PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Real time audio encoding/decoding system using MPEG-H 3D Audio toward advancement of terrestrial broadcasting technology

Takehiro SUGIMOTO[1]; Shuichi AOKI; Satoshi OODE; Tomomi HASEGAWA;
Hiroki KUBO; Hiroyuki OKUBO;

[1] NHK Science & Technology Research Laboratories, Japan

ABSTRACT

Aiming to enhance the audio service of terrestrial broadcasting, it is considered that a broadcasting system is migrated to an object-based sound system. The object-based sound system in broadcasting is required to (i) deal with audio metadata to control audio data, (ii) consist of highly efficient audio coding, and (iii) be equipped with a rendering feature capable of incorporating a viewer's preferences into audio playback. To meet such requirements, a real time audio encoding/decoding system (codec) using Moving Picture Experts Group (MPEG)-H 3D Audio (3DA) was developed. The developed audio codec supports up to 24 channels of back ground sound with 32 additional audio objects. These audio data are input into the audio codec along with the audio metadata in a serial representation of the Audio Definition Model (S-ADM). The renderer implemented in the audio codec renders decoded audio data into popular audio formats such as 22.2 ch, 7.1.4 ch, 5.1 surround and stereo. Dialogue enhancement and dialogue replacement functions are also made available by the renderer. A music program and sports program were produced to verify the effect of the object-based sound system.

Keywords: Object-based sound system, MPEG-H 3D Audio, Terrestrial broadcasting

1. INTRODUCTION

22.2 multichannel (22.2 ch) sound [1] has been broadcast as a part of the 8K Super Hi-Vision (SHV) satellite broadcasting service in Japan since December 2018. 22.2 ch sound is the largest audio format of practical immersive audio [1]. Currently, it is provided by a channel-based sound system in which each audio signal of a program is one-on-one related to each channel of the audio format. The other audio formats of broadcasting in Japan, such as 5.1 surround and stereo, are also channel-based. As for a channel-based sound system, mixing parameters such as level and localization of the sound source are designed by a balance engineer according to a target audio format. As a result, the mixing parameters cannot be modified after a program is finalized. Given this restriction, all the viewer can do is to listen to the program without being able to change to their preferred mixing balance.

The authors are currently engaged in advancing terrestrial broadcasting technology so that 8K SHV can also be broadcast via terrestrial broadcasting. To meet that challenge, we are aiming to not only extend 22.2 ch sound from satellite broadcasting to terrestrial broadcasting but also establish a scheme for an audio service for incorporating the viewer's preference into the audio playback. We previously studied the listener's preference in regard to the balance between the levels of the dialogue and back ground sound [2]. The result of a subjective evaluation performed in that study revealed that the listener's preference in regard to the level balance is diverse among listeners and a fixed level balance does not always satisfy all listeners. In response to the results, we began to examine adopting the object-based sound system in broadcasting to allow viewers to customize the audio of programs according to their preferences by modifying the mixing parameters.

In this paper, a future vision of broadcast audio brought by an object-based sound system is outlined by introducing examples of programs produced for the object-based sound system. As a core broadcasting technology, Moving Picture Experts Group (MPEG)-H 3D Audio (3DA) [3] was adopted, and a real time encoding/decoding system (codec) based on that technology was developed.

[1] sugimoto.t-fg@nhk.or.jp

The configuration and specifications of the audio codec are presented in detail.

2. BROADCAST CHAIN FOR OBJECT-BASED SOUND SYSTEM

The upper part of Fig. 1 shows the three steps of the broadcast chain of audio which has been unchanged since the early days of broadcasting: first, program production; second, transmission; and third, reproduction. To apply an object-based sound system to broadcasting, these three steps should be supplemented with the new features shown in the lower part of the figure. These three features are described in detail below.

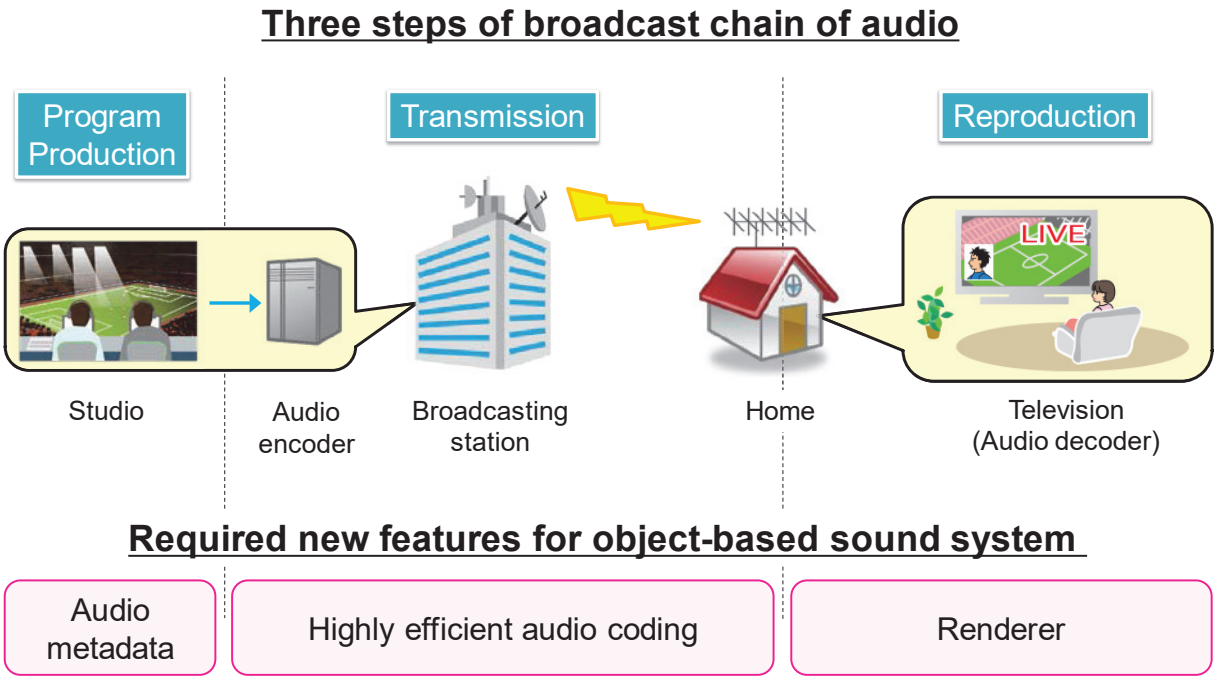

Figure 1 – Three steps of the broadcast chain of audio and required new features for an object-based sound system.

2.1 Program Production: Audio Metadata

As for program production by using an object-based sound system, audio metadata are necessary for making a program because the audio of a program is expressed by the audio data controlled by the audio metadata. Audio metadata are a data set composed of various attributes of mixing parameters, such as reproduction level, timing, and localization of audio data. Moreover, to properly manage the signal level of the program, the loudness levels of all audio data are also provided by the audio metadata [4].

The International Telecommunication Union Radiocommunication Sector (ITU-R) has specified a structure of audio metadata called Audio Definition Model (ADM) [5]. ADM is provided in the XML (Extensible Markup Language) format, which is not binary but text-based and readable. It also has a serial representation, called S-ADM, by which audio metadata is segmented into a time series of frames [6]. S-ADM is essential for broadcasting of live productions because the original ADM is static and file-based, so it is inappropriate for sequential transmission. We have adopted S-ADM as a structure for audio metadata for program production, and we are developing an S-ADM authoring system.

2.2 Transmission: Highly Efficient Audio Coding

An object-based sound system enhances flexibility of listening by using many audio data. The amount of audio data is directly linked to the quality and variety of an audio service. Thus, to transmit many audio data at low bit rate, a highly efficient audio coding scheme is required.

Accordingly, MPEG-H 3DA was adopted because it is the latest audio coding standard compliant to

the object-based sound system and is highly efficient compared with conventional audio coding technologies [7].

2.3 Reproduction: Renderer

Reproduction in the case of the object-based sound system consists of four processes:

A) Interpreting audio metadata
B) Controlling audio data with audio metadata
C) Incorporating viewer's preference
D) Rendering to home reproduction system

An audio system with the above-mentioned functions is called a renderer.

A renderer based on the MPEG-H 3DA standard was developed as a part of the audio codec. Moreover, a highly functional custom-built renderer suitable for more-advanced audio service is also being developed.

3. EXAMPLES OF PROGRAMS PRODUCED BY OBJECT-BASED SOUND SYSTEM

To verify the effect of the object-based sound system, the following programs were produced. In these productions, convenience of playback mode for viewers was focused on. As for a requirement, general viewers must not have to use an interface (IF), like a mixing console, to separately control all audio data. A simpler IF for switching playback modes would be more effective for viewers.

3.1 Music Program

A music program featuring a pop song, composed of a vocal, guitar, keyboard, bass, and drums, is introduced as an example here. The four playback modes listed below are provided. Combinations of audio data corresponding to respective playback modes are given in Table 1.

A) Recommended mode: mixing by a balance engineer
B) Vocal mode: solo vocal
C) Karaoke mode: band sound
D) Guitar mode: solo guitar

A viewer can repeatedly enjoy the same program in different playback modes.

Table 1 – Combinations of audio data corresponding to respective playback modes of a music program.

Playback mode	Vocal	Guitar	Keyboard	Bass	Drum
Recommended	✓	✓	✓	✓	✓
Vocal	✓				
Karaoke		✓	✓	✓	✓
Guitar		✓			

3.2 Sports Program

A sports program is usually composed of dialogues by commentator and the ambience of a playing field. To offer various playback modes, alternative dialogues and ambiences should be prepared and they are combined as viewers prefer. The following playback modes are for baseball programs, and combinations of each playback mode are given in Table 2.

A) Recommended mode: mixing by a balance engineer.
B) Supporter mode: the audience of a team to support and the dialogue focused on that team.
C) Immersive mode: loud ambience.
D) Multilingual mode: selected language.
E) Variety show mode: humorous chat about the game.

This program allows both serious baseball fans and neutral viewers to enjoy an identical program.

Table 2 – Combinations of audio data corresponding to respective playback modes of a sports program.

Playback mode	Dialogue						Ambience		
	Neutral	Team A	Team B	English	Russian	Variety	Neutral	Team A	Team B
Recommended	✓						✓		
Supporter (Team A)		✓						✓	
Supporter (Team B)			✓						✓
Immersive							✓ (loud)		
Multilingual					✓		✓		
Variety show						✓	✓		

4. REAL TIME AUDIO ENCODING/DECODING SYSTEM USING MPEG-H 3DA

4.1 Configuration

The configuration of the audio codec using MPEG-H 3DA—consisting of a 3DA/MPEG Media Transport (MMT) [8] encoder and a 3DA/MMT decoder—is shown in Fig. 2. Three steps of the broadcast chain of audio is illustrated again in the figure to clarify the role of each part the audio codec. The 3DA/MMT encoder receives audio data, for example 22.2 ch back ground sound and additional audio objects, and audio metadata in S-ADM format. The additional audio object corresponds to alternative audio data such as dialogues and it does not include back ground sound. S-ADM is conveyed using the last four channels of the serial multichannel audio digital interface (MADI) [9], i.e., from channels 61 to 64. Accordingly, audio data with audio metadata can be transmitted by a single MADI stream under the condition that the audio data does not exceed 60 channels.

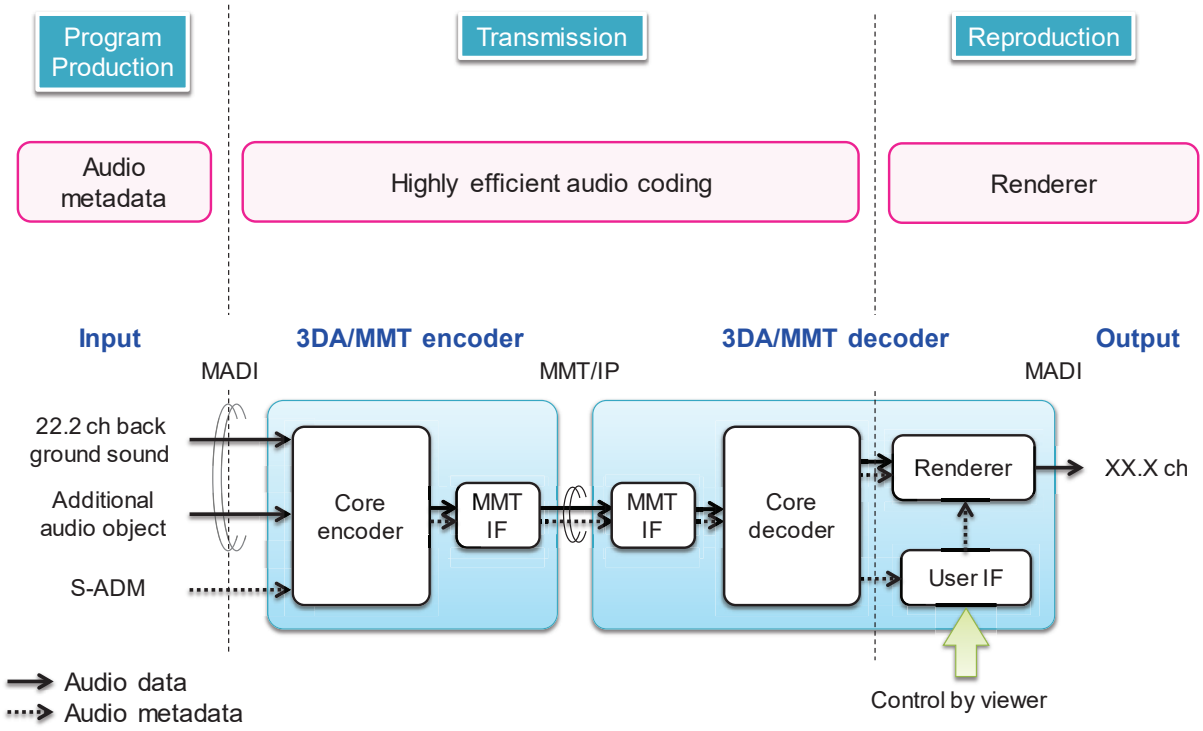

Figure 2 – Configuration of audio codec using MPEG-H 3DA

Input audio data and S-ADM are encoded by a core encoder and embedded in one audio stream of MPEG-H 3DA. The audio stream is transported on the MMT/Internet Protocol (IP) format through the MMT Interface (IF), which provides information in regard to MMT-packetized streaming. The 3DA/MMT decoder receives the audio stream and then decodes into audio data and audio metadata. Note that the decoded audio metadata is no longer in the format of S-ADM; instead, it is in the format of MPEG-H 3DA that can be interpreted by the subsequent renderer. A user IF offers available playback modes by interpreting the audio metadata. The viewer's preference is handed to the renderer through the user IF. The renderer then mixes the decoded audio data according to both the audio metadata and viewer's preference as well as the specification of home audio system.

The developed 3DA/MMT decoder is shown in Fig. 3. A Mac mini (Apple) personal computer (PC) with a 3.0 GHz Core i7 was used as a platform for the both 3DA/MMT encoder and 3DA/MMT decoder.

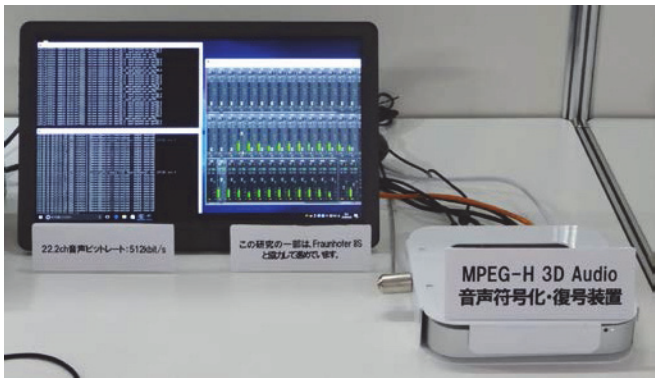

Figure 3 – 3DA/MMT decoder built on a Mac mini

4.2 Specifications

The specifications of the 3DA/MMT encoder are listed in Table 3. The 3DA/MMT encoder accepts stereo, 5.1 surround, and 22.2 ch sound as the input audio format. These audio formats were selected because they are adopted by the current 8K SHV satellite broadcasting. Bit rate of the back ground sound does not include that of the additional audio objects. The bit rate of an audio object is flexibly assigned in the range from 24 to 96 kbit/s. The actual value depends on the relation between audio objects and back ground sound; thus, the bit rate of an audio object is underspecified. That is because the core encoder adaptively judges appropriate bit distribution to all audio data to maximize the total audio quality of the program. The 3DA/MMT encoder accepts 56-ch audio data at maximum; thus, in the case of 22.2 ch back ground sound, 32 additional audio objects are available.

Table 3 – Specifications of 3DA/MMT encoder

Input audio format	Bit rate of back ground sound [kbit/s]	Number of additional audio objects
Stereo	48～192	54
5.1 surround	128～320	50
22.2 ch sound	512～1400	32

Target layouts of the 3DA/MMT decoder, i.e., audio formats that the renderer can render, are given in Table 4. It should be noted that rendering is valid in the case of converting an input original audio format to lower audio formats, e.g., from 22.2 ch sound to 7.1.4 ch. In addition, a binaural rendering for listening with headphones or earphones was implemented.

The renderer is also equipped with two other user-friendly functions: dialogue enhancement and

dialogue replacement. Dialogue enhancement is a function for improving the audibility of a dialogue by raising its level independently of the other audio data. Dialogue replacement is mainly for the multilingual service and audio description available by replacing dialogues. These two functions will play key roles in enhancing the intelligibility of and the accessibility to broadcast content in the future.

Table 4 – Target layouts of 3DA/MMT decoder

Target layout		
Mono	Stereo	5.1 surround
7.1 ch	5.1.2 ch	5.1.4 ch
7.1.4 ch	22.2 ch	Binaural

5. CONCLUSIONS

A future vision of audio services based on advanced terrestrial broadcasting technology was presented. Adopting the object-based sound system for broadcasting will allow viewers to flexibly customize an audio of programs. A real time audio codec using MPEG-H 3DA was developed to verify the proposed broadcasting scheme based on the object-based sound system. We will put the developed audio codec in a transmission experiment of terrestrial broadcasting hereafter.

ACKNOWLEDGEMENTS

The authors would like to thank the Fraunhofer Institute for Integrated Circuits for implementing the MPEG-H 3DA technology.

REFERENCES

1. “Advanced sound system for programme production,” Recommendation ITU-R BS.2051-2, International Telecommunication Union, Geneva, 2018.
2. T. Sugimoto, Y. Nakayama, T. Komori, T. Chinen, and M. Hatanaka, “Dialogue channel control for 22.2 multichannel sound broadcasting: Broadcast chain scheme and subjective evaluation of effectiveness,” J. Audio Eng. Soc. vol. 65, no. 6, pp. 507-516, 2017.
3. “Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D Audio,” ISO/IEC 23008-3:2015, 2015.
4. “Algorithms to measure audio programme loudness and true-peak audio level,” Recommendation ITU-R BS.1770-4, International Telecommunication Union, Geneva, 2015.
5. “Audio Definition Model,” Recommendation ITU-R BS.2076-1, International Telecommunication Union, Geneva, 2017.
6. “A serial representation of the Audio Definition Model,” Recommendation ITU-R BS.2125-0, International Telecommunication Union, Geneva, 2019.
7. T. Sugimoto and T. Komori, “Required bit rate of 22.2 multichannel audio signal compressed by MPEG-H 3D Audio to meet broadcast quality,” Acoust. Sci. & Tech., vol. 39, no. 3, pp. 266-269, 2018.
8. “Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 1: MPEG media transport (MMT),” ISO/IEC 23008-1:2014, 2014.
9. “AES recommended practice for digital audio engineering - serial multichannel audio digital interface (MADI),” AES10-2008, 2008.