ABU TECHNICAL COMMITTEE MEETING 2022

New Delhi India

26 - 27 November 2022

Doc T-22/23-1

Academy of Broadcasting Planning, NRTA

Audio Vivid – The next generation of audio coding standard

Jinhui Ning

Abstract: Audio Vivid is a new generation audio codec standard developed and released by UWA (UHD World Association). It is the first audio coding standard using AI technology in the world to address the 3D audio compression challenge for the immersive audio experience. The standard consists of a series of technical specifications and solutions to meet the demands from the industry, and can be used in a wide variety of audio scenarios, including consumer entertainment, cinema, AR/VR, and automobile audio systems, etc. End-to-end system have been constructed and tested based on these standards to demonstrate their technical advantages and the readiness for large-scale commercial use. More specifications for different application scenarios, test, and certification are under development and making rapid progress.

Key Word: Ultra HD video industry, Audio Vivid

1. Introduction

With the popularization of 5G technologies and the further improvement of terminal display capabilities, the era of ultra-high definition (UHD) audio and video experience has arrived.

From the perspective of industry development, the coordinated development of the audio and video industries is accelerating. Compared with dual-channel stereo and multi-channel surround sound, 3D audio provides a richer spatial sound field and a sense of immersion. It is one of the six core experiences of UHD technology and a key component of spatial audio and virtual reality experiences. It includes not only sound channel information, but also object and scene information. It requires end-to-end collaboration of sound collection and production, encoding and transmission, rendering and playback technologies to provide optimal auditory experience.

Audio Vivid is a new generation audio codec standard released by UWA (The UHD World Association), a non-governmental organization voluntarily initiated by leading enterprises engaged in the manufacture, transmission, content production, application, and service of ultra HD video products. Members of the Alliance formulate product standards that meet the same technical specifications through consensus.

Audio Vivid is the next generation high-efficiency compression methods of immersive audio in a wide variety of scenarios. It can be applied in the home environment, cinema environment, personal, AR/VR, and vehicle-mounted environments.

2. Characteristics of Audio Vivid

2.1 Overview of the coding system

Audio Vivid coding system includes lossy coding, lossless coding, speaker array renderer and binaural renderer. Supported signal formats include channel based signal, object based signal, higher-order ambisonics (HOA) based signal, metadata. **Error! Reference source not found.** shows the framework of the Audio Vivid decoding system.

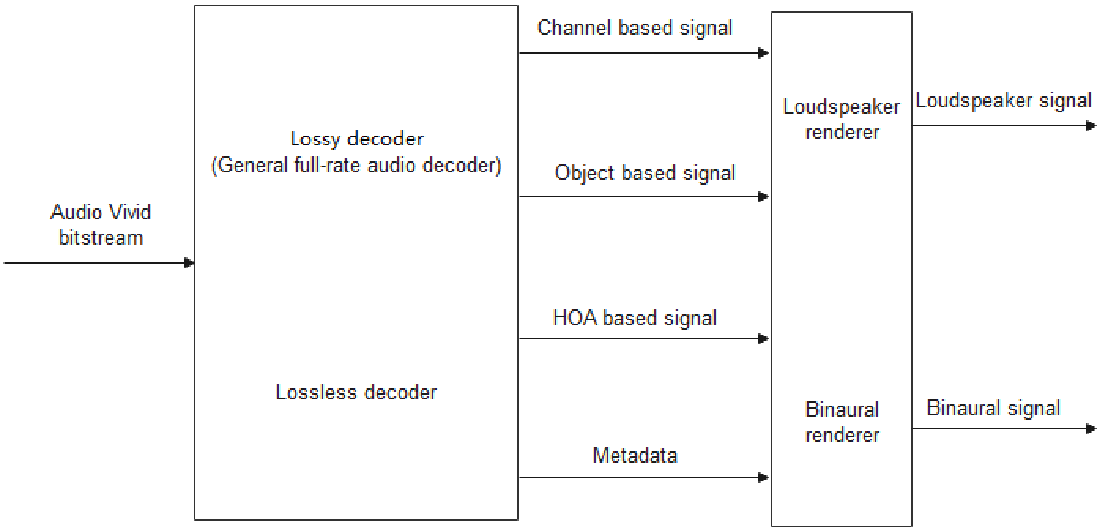

Figure 1. Framework of the Audio Vivid decoding system

2.2 General full-rate audio coding

General full-rate audio coding is the key module of Audio vivid system. 2 shows the basic architecture of a general full-rate audio coder. It supports the mono, stereo, multi-channel, object, HOA and metadata as the input, supports sampling rate from 32 kHz to 192 kHz, the bit width from 16 bits to 24 bits, bitrate from 32 kbit/s to 1.6 mbit/s. The general full-rate audio encoder consists of transient state detection, window type judgment, time-frequency transform, frequency-domain noise shaping, temporal noise shaping, bandwidth extension, downmixing, neural network transform, quantization, and range encoding. It encodes channel signals and object signals into bitstreams. The HOA spatial encoder and core encoder encode HOA signals into bitstreams.

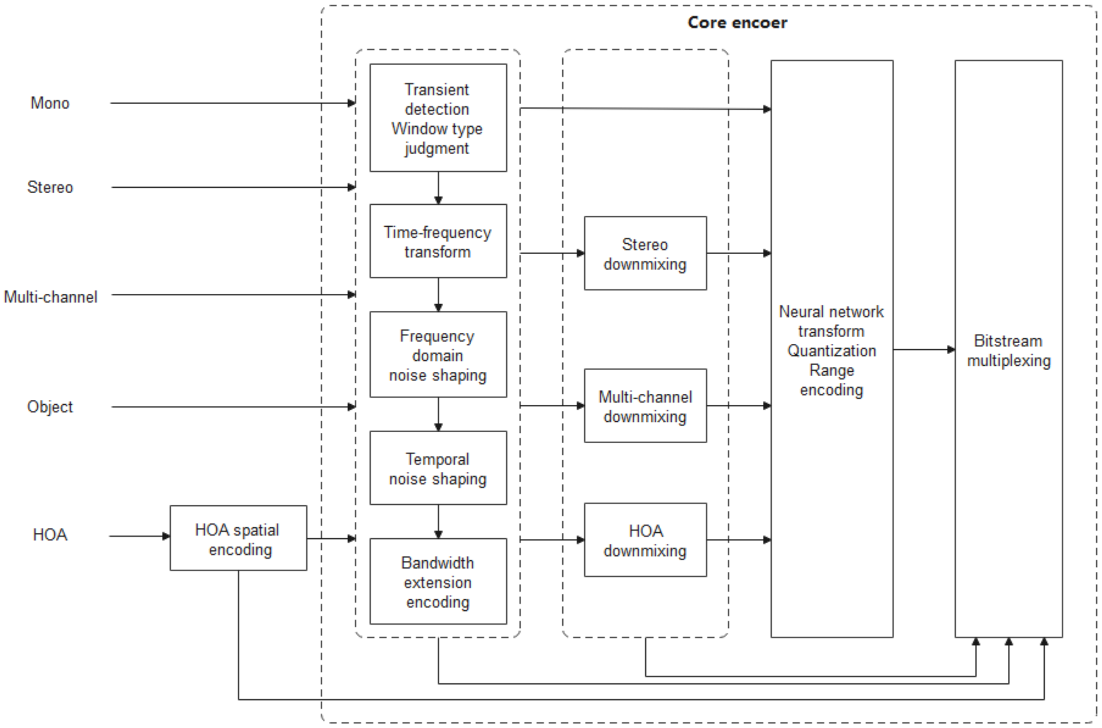

Figure 2. Architecture of a general full-rate audio coding

2.3 First 3D audio coding standard based on the AI

Audio Vivid is the world's first audio coding standard using AI technology. With AI technology Audio Vivid not only improves audio compression efficiency, but also overcomes long-standing patent barriers in some countries. Audio Vivid uses a hybrid AI encoding and decoding architecture. Traditional encoding and decoding technologies are used in the preprocessing, and AI-based technologies are used in the feature transformation and quantization entropy encoding.

This hybrid AI architecture not only combines the essence of traditional audio compression theory (psychoacoustics theory) and the advantages of deep learning to extract abstract features, but also strikes a reasonable balance between performance and overhead. In the hybrid AI architecture, an audio signal is converted from a time domain signal to a frequency domain MDCT signal in a preprocessing stage, and frequency domain noise shaping and time domain noise shaping are performed on the MDCT signal, and possible downmixing processing is performed on the MDCT signal, and then output to the AI processing stage. In the AI processing phase, a deep neural network is used to convert the MDCT signal into a hidden feature signal, and then quantize and entropy code the hidden feature signal. The purpose of generating hidden feature signals is to obtain features that are more favorable for efficient entropy coding. The scalar quantized hidden feature signal is sent to an AI-based entropy coding module. The entropy encoding module generates a context of the to-be-encoded implicit feature signal by using another deep neural network, and selects a corresponding codebook according to the context to perform entropy encoding on the implicit feature signal. The foregoing two deep neural networks are jointly trained, and a relationship between a to-be-encoded feature, a context, and each codebook is jointly searched under a constraint of minimizing information entropy, thereby fully utilizing a powerful abstraction capability of the deep neural network. To facilitate deployment of the decoder on a plurality of platforms, especially a mobile platform, and minimize storage and calculation overheads of the decoder, the foregoing two deep neural networks further adopt asymmetric designs at the encoder end and the decoder end. The encoder end uses a larger neural network to ensure relatively high compression efficiency. The decoder uses a smaller neural network to reduce the overhead.

2.4 HOA spatial coding based on virtual loudspeaker projection

Audio Vivid greatly improves the coding efficiency of HOA signals by providing a HOA spatial encoding algorithm. The HOA spatial coding algorithm assumes that several virtual speakers are distributed around scene. Figure 3 shows an example of virtual loudspeaker distribution.

The HOA signal is approximated by a linear combination of several virtual loudspeaker signals and the difference between input signal and the reconstruction signal which is recovered by the selected virtual loudspeakers. An innovative virtual loudspeaker selection is proposed. The principle for selecting the virtual loudspeakers is that using virtual loudspeakers to represent the sound source in the scene and the selected virtual loudspeakers minimizes the residual signal that difference between input signal and the reconstruction signal. With the HOA spatial coding the input signal is transformed into a few of virtual loudspeaker signals and residual signals and side information, and then send them to the coder. The quantity of the virtual loudspeaker signals and residual signals is far less than the quantity of the input signal, and the residual signal can be coded by few bits, therefore the HOA spatial coding greatly improves the HOA coding efficiency.

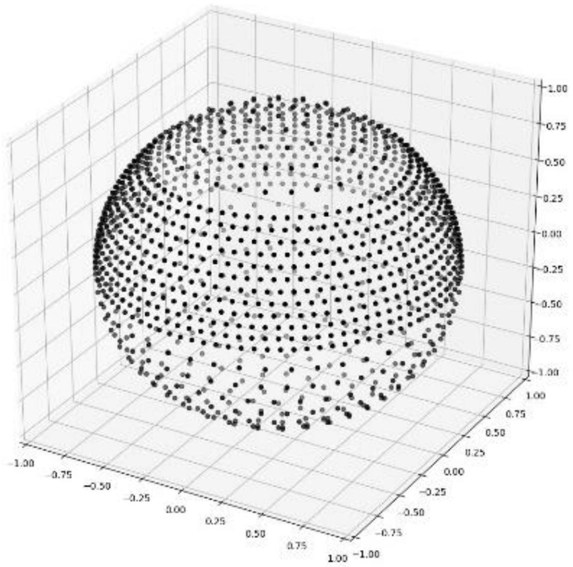

Figure 3. Virtual loudspeaker distribution

2.5 Flexible metadata system

Metadata system is one of the important features of Audio Vivid system. Metadata describes audio signal format, audio signal content, a physical attribute of the audio signal in a playback space, a sound effect parameter of the audio signal, the mechanism how the audio signal can interact with a user, and so on. Audio Vivid metadata system is designed as a hierarchical structure consisting of a base layer and an extension layer. The basic layer metadata multiplexes the attributes and elements of the audio signal content and format defined in the international standard ITU-R BS.2076-2 (ADM), and is used to transfer the content and control information related to audio signals such as bed, matrix, object, HOA, and binaural. The innovative extension layer metadata provides enhanced binaural rendering features and supports binaural rendering to better restore director's artistic intents through metadata such as acoustic environment and rendering sound effect post-processing. Audio Vivid metadata system not only meets the global interconnection of metadata, but also provides sufficient flexibility and extensibility. It provides powerful representation and interaction capabilities for the new generation of audio systems.

2.6 Compatible with more endpoint rendering and playback technologies

Audio Vivid also provide basic rendering technologies including loudspeaker renderer and binaural renderer. It is compatible with more terminal rendering and playback technologies. Standards ensure interconnection and interoperability, promote the prosperity of the technology ecosystem, and provide rich and differentiated experience for different application scenarios.

3. Application of Audio Vivid

3.1 Ecosystem-friendly and applicable to all scenarios

Audio Vivid is a global and advanced technical standard and solution. It is friendly to industry ecosystem policies and more suitable for end-to-end industry deployment by all parties in the UHD industry ecosystem. In terms of application scenarios, Audio Vivid serves the entire process of sound from construction to restoration.

3.2 Rollout of Audio Vivid

"One Hundred Cities and Thousand Screens" is an ultra-high definition (UHD) video promotion activity jointly deployed by the Ministry of Industry and Information Technology, the Ministry of Transport, the Ministry of Culture and Tourism, the State Administration of Radio and Television, and the China Media Group . At present, more than 100 ultra-high definition (UHD) screens have been deployed in 35 cities across the country. The public can watch China Media Group 8K ultra-high definition (UHD) TV channels on the public screens. At the same time, to cooperate with the promotion of ultra-high definition (UHD) videos, CMG provides the audio service of ultra-high definition (UHD) videos through mobile terminals on the basis of the original ultra-high definition (UHD) videos without causing sound interference.

In August 2022, the headquarters used the 3D Vivid technology to listen to 3D audio with accurate audio and video synchronization on the mobile platform of the 100-

city thousand-screen Walkman. On September 10, the China Media Group Mid-Autumn Festival Gala was broadcast simultaneously by Audio Vivid through the broadcast platform of "hundred cities and thousands of screens" for the first time, bringing an audio-visual feast to the public.

4 Conclusion

As a new force in the field of audio coding standard, Audio Vivid breaks the channel limits and gives each sound object a unique personality, allowing the sound to linger around and above the listener for stunning realism. Compatible metadata is introduced throughout content creation, service distribution, and terminal presentation to ensure end-to-end effect transmission. The standard will create greater value for the development of the ultra-high definition (UHD) video industry and bring infinitely wonderful audio-visual feasts to the industry and consumers.