UHD World Association
世界超高清视频产业联盟

UHD World Association
世界超高清视频产业联盟

UHD Word Association
www.theuwa.com

Audio Vivid

Technical White Paper

(V1.0)

UHD World Association

About This Document

The Secretariat of the UHD World Association (UWA) oversaw efforts to formulate this document and is responsible for its interpretation.

This document was published on August 29, 2022.

This document is proposed and managed by the UWA.

This document is jointly drafted by: China Media Group, China Electronic Technology Standardization Research Institute, Huawei Technologies Co., Ltd., Beijing DouYin Information Service Co., Ltd., CCTV New Media Culture Media (Beijing) Co., Ltd., Beijing Sinecore MicroSemi Technology Co., Ltd., China Television Information Technology (Beijing) Co., Ltd., Academy of Broadcasting Science of the National Radio and Television Administration, Academy of Broadcasting Planning of the National Radio and Television Administration, Migu Co., Ltd. of China Mobile, Smart Home Operation Center of China Mobile, Fifth Electronics Research Institute of Ministry of Industry and Information Technology, Shenzhen Tencent Computer System Co., Ltd., Beijing iQIYI Science & Technology Co., Ltd., Sharp Corporation, China United Network Communications Group Co., Ltd., TCL Industrial Holdings Co., Ltd., Konka Group Co., Ltd., Hisense Visual Technology Co., Ltd., Shenzhen Skyworth-RGB Electronics Co., Ltd., Qingdao Haier Multimedia Co., Ltd., Sichuan Changhong Electric Co., Ltd., BOE Technology Group Co., Ltd., Guangzhou Shiyuan Electronic Technology Co., Ltd., Unilumin Group Co., Ltd., Hangzhou Arcvideo Technology Co., Ltd., Sumavision Technologies Co., Ltd., Beijing Bohui Science & Technology Co., Ltd., Youku Information Technology (Beijing) Co., Ltd., Beijing Baidu Netcom Science and Technology Co., Ltd. Beijing Xiaomi Electronic Products Co., Ltd., Hunan Broadcasting System, Guangdong Television, Beijing Golden Times Media Technology Corp., Ltd., Shanghai National Engineering Research Center of Digital Television Co., Ltd. Zhejiang Radio and Television Group, Zhejiang Huace Film & TV Co., Ltd., Zhongtu Yunchuang Intelligent Technology (Beijing) Co., Ltd., Beijing Sports University 5G High-Tech Video Sports Integration Innovation Application Laboratory of State Administration of Radio and Television, and Shanghai Media Group.

The main authors are: Jiang Wenbo, Cheng Duofu, Guan Chaoyang, Li Yan, Chen Renwei, Li Jingxin, Wang Zhe, Huang Chuanzeng, Guo Xiao, Wu Jian, Xu Jianxin, Sun Jian, Ning Jinhui, Zhou Yun, Jiao Jianbo, Lu Nannan, Pang Chao, Zheng Qiang, Fan Shengli, Zhang Jiandong, Liu Derong, Wu Qiang, Wang Qi, Pan Xinghao, Xie Yugui, Li Kangjing, Cheng Zhipeng, Yang Zhongyao, Shan Huaqi, Xing Gang, Han Jian, Guo Peipei, Chen Wei, Cheng Jian, Wang Peng, Lin, Zhang Hongwei, Du Zhengzhong, Sun Lei, Zhu Zirong, Li Dalong, Liu Changtao, Zhao Xinglong, Su Yunquan, Wang Haiying, Xu Yaoling, Tong Hai, Geng Yidan, Chen Xun, Dun Shengbao, Pan Li, Chen Chundan, Li Guangzhou, Xiong Zhaomin, Lai Fan, Wang Ziqian, Wu Xiaodong, Fu Binxing, Li Yunfeng, Wei Shengyu, Tan Shenglin, Chen Yong, Dong Jie, Zhou Cheng, Zou Jianyu, Xie Yang, Zeng Zejun, Yin Huiqing, Wang Xuehui, Xu Yangfa, Chen Jiaxing, Chen Zuole, Chen Jiaxin, Jiang Jianliang, Wang Rongfang, Li Fa, Xing Huaifei, Zha Li, Yu Lei, Gao Weibiao, Qin Yu, and Tang Xun

Disclaimer:

1. This document is available for use free of charge. It is used for reference only and is not liable for any products developed by referencing this document.
2. Updates to this document will be uploaded to the official UWA website without notice.

Table of Contents

1. Background

Audio technologies have steadily developed as people continue to pursue true-to-life audio, and as science and technology continue to advance. From the emergence of the first gramophone to optical disks, magnetic recording, and electronic recording, and from mono and stereo audio to surround sound and 3D audio; audio processing technologies have evolved from purely electronic technologies to a combination of electronic technologies and the latest processing technologies, such as AI. These technologies continue to evolve.

In recent years, evolution to ultra high definition (UHD) has allowed people to enjoy immersive, true-to-life audio and video experiences. UHD offers high resolution, high frame rate, high color depth, wide color gamut, high dynamic range, and 3D audio. As a key element of the UHD experience, 3D audio can create a true sense of space and direction. This makes it highly immersive and true-to-life, bringing audiences a more impressive, personalized, and interactive experience.

From the industry development perspective, the coordinated development of the audio and video industries has started accelerating. For example, standard-definition televisions with stereo audio made the leap to high-definition televisions with surround sound, and we are now seeing UHD televisions with 3D immersive audio. In addition, large outdoor screens support 3D audio. These developments have greatly promoted overall industry development. Now, the priority is to seize development opportunities by improving audio performance, quickly catch up with advanced international technologies, and meet people's requirements for high-quality audio content, thereby promoting the development of the audio and video industry. The promotion of 3D audio will have significant economic and social implications.

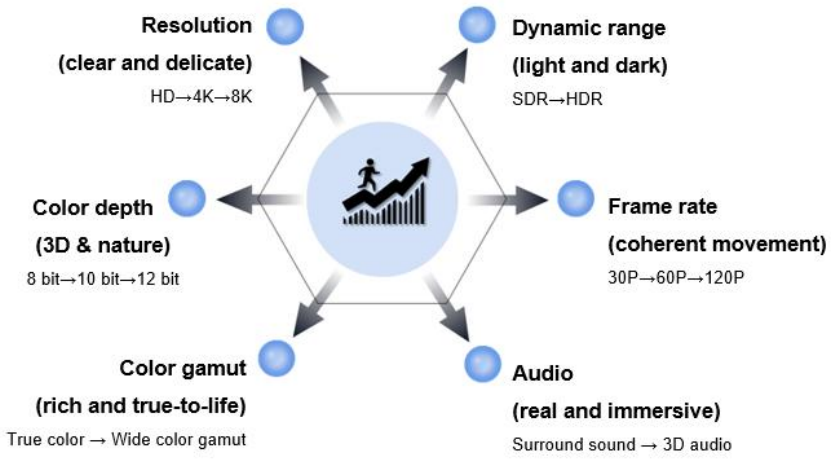

Figure 1 Audio: One of Six Key Elements of UHD Videos

2. Technology Introduction

2.1 Application Scenarios

Compared to traditional audio, 3D audio creates a greatly enhanced sense of space and direction, allowing the accurate reproduction of sounds that people hear in real life. This meets requirements for highly-immersive, true-to-life audio, while offering a more personalized and interactive experience. Audio Vivid supports the entire process, from sound production to reproduction, and can be applied to home, cinema, personal, AR/VR, and vehicle audio scenarios.

2.2 Sound Reproduction Technologies

There are three technical methods for reproducing a sense of space and direction when dealing with sounds. The first is based on channels; the second is based on beds and objects; the third is based on sound fields.

1. **Channel-based method**

This method requires the configuration of audio channels in order to map the audio source of each channel to a specified speaker. The limitation of this method is that each channel configuration corresponds to only one speaker, so audio being rendered in one channel configuration will only have the optimal effect only with the corresponding speaker.

2. **Bed- and object-based method**

The bed carries the basic ambient sound, while the object is a collection of mono audio elements and corresponding metadata. Metadata contains the location, strength, and size of the object. During playback, an object is mapped to one or more speakers, or to headphones, through binaural rendering based on the metadata in order to achieve the desired spatial audio effect.

3. **Ambisonics**

Sound is a vibration that propagates as an acoustic wave. At a given time, each point of a sound field can be represented by a pressure function of multiple waveform fields. When the pressure value of a point in the space is obtained, the sound in that space can be reproduced. To ensure and improve sound field-based audio quality, each coefficient of the pressure function of the point must be accurately obtained, so as to improve the coding quality of the sound field spatial coefficient. Higher Order Ambisonics (HOA) is a format for modeling 3D sound fields that are defined on the surface of a sphere. It can accurately capture, process, and reproduce sound fields on any device, such as headphones,

loudspeakers, and loudspeaker boxes. The performance of the HOA system increases as the HOA order increases. This means an increase in the number of HOA signals, in turn requires more powerful technologies to better process signals. The introduction of AI and other new technologies allows this requirement to be met.

4. **3D audio reproduction with speakers**

A 3D audio reproduction environment plays an important role in 3D audio production. The number of speakers used greatly affects the scope of the area in which it is possible to achieve optimal audio effects. The fewer the speakers, the smaller the area.

5. **3D audio reproduction with binaural rendering**

3D audio reproduction can also be achieved with headphones. Producing the very best 3D audio through common stereo headphones requires the binaural rendering of the aforementioned three methods. Binaural rendering technology can better promote 3D audio, and is also the main rendering technology for 3D audio in VR/AR scenarios.

3. Audio Vivid Solution

3.1 Introduction to Audio Vivid

The UHD World Association (UWA) worked with players from across the industry ecosystem to promote the release of the draft Audio Vivid technical group standards, which is intended to quickly promote the development of the UHD video industry and enhance the influence of related core and key technology standards.

Compared with other 3D audio technologies, the Audio Vivid standard is a more technologically-advanced, more open technical standard for players worldwide. The standard meets industry security requirements and is backed by a friendly industry ecosystem policy. This makes it more suitable for implementation across the UHD industry ecosystem. Through the joint efforts of all parties, the Audio Vivid technical standard has been developed and is set to be applied at scale in the market.

3.2 Solution

Audio Vivid involves tools, chips, devices (e.g., TVs and mobile phones), platforms, and other elements. The following figure displays the integration.

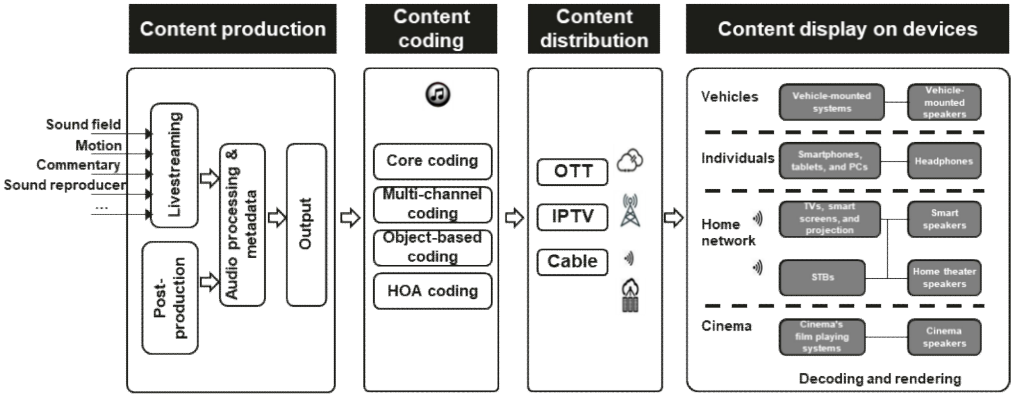

Figure 2 Audio Vivid Integration

The Audio Vivid standard covers the codecs and rendering related to content generation and reproduction. Currently, the standard covers content production, encoding, and decoding as well as rendering on devices.

3.3 Technical Principles

Based on Audio Vivid, different technical tools are used to encode and decode different types of input signals. Multi-channel coding technology is used to eliminate information redundancy in multi-channel signals. HOA spatial coding technology is used to eliminate spatial geometric information redundancy between HOA channels. Psychoacoustic model-based preprocessing and AI-based quantization and entropy coding technologies are used to eliminate information redundancy in mono and object-based audio signals. Speakers or headphones are used to complete final rendering and output. For further details, refer to *3D Audio Technology Specification: Part 1: Coding, Distribution and Presentation* (Audio Vivid) released by the UWA.

3.4 Key Features

As one of the world's leading UHD audio and video technologies, Audio Vivid has the following key features:

1. Supports mono, stereo, surround sound, and 3D audio (multi-channel bed, object, and ambisonic sound field)
2. Supports lossy and lossless codecs
3. Supports speakers and binaural rendering
4. Supports the representation of Three Degrees of Freedom (3DoF)
5. Supports the HOA spatial coding tool, which greatly improves the coding efficiency of HOA signals

6. Features decoder complexity similar to that of industry benchmarks
7. Supports 16-channel codecs (supporting up to the third order of HOA, with the possibility of extending to higher orders)
8. Supports a sampling rate range of 32 kHz to 192 kHz
9. Supports bit depths of 16-bit and 24-bit
10. Supports a bitrate range of 32 Kbps to 1.6 Mbps
11. Offers codecs algorithm latency of less than 80 ms (typical scenario)
12. Supports up to 7th-order HOA binaural rendering and real-time rendering of 128-track audio
13. Offers binaural rendering latency of less than 40 ms (typical scenario)

3.5 Codec Solution

The Audio Vivid audio coding system supports channel signal coding, object signal coding, HOA signal coding, and metadata coding. [1]

Coders are composed of multiple coding tools, as shown in Figure 3, including a general full-bitrate audio coding tool and a lossless audio coding tool.

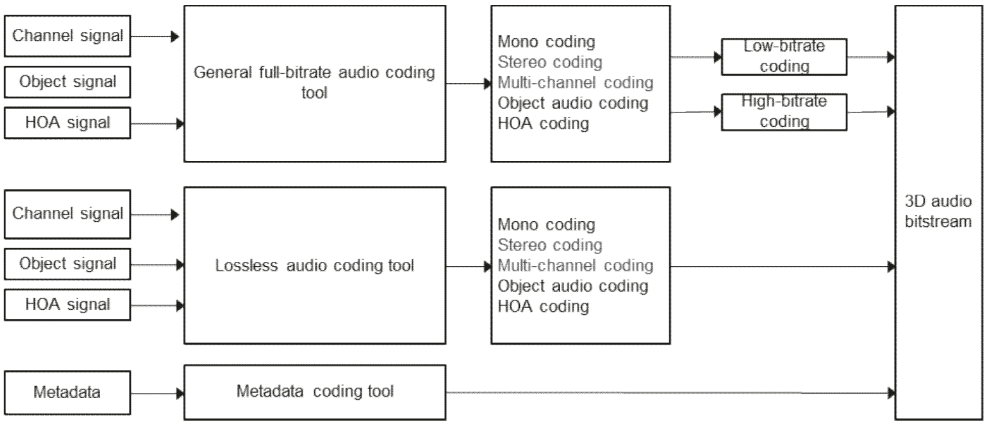

Figure 3 3D Audio Encoder

When dealing with audio signals with different characteristics or application scenarios, users can use the general full-bitrate audio coding tool, lossless audio coding tool, or metadata coding tool based on the input type and bitrate range.

1. **General full-bitrate audio coding tool**

The tool adopts neural network transformation, quantization, and entropy coding technology, multiple-channel downmixing and bit allocation technology based on inter-channel correlation, HOA spatial

coding technology which is built upon virtual speakers, etc. This tool is applicable to mono, stereo, multi-channel coding, object audio coding, mixed audio coding, and HOA coding.

The sampling rate ranges from 32 kHz to 192 kHz, and bit depths of 16-bit and 24-bit are supported. The coding output bitstreams are as follows:

(1) Mono/object audio: 32 Kbps, 44 Kbps, 56 Kbps, 64 Kbps, 72 Kbps, 80 Kbps, 96 Kbps, 128 Kbps, 144 Kbps, 164 Kbps, and 192 Kbps;

(2) Stereo sound: 32 Kbps, 48 Kbps, 64 Kbps, 80 Kbps, 96 Kbps, 128 Kbps, 144 Kbps, 192 Kbps, 256 Kbps, and 320 Kbps;

(3) 5.1 multi-channel audio: 96 Kbps, 128 Kbps, 144 Kbps, 160 Kbps, 192 Kbps, 256 Kbps, 320 Kbps, 384 Kbps, 448 Kbps, 512 Kbps, 640 Kbps, and 720 Kbps;

(4) 7.1 multi-channel audio: 128 Kbps, 160 Kbps, 192 Kbps, 256 Kbps, 384 Kbps, 480 Kbps, 576 Kbps, and 640 Kbps;

(5) 5.1.4 3D audio: 176 Kbps, 256 Kbps, 384 Kbps, 448 Kbps, 576 Kbps, and 704 Kbps;

(6) 7.1.4 3D audio: 240 Kbps, 384 Kbps, 512 Kbps, 608 Kbps, and 832 Kbps;

(7) FOA: 96 Kbps, 128 Kbps, 192 Kbps, and 256 Kbps;

(8) 2nd-order HOA: 192 Kbps, 256 Kbps, 320 Kbps, 384 Kbps, 480 Kbps, 512 Kbps, and 640 Kbps;

(9) 3rd-order HOA: 256 Kbps, 320 Kbps, 384 Kbps, 512 Kbps, 640 Kbps, and 896 Kbps;

(10) Bed + object hybrid signals: various combinations of bitrates of the above stereo, multi-channel, and 3D audio signals, and mono and object signals;

To ensure the audio quality required for the commercialized systems that adopt Audio Vivid, the following coding bitrates are recommended for typical audio signal configurations:

(1) 5.1 configuration: ≥ 192 Kbps; ≥ 320 Kbps for applications requiring ultra-high voice quality

(2) 5.1.2 configuration: ≥ 320 Kbps; ≥ 480 Kbps for applications requiring ultra-high voice quality

(3) 5.1.4 configuration: ≥ 384 Kbps; ≥ 576 Kbps for applications requiring ultra-high voice quality

(4) 7.1.4 configuration: ≥ 384 Kbps; ≥ 608 Kbps for applications requiring ultra-high voice quality

(5) Each object: ≥ 96 Kbps; ≥ 164 Kbps for applications requiring ultra-high voice quality

2. **Lossless audio coding tool**

The lossless audio coding tool supports up to 128 channels at any sampling frequency. This tool also supports bit depths of 8-bit, 16-bit, and 24-bit.

3. **Metadata coding tool**

The tool is used for the coding of audio metadata.

4. Decoding and rendering system

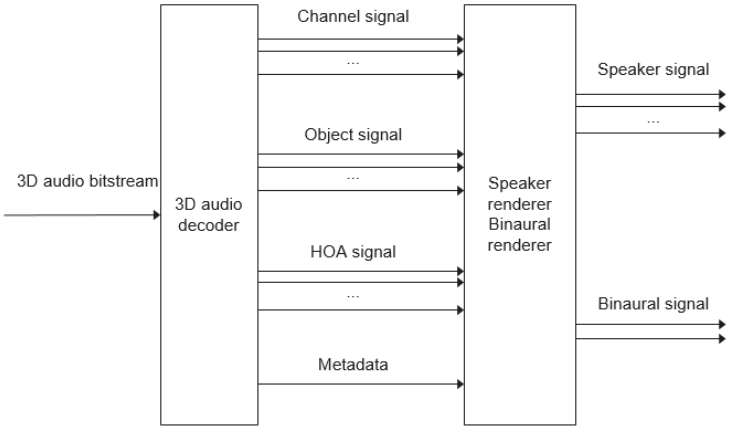

Figure 4 3D Audio Decoding and Rendering System Framework

5. Example of general full-bitrate audio codec

General full-bitrate audio codec covers mono signal coding, stereo signal coding, multi-channel signal coding, object-based signal coding, mixed signal coding, HOA signal coding, and metadata coding.

5.1 Encoder architecture

The basic architecture of a general full-bitrate audio encoder can be seen in Figure 5. The encoder consists of core functions, such as transient detection, window type decision, time-frequency transformation, frequency-domain noise shaping, time-domain noise shaping, bandwidth extension coding, downmixing, neural network transformation, quantization, and range coding. The encoder encodes channel signals and object signals into bitstreams; the HOA spatial encoder and core encoder encode HOA signals into bitstreams; and the metadata encoder encodes metadata into bitstreams.

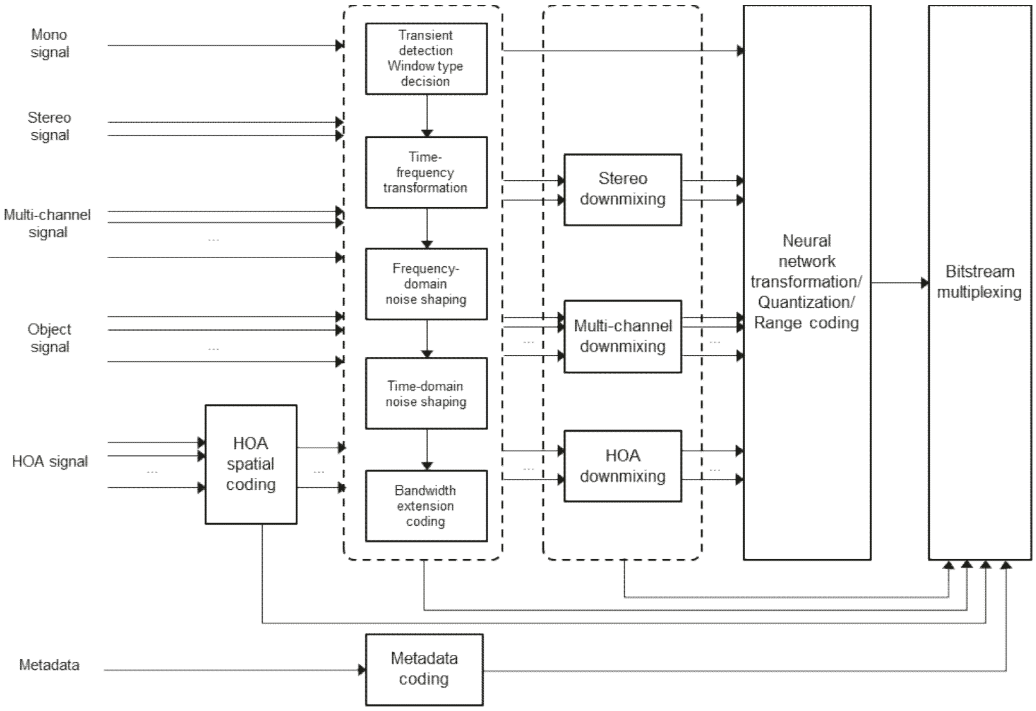

Figure 5 General Full-Bitrate Audio Encoder Framework

5.2 Decoder architecture

The basic architecture of a general full-bitrate audio decoder can be seen in Figure 6. The decoder consists of core functions, such as range decoding, inverse quantization, inverse neural network transformation, bandwidth extension decoding, inverse time-domain noise shaping, inverse frequency-domain noise shaping, upmixing, and inverse time-frequency transformation. The decoder decodes bitstreams into channel signals and object signals; the HOA spatial decoder and core decoder decode bitstreams into HOA signals; and the metadata decoder decodes bitstreams into metadata.

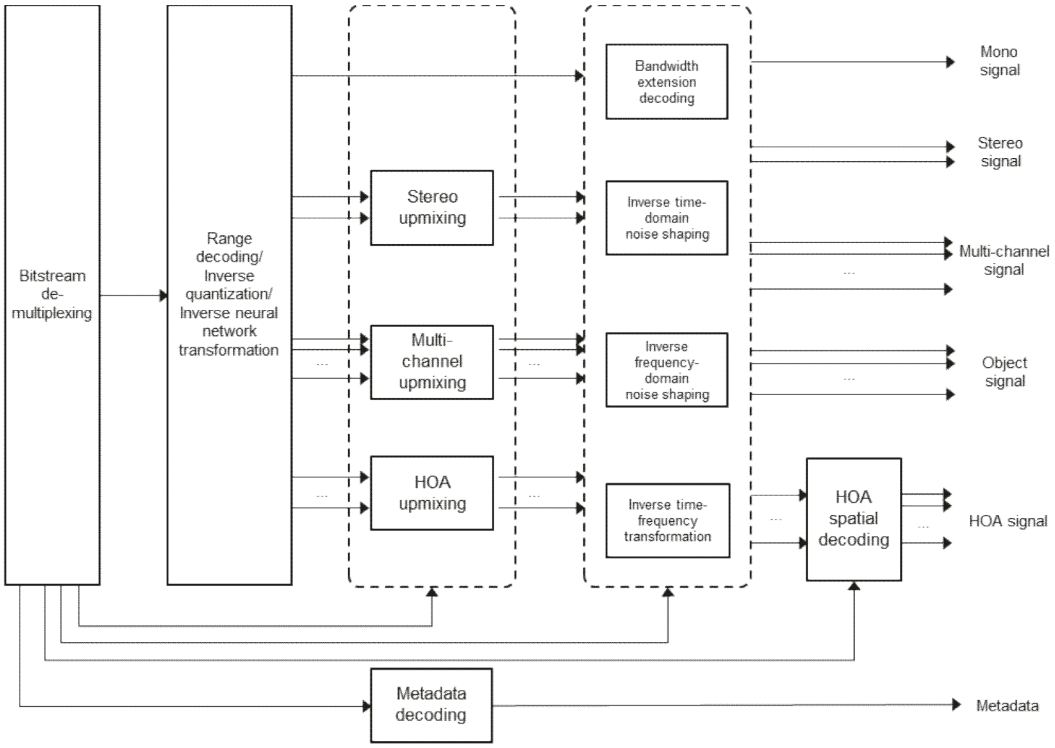

Figure 6 General Full-Bitrate Audio Decoder Framework

5.3 Workflow

A general full-bitrate audio encoder handles audio signal pre-processing, downmixing of signals in each mode, neural network transformation, quantization, range coding, and metadata coding. During the audio signal pre-processing, the encoder transforms time-domain signals into frequency-domain signals and then performs pre-processing for the signals. After the pre-processing, frequency-domain signals are subject to downmixing based on different coding modes to remove inter-channel correlation. After the downmixing, the signal spectrums are subject to transformation and coding with neural networks in the phase of neural network transformation, quantization, and range coding. Metadata coding performs quantization and coding on metadata based on their structure and specific quantization specifications. A general full-bitrate audio decoder handles audio signal post-processing, upmixing of signals in each mode, inverse neural network transformation, inverse quantization, range decoding, and metadata decoding. Decoding is the reverse of encoding.

3.6 Metadata Solution

3D audio metadata is compatible with the audio models defined in ITU-R BS.2076, while extended metadata is newly defined. The metadata system consists of two parts: basic metadata (<AudioformatExtended>) and extended metadata (<VRext>). The basic metadata part references the

ITU-R BS.2076-2 standard, while the extended metadata part is newly defined in this document. Based on this architecture, the metadata system described in this document is forward-compatible and backward-expandable. In addition to meeting global interconnection requirements regarding metadata, this metadata system is flexible and scalable, and can provide a powerful representation capability for the immersive audio system outlined within this document. For further details about the metadata system architecture, please see Figure 7.

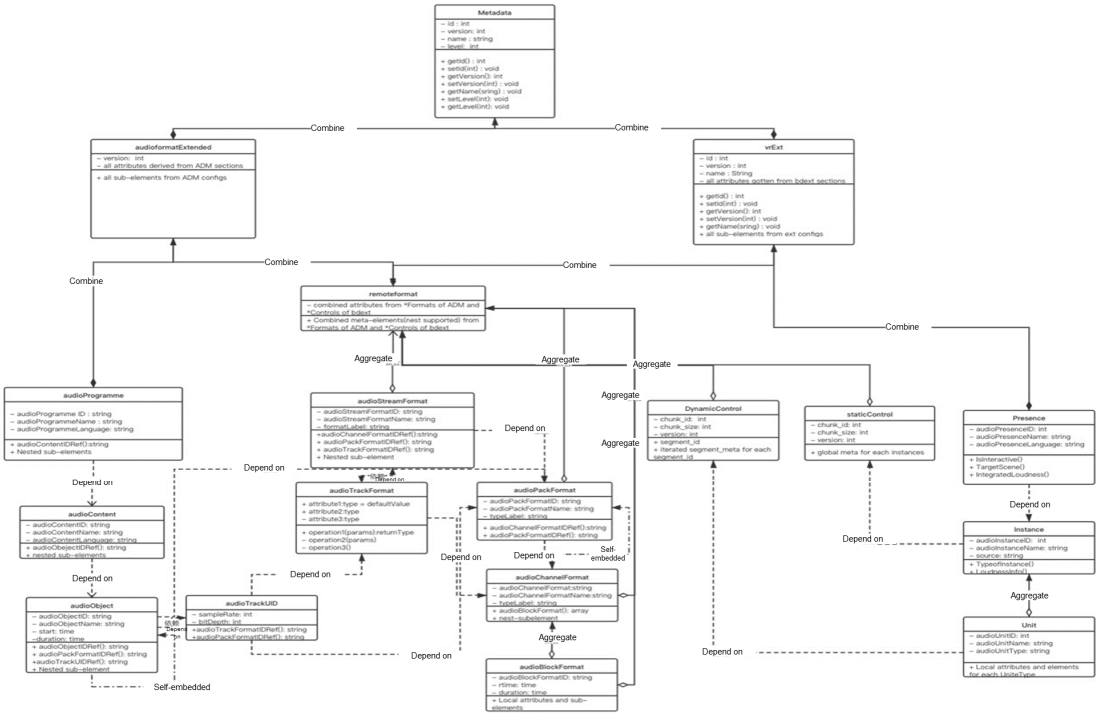

Figure 7 Metadata System Architecture

For details about the codecs of Audio Vivid and other examples, see the released coding specifications in the appendix.

3.7 Rendering Solution

3.7.1 Speaker rendering

The Audio Vivid renderer can be employed for both audio content production and audio rendering and playback on devices. The renderer can be used to directly parse ADM metadata, as defined in ITU-R BS.2076, or parse third-party metadata. Providing an open reference renderer for ADM metadata parsing during audio evaluation, production, and monitoring will be beneficial to the health of the audio content ecosystem as a whole.

We should adhere to the following rendering principles when using and processing metadata and related audio data:

(1) Properly process loudness-related metadata to maintain target loudness.
(2) Preserve the artistic intent, through actions such as the correct usage of upmixing and downmixing matrices for transmission.
(3) Properly present the spatial attributes (such as the location and spatial range) of objects.
(4) Properly process diffuse reflection and divergence azimuthRange, and zone exclusion - metadata.
(5) Properly process high-priority audio objects.

The overall architecture of a speaker renderer consists of several core components and processing steps, as shown in the above figure. The specific functions are as follows:

(1) It transforms ADM metadata to renderable TypeMetadata.
(2) It calculates gains based on the TypeMetadata used for rendering.

Based on type definition of type metadata, rendering can be classified into the following types:

- ✓ Object-based rendering
- ✓ Direct speaker signal rendering
- ✓ HOA-based rendering

The matrix type is processed during the creation of rendering items and is part of the direct speaker type renderer; and the binaural type also serves as a direct speaker.

1. **Object-based rendering**

The structure of object-based rendering is as follows:

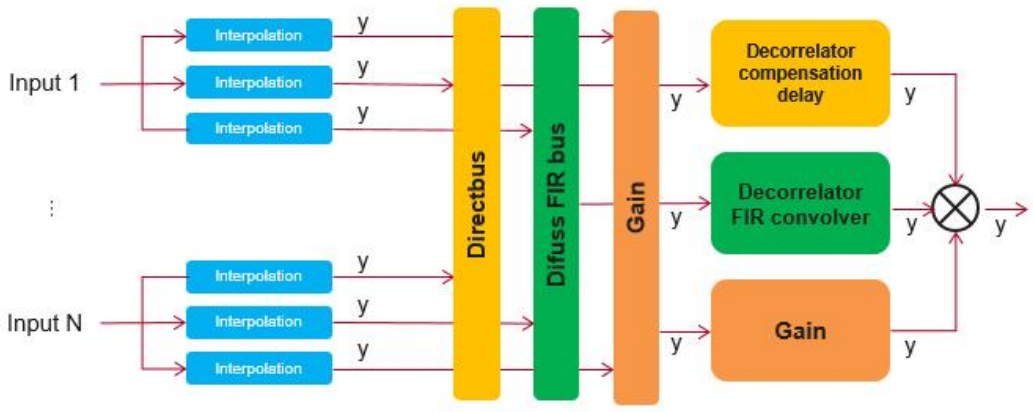

Figure 8 Object-Based Rendering

In cases when the output of the object audio data and metadata is requested through the object output interface, the metadata and audio data inputs can be provided, without the processing of diffuse reflection. Metadata is input into the renderer in the form of objectsTypemetadata. The object-based metadata and audio data is output after the preprocessing of the element metadata and calculation of "objects_gains". Those data will be enabled for playback when the interface is requested to output data.

2. **Channel-based rendering**

The structure for rendering based on the DirectSpeakers channel is as follows:

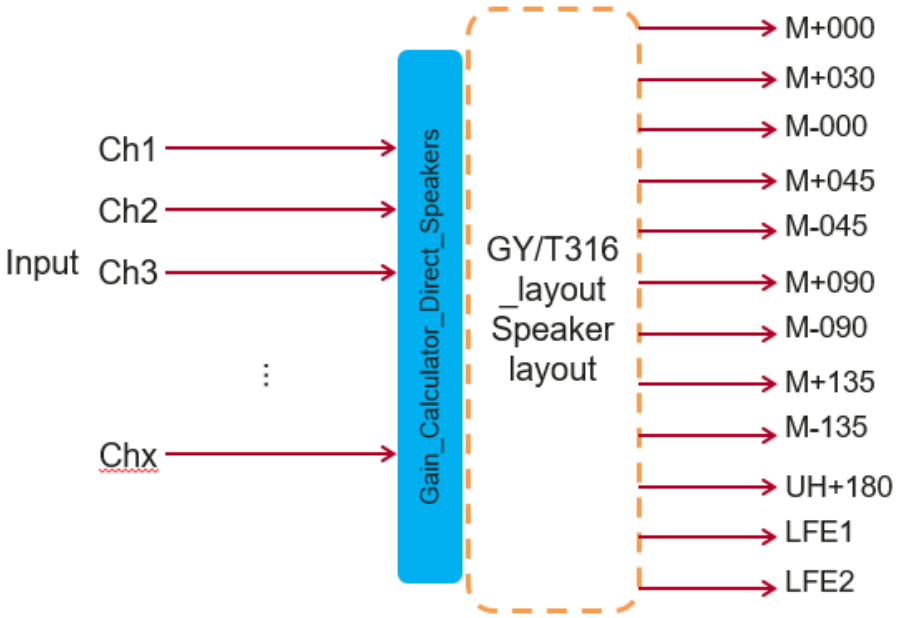

Figure 9 Channel-Based Rendering

The gains that are calculated with Gain_Calculator_Direct_Speakers for DirectSpeakers are directly applied to the audio input channels to generate audio output channels. The metadata of DirectSpeakers should not be dynamic. Therefore, interpolation should not be implemented for the gains in audioBlockFormat. However, if the metadata is modified by users, interpolation should be implemented for the gains.

During gain calculation, the number of input channels is independent of the number of speaker layout channels for output playback. Regardless of whether the number of input channels equals the number of output channels, the renderer performs rendering based on the number of output channels defined by BS.2051. When there are fewer DirectSpeakers input channels than output channels, the gains for upmixing are calculated. When there are more DirectSpeakers input channels than output channels, the

gains for downmixing are calculated. Since the basic signals of matrix-based audio types refer to those in the middle and two side signal channels, upmixing is generally performed on the side signals.

3. **HOA-based rendering**

The structure for HOA-based rendering is as follows:

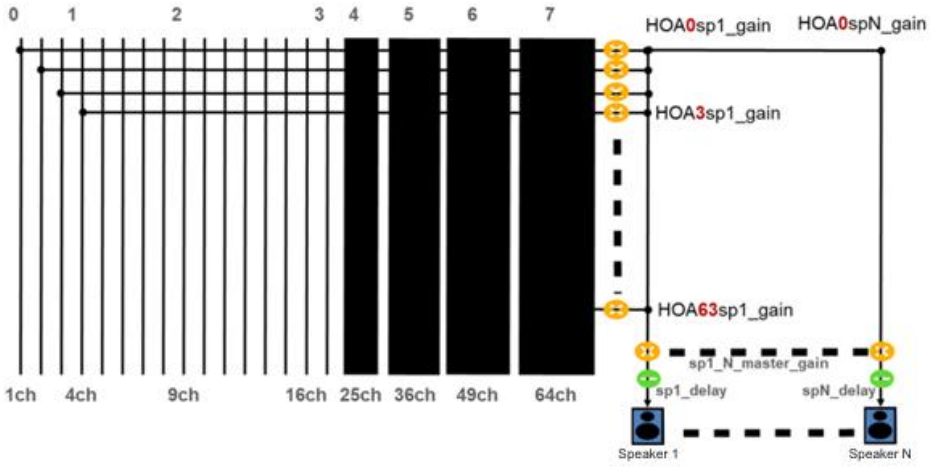

Figure 10 HOA-Based Rendering

The renderer selects the HOA decoding method based on the AllRAD algorithm. The HOA decoder D calculates the HOAnspN_gain value that is provided for each loudspeaker box by each HOA track, and then delivers multi-track HOA signals to independent loudspeaker boxes. The decoding matrix D, which is calculated based on Gain_calculator_HOA for HOA channels, is directly used in input audio channels to produce output audio channels. Like DirectSpeakers, HOA metadata should not be dynamic (there should be an audioBlockFormat for each audioChannelFormat). Therefore, interpolation should not be implemented for the gains in the block. However, if the metadata is modified by users, interpolation should be implemented for the gains.

The spN_delay function for signal receiving should be adjusted before signals are displayed on the interface, to make sure that the signals have the same delay as regular output signals. The HOA interface and the HOA renderer do not run simultaneously. If the signals need to be outputted through the HOA interface, the signals that are outputted through the HOA render should be muted without producing audio, and vice versa.

3.7.2 Binaural rendering

The audio rendering system is controlled by information that describes the audio content and rendering technology of the metadata system found in Section 3.7 of this document. This information contains

whether the audio payload is inputted in the form of mono-, stereo-, or multi-channel audio, object-based audio, HOA audio, or information about dynamic sound sources and listener locations, and rendered acoustic environment information, such as the shape, size, and wall materials of a house. The core rendering system renders playback devices and environments based on different representations of audio signals and corresponding metadata parsed from the metadata system. The following figure shows the framework of the rendering system.

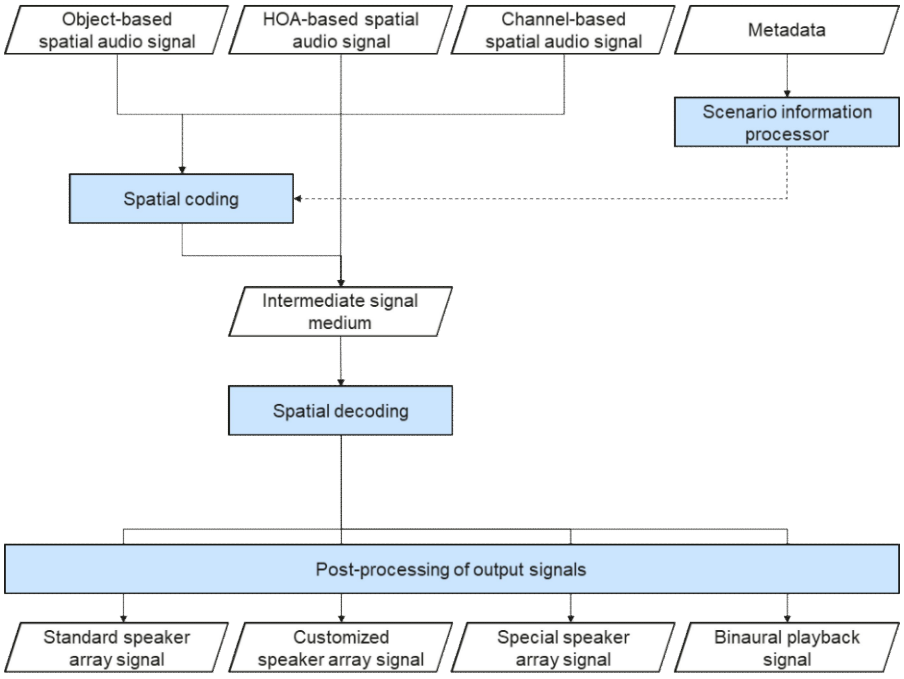

Figure 11 Overall Framework of the Rendering System

The rendering engine will process audio data in blocks (for example, a block of 1,024 sampling points). The size of a block is determined during engine initialization and does not change afterwards. Metadata should be parsed and scenario information should be adjusted based on the metadata in real-time processing. The rendering of multi-channel audio, object-based audio, and HOA audio is detailed below.

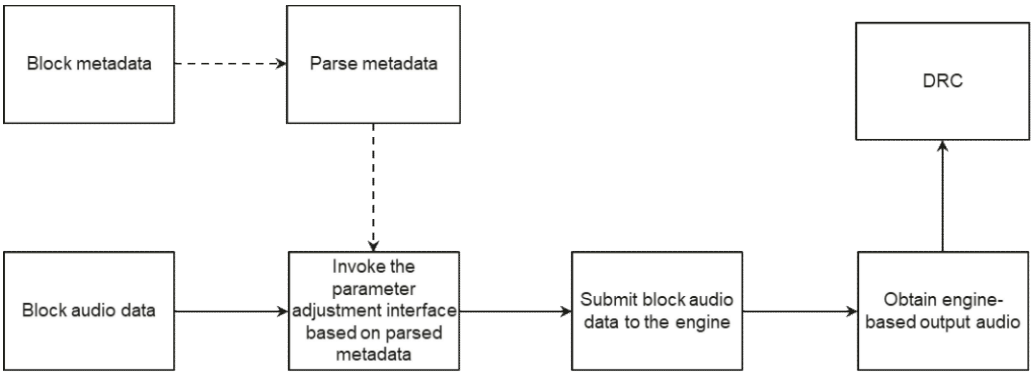

Figure 12 The Processing Link

1. **HOA signal rendering**

Currently, renderers can render HOA signals up to the seventh order. Since the channels of HOA signals in ADM are separated and marked with corresponding orders and degrees, the engine only supports channel ordering in Ambisonics Channel Number (ACN) format, and cannot support channel ordering in Furse-Malham (FuMa) higher-order format. Therefore, orders and degrees must be converted accordingly before the input of channels in FuMa format. In terms of regularization regarding ACN, the engine supports two regularization modes: N3D and SN3D. For regularization related to FuMa, conversion must be completed in advance.

Renderers can simultaneously render multiple HOA signals. During block processing, if signals of multiple channels have the same order and degree, the signals will be mixed.

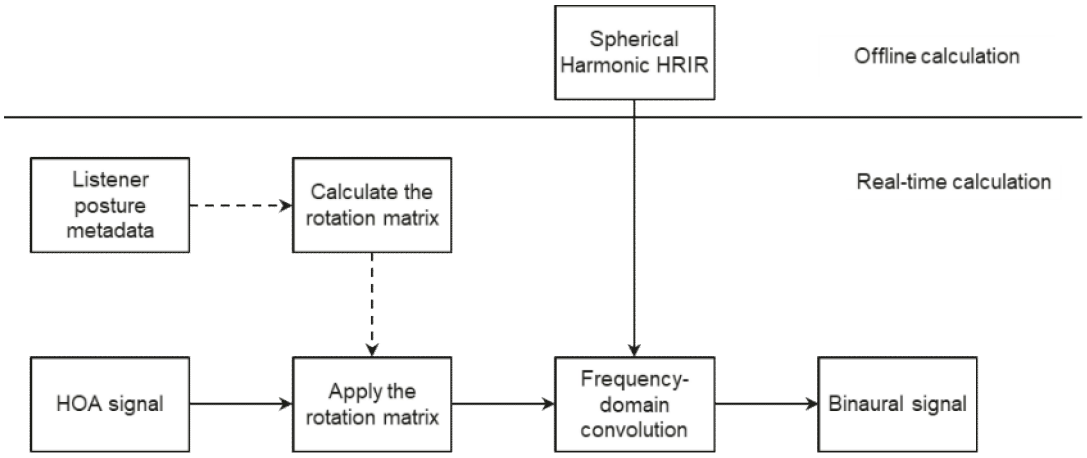

Figure 13 HOA Signal Rendering

The above figure summarizes the binaural rendering of HOA signals. The renderer has a set of head-related impulse responses (HRIRs) from the first to the seventh order that are calculated based on spherical harmonics. During rendering, the renderer first determines the rotation matrix based on the posture of a listener, and then converts HOA signals and performs HRIR convolution in each HOA channel.

2. **Object signal rendering**

The renderer encodes object signals into HOA signals, and then reuses the HOA signal rendering process to perform binaural rendering of object-based audio.

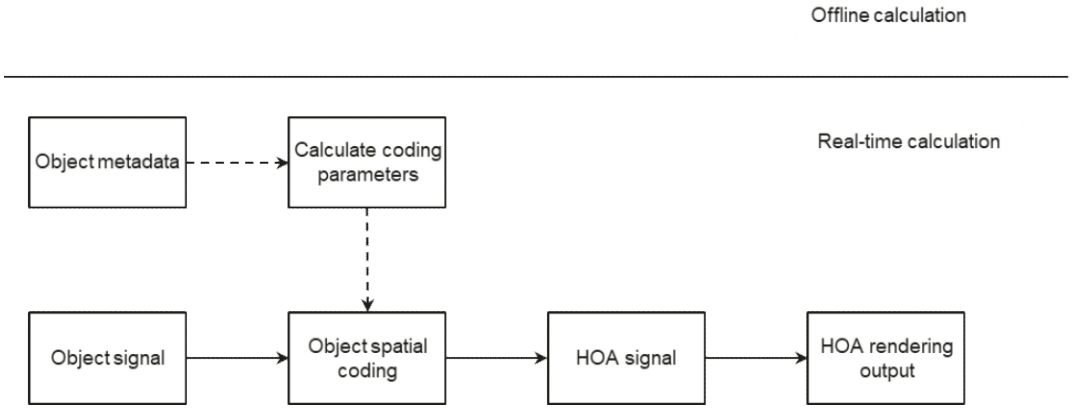

Figure 14 Object Signal Rendering

3. **Multi-channel signal rendering**

We can refer to the method of object-based audio rendering when rendering multi-channel audio into binaural audio. This means that we should regard multi-channel audio as multiple pieces of object-based audio, and then perform HOA signal rendering of the audio for binaural playback.

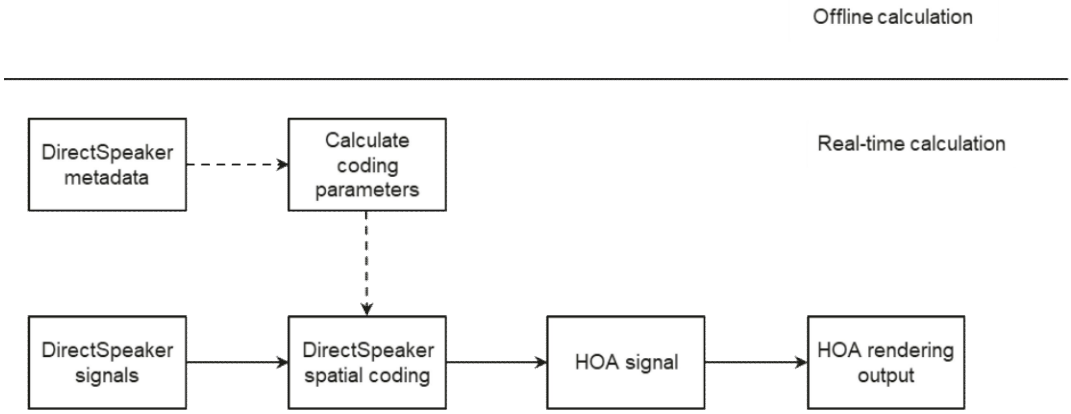

Figure 15 Multi-Channel Signal Rendering

4. Audio Vivid User Experience

4.1 Example of Speaker Layout

Take the 5.1.4 speaker setup as an example. This setup is the basic configuration for home audio and video services. Traditional 5.1 surround sound system consists of a left speaker, a right speaker, a central speaker, a left back speaker, a right back speaker, and a subwoofer. The 5.1.4 setup is based

on the 5.1 setup, and then adds four overhead speakers (left top front, right top front, left top back, and right top back). Figure 16 shows the layout of speakers in the 5.1.4 setup.

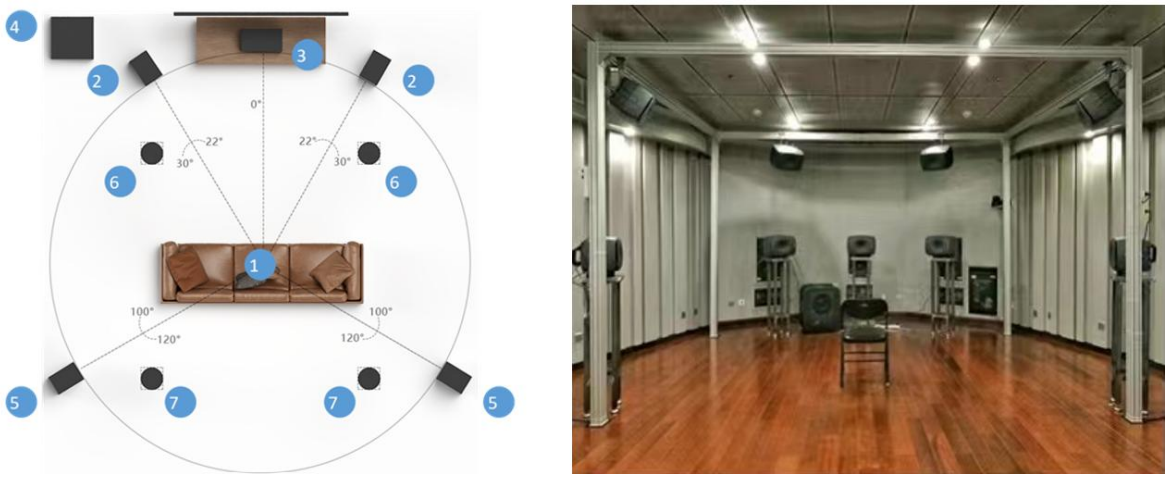

Figure 16 Speaker Layout

Playing audio and video in a well-arranged room that has an appropriate device layout provides the audience with an immersive, 3D audio experience that offers a sense of space.

1. Best position for optimal sound effect
2. Left and right speakers
3. Central speaker
4. Subwoofer
5. Left and right back speakers
6. Left and right top front speakers
7. Left and right top back speakers

4.2 Example of Binaural Spatial Audio

Since sound may change its transmission route as it gets closer to auricles or the head before entering a person's ears, it may reach the left and right ears at slightly different times. This can cause the sound that is heard by a person to be different from the original sound, as demonstrated in the figure below.

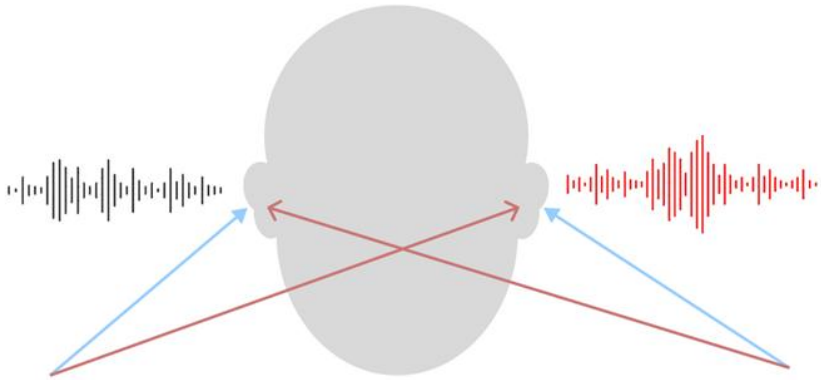

Figure 17 Binaural Rendering

Through binaural rendering, individuals can hear the same sound while wearing a headphone as they would without and obtain a fully immersive 3D audio experience that offers a sense of space.

4.3 Subjective Experience and Verification

3D audio experiences are generated based on the speaker solution, as seen below.

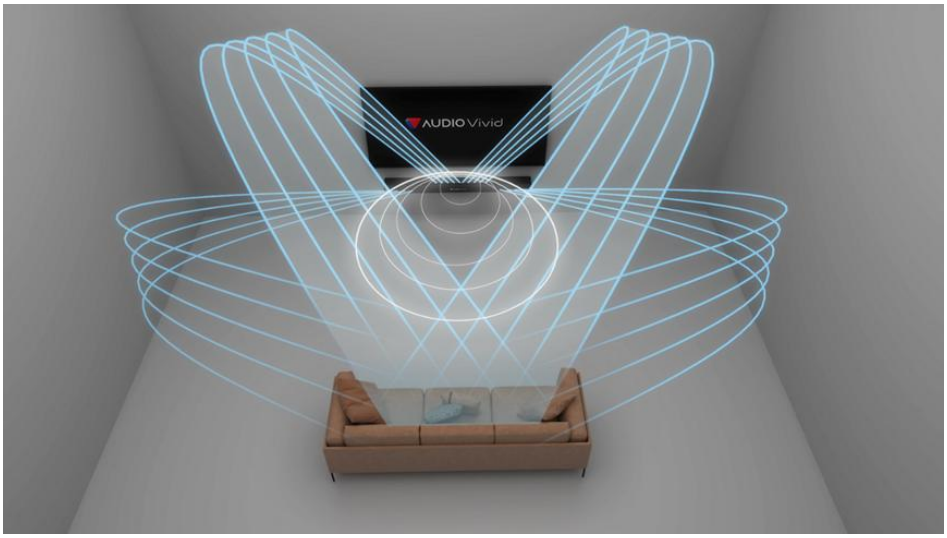

Figure 18 General 3D Sound Experiences

Individuals can perceive the direction, ambient effects, and height of the sound and evaluate it according to the following dimensions.

Function	**Evaluation Dimension**
Music quality	Clarity
	Balance
	Brightness
	Distortion
	Pleasantness (fullness, richness, and softness)
	Dynamics
Space	Reverberation perception
	Spatial perception
Direction	Verticality
	Horizontality
Distance	Loudness
	Sound pressure level
Volume of sound source	Sound width and depth
Sound externalization	(For headphones)

Table 1 Sound Evaluation (for Reference)

5. Audio Vivid Content Production Renderers and Usage Examples (Based on Speakers)

5.1 Environment Setup

There are two types of content production tools: plug-in renderers for digital audio workstations (DAWs) and hardware rendering devices.

The following figure shows an example of a DAW plug-in renderer:

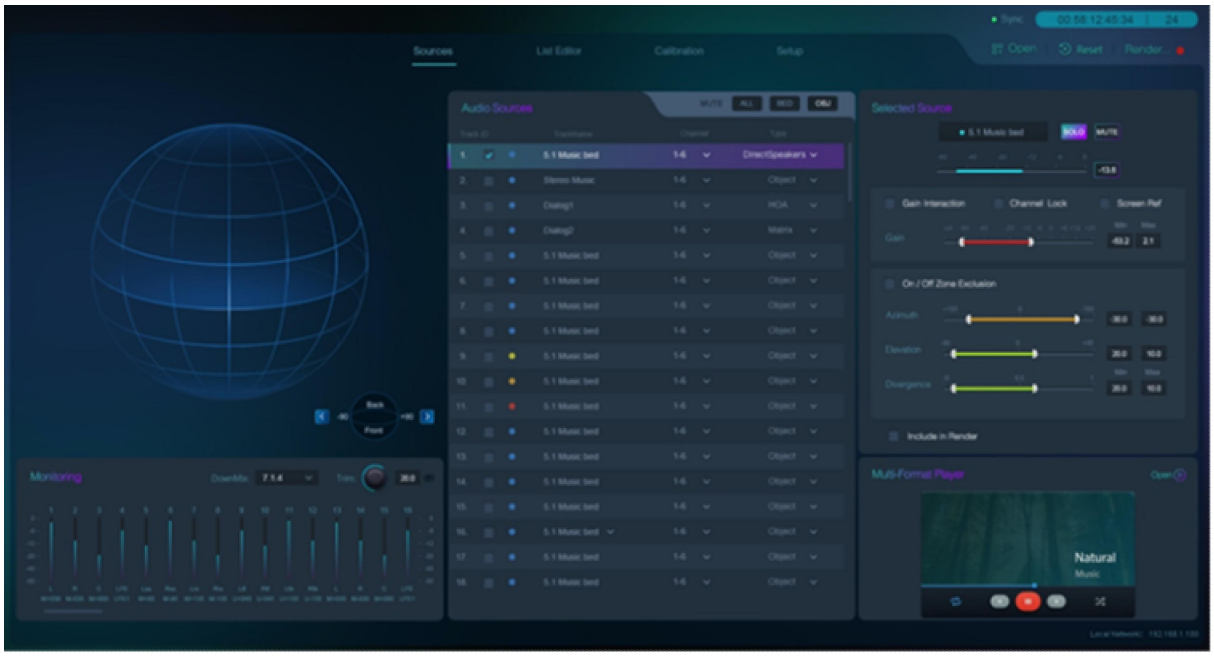

Figure 19 DAW Plug-In Renderer Example

The figure below shows a hardware rendering device workflow diagram:

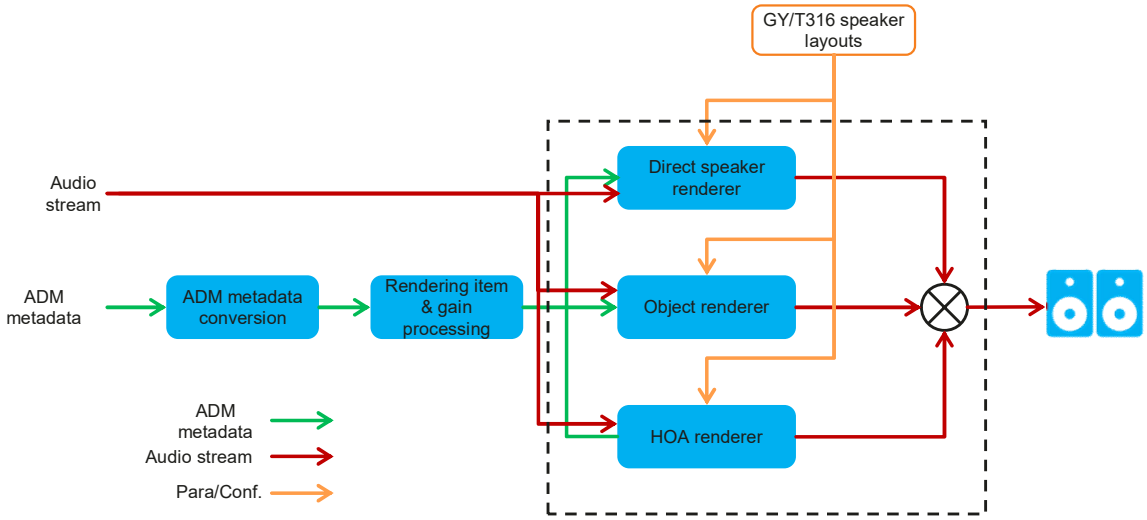

Figure 20 Hardware Rendering Device Workflow Diagram

5.2 Workflow

The workflow, from content production to playback, is shown below:

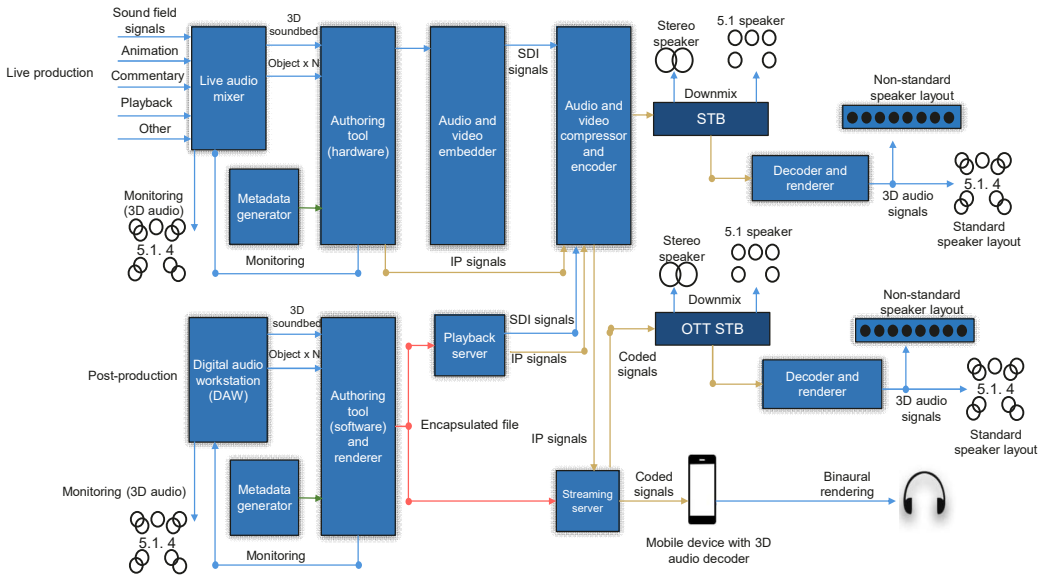

Figure 21 Workflow Diagram

5.3 Operation Procedures

(1) Create an audio mix in the DAW, and perform editing and signal processing during content production.

(2) Route audio channels to a new software renderer or rendering device, rather than the output bus in the DAW.

(3) Position each audio track in a 3D space using the built-in 3D Panner (if available) in the DAW.

(4) Send panning information to the renderer as separate metadata, alongside the audio signal.

(5) Connect Studio speakers (surround height speakers) to the output of the renderer. The renderer software processes the audio signal and its individual panning information in real time, playing the audio as a proper 3D sound field through the speakers.

(6) Store the audio mix in a specific ADM file, separating the audio signal from its panning information.

(7) Open the ADM file in the renderer to monitor and edit the ADM metadata configuration, or convert the ADM file into various media files.

(8) Upload the ADM file (representing the DAW audio mix) into the renderer, which can be converted into various media file formats and used as deliverables for various distribution channels.

(9) Code deliverables in a special bitstream format, so that the DAW audio mix can be sent to the end user via streaming services.

(10) The consumer receives the DAW audio mix as an encoded file, which still keeps the audio signal and its panning information separate, and the playback device performs actual rendering to available speakers or headphones.

6. Audio Vivid Application in the Thousands of Screens in Hundreds of Cities Program

6.1 Introduction to the Thousands of Screens in Hundreds of Cities Program

The Thousands of Screens in Hundreds of Cities Program is a campaign to promote UHD video, which is jointly implemented by six Chinese government departments: Ministry of Industry and Information Technology, Publicity Department, Ministry of Transport, Ministry of Culture and Tourism, National Radio and Television Administration, and China Media Group. To date, more than 100 UHD screens have been installed in 35 cities across China, allowing the general public to watch the China Media Group's 8K UHD TV channels. An audio service – Portable Audio – is simultaneously being provided for mobile devices, in order to better promote UHD videos in the Thousands of Screens in Hundreds of Cities Program. This means users can enjoy audio services anytime and anywhere while watching UHD videos on large outdoor screens, free from sound disturbance. In August 2022, supported by Audio Vivid technologies, the China Media Group allowed users of the Portable Audio service to receive 3D signals that perfectly synced audio and video. On September 10, 2022, for the first time, the China Media Group Mid-Autumn Festival Gala was broadcast simultaneously to screens installed as part of the Thousands of Screens in Hundreds of Cities Program. This was achieved through the application of Audio Vivid audio encoding and decoding technologies, bringing an audio-visual feast to the public.

6.2 Portable Audio Platform in the Thousands of Screens in Hundreds of Cities Program

The primary purpose of the Portable Audio Platform is to support the encoding and transmission of TV audio to mobile devices, thereby allowing the audio and video content of TV channels to be transmitted separately but presented in sync. This can greatly improve the viewing experience of users watching large outdoor screens. The audio service provided by the Portable Audio Platform was upgraded from stereo to Audio Vivid 3D audio on August 18, 2022, marking the beginning of 3D audio's commercial use in live broadcasts. The figure below shows the Portable Audio Platform's system architecture:

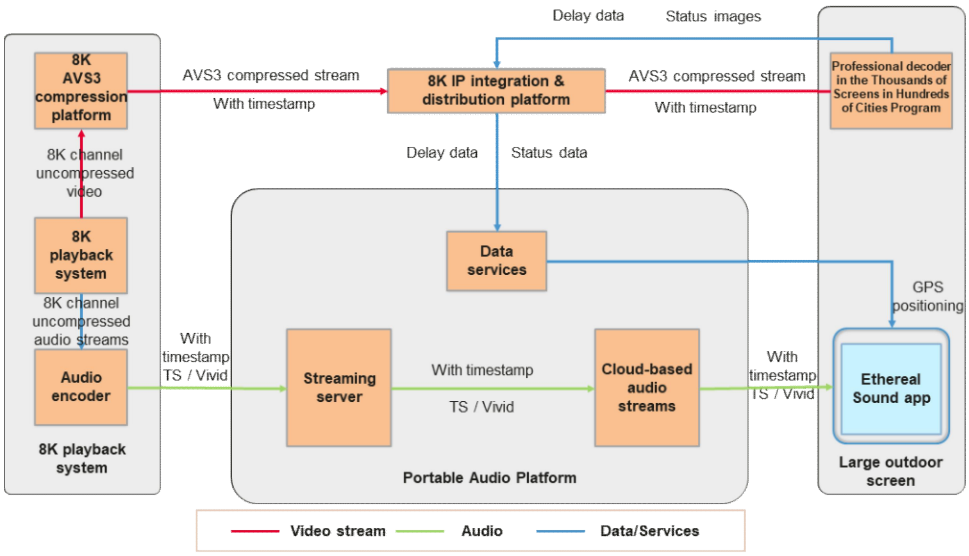

Figure 22 Portable Audio System Architecture

The Portable Audio Platform uses Audio Vivid audio from end to end, from encoding and encapsulation to transmission and distribution, and supports playback on a mobile device that integrates the Portable Audio service after decoding and binaural rendering. The key advantages include:

1. Broadcast-level Audio Vivid encoders that use x86 servers have been developed to support multi-protocol encapsulation and multiple signal source formats, allowing Audio Vivid 3D audio streams to be output through the 8K playback system of the China Media Group.
2. Based on MPEG transport stream encapsulation, Audio Vivid servers have been developed for content distribution and transmission over both the Internet and mobile Internet. These servers support functions like Audio Vivid audio stream encapsulation, stream pushing, and streaming services, making it possible to provide Audio Vivid audio services for users over the Internet.
3. In the Thousands of Screens in Hundreds of Cities Program, audio and video signals need to be transmitted through different links and played/displayed on different terminals. Therefore, a customized audio/video sync mode – synchronous audio/video transmission over heterogeneous networks – has been implemented as part of the Program. With this mode, the absolute timestamp of the broadcast time is embedded in AVS3 video streams, and the Portable Audio Platform calculates the link delay based on the local time and reports the information during decoding. The Portable Audio Platform then relays the delay information of the related decoder to the mobile device based on the device's positioning information. The decoder can sync audio and video signals using the delay and timestamp provided in the Audio Vivid audio streams, meeting the national standard (-90 ms to 120 ms).

6.3 Portable Audio Service in the Thousands of Screens in Hundreds of Cities Program

The Portable Audio service is integrated within the Ethereal Sound app. Ethereal Sound is a mobile audio app launched by the China Media Group based on 5G technology. Focusing on content like news, general knowledge, and culture, Ethereal Sound collects high-quality content from the China Media Group and offers self-made audio programs and high-quality audio books. Ethereal Sound is committed to providing full-scenario audio services for a variety of users, including those that use mobile phones, head units, tablets, and smart wearables.

In the Thousands of Screens in Hundreds of Cities Program, users can link the Ethereal Sound app to a designated large screen by scanning the relevant QR code or through manual selection. They can then obtain the audio stream URL for that large screen and the corresponding playback delay (tvideo) caused by video decoding through the China Media Group's media gateway interface. Following this, audio stream data that is encoded using Audio Vivid and encapsulated in transport streams is read using the obtained URL. Finally, Audio Vivid audio signals are played in sync with that large screen through steps like encapsulation, decoding, and rendering. Specifically, the playback process includes the following steps:

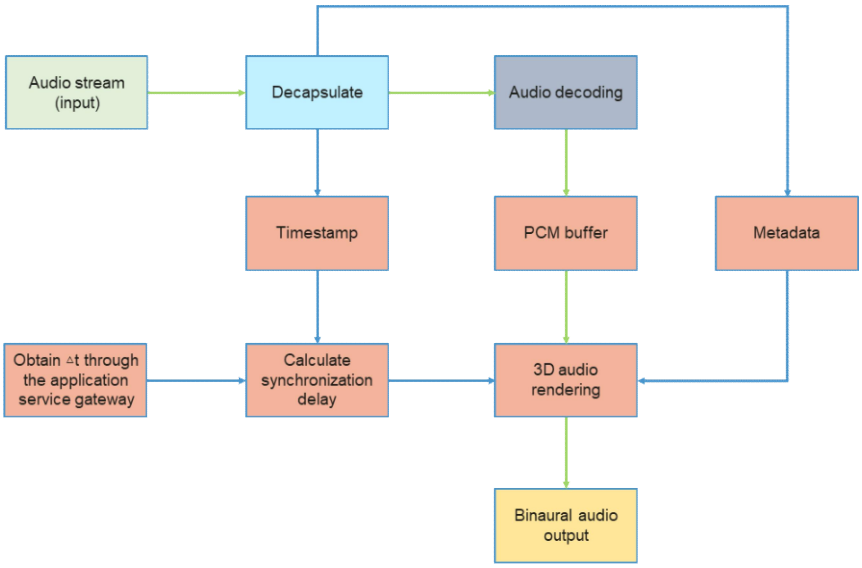

Figure 23 The Ethereal Sound App Playback Workflow

1. Audio stream input: The audio player obtains the digital audio stream that is encoded using Audio Vivid and encapsulated in transport streams according to the network protocol used by the URL.

2. Audio decapsulation: The audio player decapsulates the audio data encoded using Audio Vivid, encoded timestamp information, and audio metadata (if any) from the transport streams.
3. Audio decoding: The audio player decodes the decapsulated Audio Vivid audio data into multi-channel PCM data, and then stores the data in the rendering and playback buffer queue.
4. Syncing audio and video: The audio player periodically obtains the playback delay (tvideo) caused by video decoding of the associated large screen from the China Media Group's application gateway. Simultaneously, the player calculates the difference between the encoded timestamp obtained from decapsulation and the local timestamp, in order to obtain the audio decoding delay (taudio). Based on the difference between tvideo and taudio, the player calculates the synchronization delay time that is required for the audio and then adjusts the rendering and playback buffer queue accordingly.
5. 3D audio rendering: The player obtains the to-be-played, multi-channel PCM data from the rendering and playback buffer queue, performs binaural audio rendering according to metadata that contains audio bed information, and outputs the rendered dual-channel PCM data to the sound card device.

7. Audio Vivid Development Recommendations

Building an open ecosystem

Bring together top industry players from around the world to build the Audio Vivid ecosystem and promote the iteration of international standards. Drive efforts to promote the Audio Vivid standards in sectors such as radio and television, new media, and entertainment. Grow the influence of the technical specifications within vertical industries. Promote closer cooperation across the industry, both upstream and downstream, to integrate the end-to-end industry value chain.

Demonstrating technical standards in real-world projects to drive Audio Vivid adoption

Apply Audio Vivid standards during major sports events and activities to drive their adoption and popularity. Promote their application in the entertainment and manufacturing sectors to create premium case studies. Focus on leveraging these technical standards during key events to achieve early adoption. Produce case studies to facilitate deeper cooperation between upstream and downstream industry players and generate awareness among audiences.

Encouraging businesses to accelerate application and R&D

Over the next two to three years, double down on deploying services such as home theaters, in-vehicle audio and video systems, personal audio and video services, and virtual reality services; motivate the content production sector and distribution platforms to accelerate Audio Vivid content production and distribution; redouble efforts to build the industry value chain; drive industry evolution; and create a solid foundation for the future UHD audio and video industry in terms of both technology and talent.

8. Appendix

8.1 Abbreviations

The following terms and definitions apply to this document:

Sampling frequency (fs)

Audio and video (AV)

Decibels relative to full scale (dBFS)

Fast Fourier Transform (FFT)

Left (L)

Low frequency enhancement (LFE)

Left Rear Surround (Lrs)

Left Side Surround (Lss)

Left Top Back (Ltb)

Left Top Front (Ltf)

Moving Picture Experts Group (MPEG)

Program Map Table (PMT)

Right (R)

Right Rear Surround (Rrs)

Right Side Surround (Rss)

Right Top Back (Rtb)

Right Top Front (Rtf)

8.2 References

[1] T/UWA 009.1-2022 3D Audio Technology Specification: Part 1: Coding, Distribution and Presentation, Standard of UHD World Association (UWA), 2022

[2] T/UWA 009.3-1-20 3D Audio Technology Specification Part 3-1: Technical Requirement and Test Method – Home AV Playback Device, Standard of UHD World Association (UWA), 2022

8.3 Certification and Authorization

Certification is key to the development of an industry ecosystem. The certification system for Audio Vivid is authoritative, standardized, science-based, and open. In addition, the system strictly standardizes the display effects of regular and portable display devices, and provides the highest-quality UHD technical support for certified companies. This will drive the device industry towards higher-standard display technologies.

1. Why certification is important
    - The UWA is a symbol of advanced audio and video technology and ecosystem standards, meaning all UWA-certified products are equipped with high-quality audio and video technology.
    - The certification process is fast and simple.
2. What to expect after being certified
    - The right to use the certificate and certification mark (Audio Vivid)
    - Inclusion in the UWA's public certification directory
    - Prioritized participation in joint marketing activities organized by the UWA for certified companies
    - Long-term and effective high-quality Audio Vivid technical support
    - Participation in professional technical training for technicians and engineers organized by the UWA
3. Certification method

The basic certification process for devices or systems that support Audio Vivid is shown in the flow chart below, with the blue blocks representing the basic certification nodes specified in the *Audio Vivid Certification Rules*. For further details about the certification method, please log on to the UWA website. Following registration, you can download the *Audio Vivid Certification Rules*.

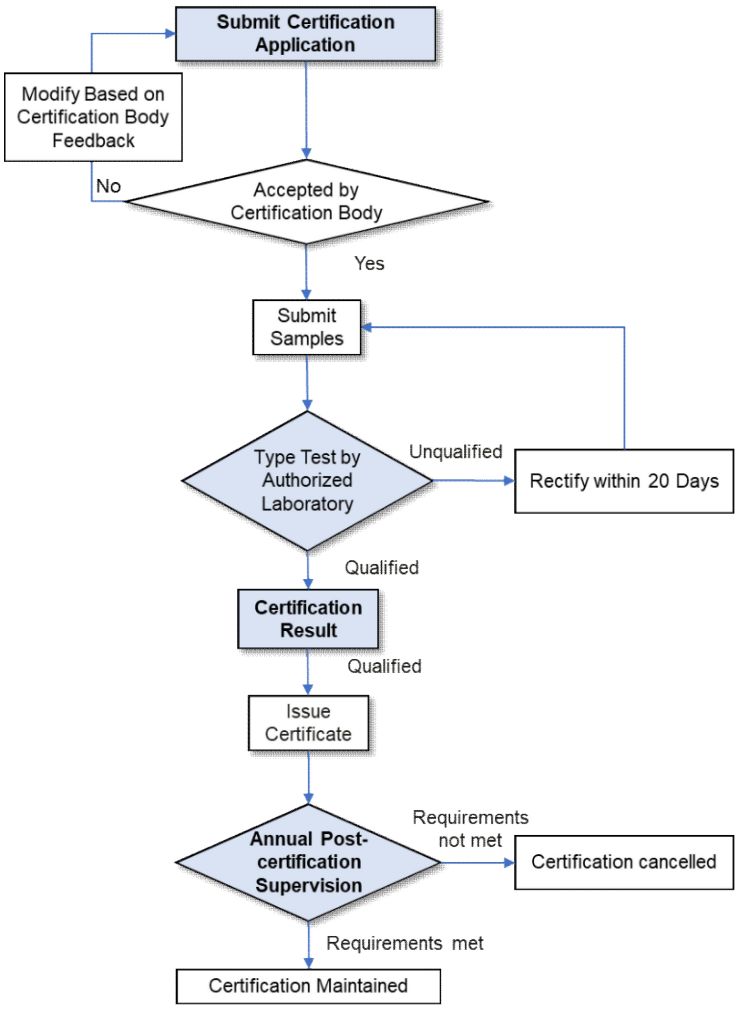

Figure 24 Certification Process

8.4 Testing and Evaluation

1. Example of test method

The general test plan of 3D audio is as follows: Provide the original audio source through the standard test stream (coded data stream) to the AV equipment being tested, and then test it using the sound-generating unit of the equipment or the accompanying sound-generating unit used for test. During this time, the audio is analyzed by the audio analyzer. This test case does not cover binaural headphones. For details about the test and evaluation, refer to the Audio Vivid standard test document. [2]

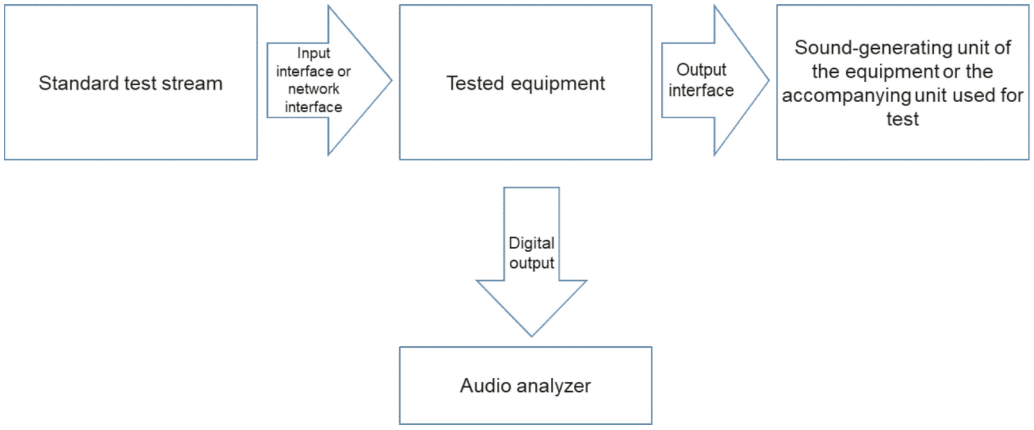

Figure 25 Schematic Diagram of General Test

To verify whether the tested equipment meets the functional requirements, test it as follows:

a) Connect the tested equipment according to the diagram.

b) Provide input to the tested equipment according to the standard test flow.

c) Check P1b – the interface between the tested equipment and the sound-generating unit of the equipment, and P3 – the interface between the tested equipment and the audio analyzer.

2. Functional indicators

An acceptance test is conducted according to the following function table.

SN	Item		Functional Requirements
1	Audio Vivid audio recognition		Devices with UI can identify and decode Audio Vivid audio streams. Specifically, they can correctly identify and decode Audio Vivid from multiple audio streams (Audio Vivid and MPEG-1 Layer II audio) multiplexed in a program. In addition, the devices do not misidentify non-Audio Vivid audio streams as Audio Vivid ones.
2	Bed decoding	Sound channel mapping	All Audio Vivid audio channels can be mapped correctly, including stereo and 5.1.4 multi-channels, and signals from all normal channels can be recreated correctly.
3		Input sampling frequency	The device supports the decoding of Audio Vivid audio with a sampling frequency of 32 kHz, 44.1 kHz,

SN	Item		Functional Requirements
			and 48 kHz, and, preferably, 96 kHz.
4		Bitrate	The device supports the decoding of Audio Vivid audio with a bit rate ranging from 64 Kbps to 832 Kbps.
5		Bit depth	The device supports a bit depth of 16-bit, and preferably, 24-bit in lossless audio decoding.
7	Lossless audio decoding		The device, preferably, supports lossless audio decoding.
8	HOA decoding		The device, preferably, supports the accurate recreation of 3rd-order HOA signals from all directions.
9	Object audio decoding		The device supports the accurate restoration of object audio in all directions.
10	Output sampling frequency		Devices with digital audio output support an output sampling frequency of 48 kHz, and preferably, 96 kHz.
11	Sound rendering		Audio Vivid audio can be downmixed with the maximum audio playback capability set of the device.

Table 2 Functional Requirements

3. Technical indicators

SN	Item	Unit	Performance Requirement
1	Audio signal-to-noise ratio (SNR)	dB	≥ 60
2	Channel pair gain difference	dB	≤ 1
3	Crosstalk of channel pairs	dB	≤ -60
4	Overload	——	No clipping

Table 3 Electrical Signal Requirements

SN	Item	Unit	Performance Requirement
1	Audio frequency response characteristics	dB	The variation (difference between the peak and the valley) within the range of 200 Hz–8,000 Hz must be less than 6 dB.
2	Maximum sound pressure at rated input	dBSPL	≥ 74 dB
3	Total harmonic distortion of sound pressure at rated input	%	≤ 5 (200 Hz–8,000 Hz)

Table 4 Acoustic Signal Requirements

SN	Receiving Terminal Type	Time Difference Range (ms)
1	AV playback & display device	-125–45
2	AV playback device	-30–20 0–20
3	Audio playback device	-
Notes: If the time difference between audio and video signals is -40 ms, this means that the audio signal is 40 ms behind the video signal after being decoded by the television receiving terminal. If the time difference between audio and video signals is 20 ms, this means that the audio signal is 20 ms ahead of the video signal after being decoded by the television receiving terminal.		

Table 5 Time Differences

SN	Item	Functional Requirement
1	Gain control	The device provides an accurate analysis of loudness metadata and accurate control of loudness, with the

SN	Item	Functional Requirement
		deviation not exceeding 2 dB.
2	Dialogue gain control	The device provides an accurate analysis of dialogue loudness metadata and accurate control of dialogue loudness, with the deviation not exceeding 2 dB.

Table 6 Metadata Processing Performance

UHD World Association
世界超高清视频产业联盟