Surround: Overview and Current Status*, New York, USA: Audio Engineering Society, 2005, pp. 770–779.

[9] ISO/IEC 23003-1:2007, *MPEG-D (MPEG Audio Technologies), Part 1: MPEG Surround*, 2007.

[10] ITU-R Recommendation BS. 2051, *Advanced Sound System for Programme Production*, Geneva, Switzerland, 2014.

[11] S. Meltzer et al., “MPEG-H 3D Audio: the Next Generation Audio System,” *IBC Conf.*, Amsterdam, Netherlands, Sept. 11–15, 2014, p. 42.

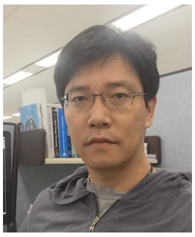
[12] P. Heiko, “Low Complexity Parametric Stereo Coding in MPEG-4,” *Int. Conf. Audio Effects (DAFX-04)*, Naples, Italy, Oct.

5–8, 2004.

[13] ITU-R Recommendation BS.1534-1, *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, Geneva, Switzerland, 2003.

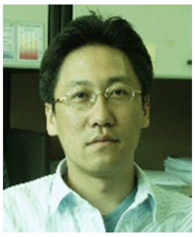
[14] ISO/IEC JTC1/SC29/WG11/N13633, *Submission and Evaluation Procedures for 3D Audio*, Geneva, Switzerland, Jan. 2013.

[15] ISO/IEC JTC1/SC29/WG11/M37167, *Proposal for Profiles and Levels for 3D Audio*, Geneva, Switzerland, Oct. 2015.



Seungkwon Beack received his BS degree in electronic engineering from Korea Aviation University, Koyang, Rep. of Korea in 1999, and his MS degree and PhD from the Department of Information and Communications Engineering at Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea in 2001 and 2005, respectively. He is currently with ETRI, Daejeon, Rep. of Korea. His research interests include audio signal processing, multi-channel audio coding, and representation.

Jongmo Sung received his BS and MS degrees in electronics engineering from Pusan Nation University, Rep. of Korea, in 1995 and 1997, respectively. He received his PhD in mechatronics engineering from Chungnam National University, Daejeon, Rep. of Korea in 2014. Since 1999, he has been working as a principal researcher in the Audio & Acoustics Research Section at ETRI, Daejeon, Rep. of Korea. His research interests cover a wide range of topics in speech and audio signal processing.



Jeongil Seo was born in Goryoung, Rep. of Korea, in 1971. He received his PhD in electronics from Kyoungpook National University, Daegu, Rep. of Korea in 2005 for his work on audio signal processing systems. He was a member of the engineering staff at the Semiconductor Laboratory of LG-Semicon, Cheongju, Rep. of Korea from 1998 to 2000. Since 2000, he has worked as a director at the Interactive Realistic Media Research Section of ETRI, Daejeon, Rep. of Korea. His research activities include image and video processing, audio processing, multi-modal user interface, and realistic broadcasting systems.

Taejin Lee received his BS and MS degrees in electronics engineering from Chonbuk National University, Jeonju, Rep. of Korea in 1996 and 1998, respectively, and his PhD in electronics engineering from Chungnam National University, Daejeon, Rep. of Korea in 2013. He worked for Mobens Co., Ltd., Daejeon, Rep. of Korea from 1998 to 2000. He has been at ETRI since 2000, where he is currently a principal researcher and the director of the Audio and Acoustics Research Section. From 2002 to 2003, he was a visiting researcher at Tokyo Denki University, Japan. His research interests include audio signal processing and interactive broadcasting technologies.