MPEG-I Immersive Audio
The Upcoming New Audio Standard for Virtual / Augmented Reality

Andreas Silzle[1], Sascha Disch[1], Alexander Adami[1], Nils Peters[2] and Jürgen Herre[2]
[1]*Fraunhofer IIS,* [2]*International Audio Laboratories Erlangen, E-Mail: andreas.silzle@iis.fraunhofer.de*

Abstract

Research in the field of auditory virtual environments has a long history. When combined with a visual counterpart and motion tracking, virtual environments can be used in widespread Virtual Reality (VR) or Augmented Reality (AR) application fields. Only recently, the visual part of such VR implementations achieved acceptable characteristics (resolution, latency, price, etc.) for the mass market. International standards have often helped an application field to get widely used by providing formats and software that enable both interoperability between different implementations and format stability over long periods of time. In this way, consumer electronics (CE) companies and service providers creating infrastructure and content for different platforms can jointly establish a healthy ecosystem and reach the population beyond early adopters. Currently, the MPEG Audio working group defines the new MPEG-I Immersive Audio standard for Virtual and Augmented Reality. The normative bitstream and renderer are defined to provide a real-time auditory world consisting of spatially distributed sound sources and listeners with six Degrees of Freedom (6DoF), both of which can move interactively. This includes the modelling of point sources, sized sources, coupled rooms and the realistic auralization of (room) acoustic phenomena, like reflections, diffraction, occlusion, late reverberation, Doppler and more. A specifically developed encoder input format allows the definition of such feature-rich acoustic worlds.

Introduction

The aim of an Auditory Virtual Environment (AVE) is to give humans auditory perceptions that do not correspond to their real environment, but to a virtual one. The listener should have a spatial impression of the virtual room and perceive his/her own movements inside the environment as well as the movements of the sound sources, [1]. An AVE is often combined with other modalities, such as visuals or haptics. “The term Auditory Virtual Environment is used to describe the percepts of the listener whereas the term AVE generator is used for the system that creates the signals to achieve this perception”, [2]. An AVE combined with a visual simulation is often called Virtual Reality (VR). The virtual world created with it is called a metaverse. Augmented Realities (AR) enhance the real perceptions with virtual ones, but leaves the contact of the user/listener to the real world.

In general, hearing is one of the fundamental senses that we use to experience, navigate, and interact with the world (real or virtual). Therefore, audio technologies that enable these activities virtually will be a critical component for any kind of VR or AR simulations.

MPEG-I Immersive Audio Architecture

The new upcoming ISO/IEC MPEG-I Immersive Audio standard provides high fidelity audio rendering of the soundscape that is simulated in AR and VR environments. To do so, it interactively models the acoustics of a virtual scene by simulating its relevant physics.

Figure 1 shows an overview of the MPEG-I Immersive Audio architecture. The MPEG-I encoder processes PCM audio objects, channels and HOA signals alongside with raw metadata that describes relevant acoustic properties of a VR/AR scene and the associated audio signals.

For authoring and storing the raw metadata that is input to the encoder, a dedicated scene description file format has been developed named Encoder Input Format (EIF) that holds definitions for e.g., sources and their positions, scene geometry (primitives, meshes or voxels), transforms, material acoustic coefficients, etc.

The output of the encoder is a compressed metadata bitstream that is packaged in MHAS (MPEG-H Audio Streaming) compatible data packets. The MPEG-I compressed metadata bitstream, together with MPEG-H encoded audio objects, channels and HOA signals (that are also packaged in MHAS packets) are transmitted to the decoder. All MHAS packets can be multiplexed for broadcast and streaming or simply stored in a file.

At the receiver, the decoded PCM audio and the MPEG-I bitstream are input to the MPEG-I Renderer. Additionally, the renderer has an interface to input local information about the consumption environment, scene updates and, most importantly, for instantly tracked user position, orientation, and user interaction with the scene. User interaction can be e.g., opening or closing doors or manipulating objects. In this way, the renderer is able to interactively render the scene with 6DoF for the user that moves freely within the scene.

As official test material, a set of test scenes was composed. These scenes were created by WG6 experts and are suitable to assess relevant aspects of 6DoF immersive audio rendering like localization of point sources, perception of spatially extended sources, reflections, diffraction, occlusion, late reverberation, Doppler and more.

The test scenes also have visuals associated with them to allow realistic testing conditions. The still images in the following sections have been captured from these test scenes. However, the relevant geometry for audio rendering is not always identical with the visual geometry. In most cases, the ‘audio geometry’ is a simplified version of the visual geometry.

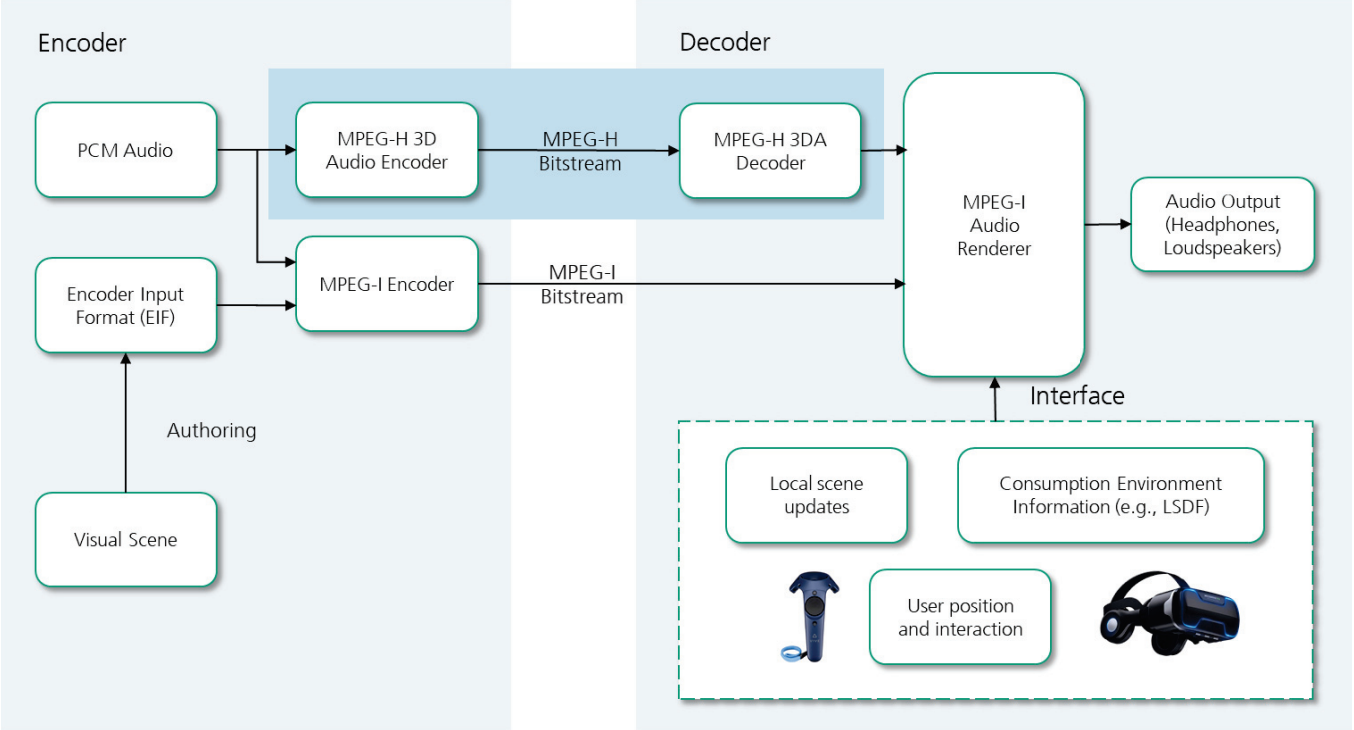

Figure 1: MPEG-I Immersive Audio architecture

Application Fields

Some of the envisioned application fields for MPEG-I are:

- Conferencing, Social VR
- Edutainment
- Enterprise solutions (B2B digital twin, training)
- Virtual museums, art presentations
- Medical solutions
- Immersive live performance

Features of MPEG-I Immersive Audio

Figure 2: Visual and acoustic geometry. For the acoustical rendering not all details of the visual rendering are necessary.

Figure 3: Visual and acoustic materials. Each mesh face can be defined with acoustic properties. This allows the calculation of reflections, coupling between rooms, acoustic transmission through e.g., closed doors and occlusion.

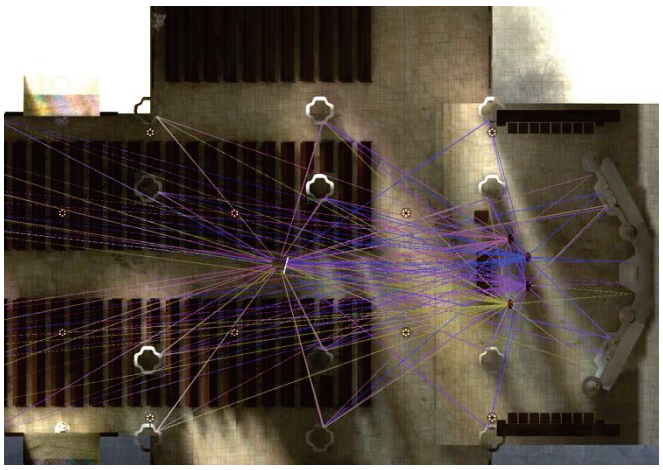

Figure 4: First and second order early reflections from four singers to the user/listener in a church (top view).

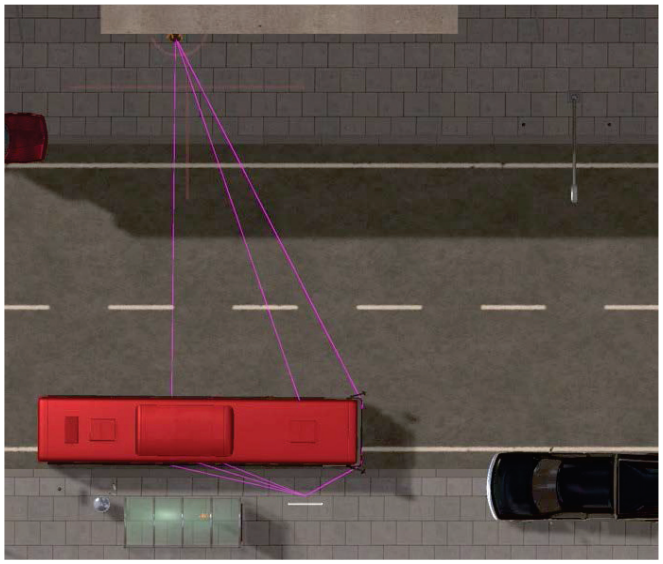

Figure 5: Occlusion and Diffraction around a red bus in an outdoor scene (top view). The listener (behind the bus at the bottom) can hear the downtown drummer (top) via diffracted sound (diffraction paths indicated by pink beams)

Figure 6: Distance and Doppler effect with a car with siren

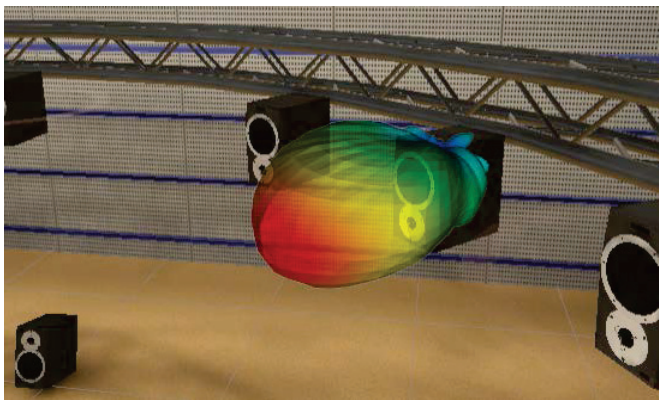

Figure 7: Radiation characteristics of a loudspeaker in a simulated soundlab

Figure 8: Fountain in park is an example of a spatially extended sound source

Augmented Reality Simulation

In an augmented or mixed reality simulation, the listener perceives the real environment plus some additional rendered augmented sound sources, e.g., online participating persons in a teleconference or musical instruments forming a virtual orchestra. To embed such sources convincingly into the real room, the acoustics of the user‘s room has to be taken into account. A Listener Space Description Format (LSDF) describes the room properties with dimensions, acoustics and anchors.

Standardization Timeline

The standardization process for MPEG-I started in 2017 by specifying audio for virtual and augmented reality applications at ISO/MPEG Audio. This included development steps:

- Definition of requirements,
- Development of a real-time testing platform in VR, the Audio Evaluation Platform (AEP),
- Definition of a specific audio scene description, i.e., the Encoder Input Format (EIF).
- Research, development, and implementation of a real-time AVE generator.

The competitive phase of the standardization ended with the successful conclusion of the MPEG-I Call for Proposal (CfP) that had been issued in April 2021. In January 2022, the competitive evaluation determined the winners of the process who then formed the so-called MPEG-I Audio Reference Model (RM0). In the present (second) phase of the MPEG-I standardization, further technology improving the performance of the existing implementation from other proponents is added based on the Reference Model technology. For more details about this process, the CfP selection and listening test results see [3]. The envisioned standardization timeline for MPEG-I is:

- October 2023: Committee Draft (CD), end of technical development
- January 2024: Draft International Standard (DIS)
- April 2024: verifications tests
- July 2024: Final Draft International Standard (FDIS)
- October 2024: International Standard (IS)

In contrast to proprietary products and systems, audio standards in general play an important role to ensure long-time stability as well as interoperability between different products implementing the same standard. They ensure that media will exhibit a well-defined quality. MPEG-I Immersive Audio and MPEG-I Immersive Video could serve as a fundament for metaverse applications and their interoperability.

Actual information to the MPEG-I standardardization and sound examples can be found here:
https://www.iis.fraunhofer.de/de/ff/amm/for/mpegi.html

Acknowledgements

The presented technical work is the result of a large-scale team effort by Ericsson, Fraunhofer IIS/International Audio Laboratories Erlangen, and Nokia. Additional technology contributions come from Dolby Laboratories, Philips, and Qualcomm.

Literature

[1] Blauert, J., Spatial Hearing, The Psychophysics of Human Sound Localization. 2nd ed. 1997, Cambridge Massachusetts: MIT Press.

[2] Silzle, A., Generation of Quality Taxonomies for Auditory Virtual Environments by Means of Systematic Expert Surveys. Institut für Kommunikationsakustik, Ruhr-Universität Bochum, Doctoral Dissertation. 2007, Aachen: Shaker Verlag.

[3] Herre, J. and S. Disch, MPEG-I Immersive Audio – Reference Model For The New Virtual / Augmented Reality Audio Standard. Accepted for publication in J. Audio Eng. Soc., 2023.