”What Is Going On In MPEG Audio?” About Virtual Worlds, Quality Evaluation & More

Sponsored jointly by the
AES TC on Coding of Audio Signals, and by the
AES TC on Perceptual and Subjective Evaluation of Audio Signals

Jürgen Herre

International Audio Laboratories Erlangen
Erlangen, Germany

Workshop Context: MPEG Audio (ISO/IEC JTC1/SC29/WG6)

- Over the past decades, ISO/MPEG standardization has been successfully driving the state of the art in perceptual audio coding, including:
  - MPEG-1 Audio incl. mp3 (1992)
  - MPEG-2 Advanced Audio Coding AAC (1997)
  - MPEG-4 High Efficiency AAC (2003 & 2004)
  - …
  - Unified Speech an Audio Coding USAC (2012)
  - MPEG-H 3D Audio (2015/17)
    Extremely versatile codec for next-generation audio (NGA) systems
- What comes / came after MPEG-H? Any new project?

The Current Main Project In MPEG Audio

- Since ca. 2017, a new project was implemented in MPEG Audio which gradually became the group's main activity:

- "MPEG-I Immersive Audio"
  - Specification for Audio for Virtual & Augmented Reality (VR/AR)
  - Follows the general trend of past MPEG Audio projects: More and more rendering (on top of a highly developed low bitrate coding kernel)
  - Not only perceptual audio coding, but mainly *rendering* of audio in a multi-modal context (3 involved senses: Audio, Visual, Proprioception)

  ⇒ New challenges (multi-modal, highly interactive / real-time responsive …)

- This workshop presents an overview of the MPEG-I Immersive Audio standardization effort and a snapshot of its results

Workshop Overview:

1. MPEG-I Immersive Audio – The Project
   (*Jürgen Herre*)
2. Quality Evaluation for Virtual/Augmented Reality
   (*Thomas Sporer*)
3. MPEG-I Immersive Audio – Where do we stand now?
   (*Jürgen Herre*)

Q&A / discussion …

Note:
Please have your headphones ready for some binaurally rendered examples!

MPEG-I Immersive Audio:

Part 1 - The Project

Prof. Dr.-Ing. Jürgen Herre

International Audio Laboratories Erlangen

Erlangen, Germany

Overview

- Virtual and Augmented Reality (VR/AR) require realistic & immersive audio rendering, both for headphone & loudspeakers reproduction
- "MPEG-I Immersive Audio" specification – currently under development
  Immersive Audio for VR/AR in 3DoF and 6DoF

- Contents of Part 1:
  - Requirements
  - System Architecture
  - Development & Evaluation Environment

MPEG-I Audio
New ISO Standard on Immersive Media (VR/AR)

Objectives

- 3 Degrees of freedom: 3DoF / 3DoF+ (Phase 1)
  - User may turn head in any way (pitch/yaw/roll)
  - Requires ***rotation*** of sound image for binaural headphone playback

  ⇒ This is already addressed by the existing MPEG-H Audio codec

- 6 Degrees of freedom: 6DoF (Phase 2)
  - Users may freely navigate (walk, teleport) and turn their head
  - Requires ***rotation*** and ***translation*** of sound image for binaural playback - plus sophisticated modelling of many position-dependent acoustic effects

  ⇒ To be developed newly – ongoing standardization process

Ongoing Work Item: MPEG-I 6DoF Audio
Some Requirements

- Audio for both Virtual Reality (VR) and Augmented Reality (AR)

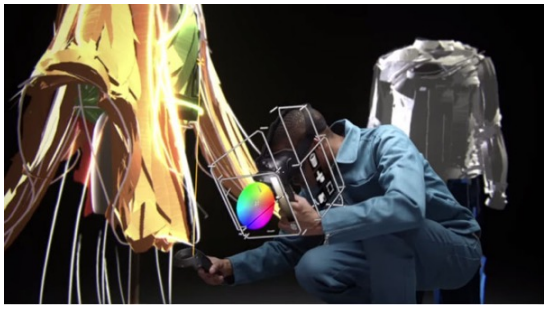
- Playback via headphones (binaural) or loudspeakers
- Spatial sound reproduction (3D sound)
- Sound source models (directivity, spatial extent)
- Convincing simulation of room acoustics (indoor / outdoor)
- Geometry-based effects (occlusion/diffraction sound changes behind obstacles & corners)
- Fast moving sources (Doppler shifts)
- Social VR: Include live sounds of other users (e.g. virtual teleconferencing) and locally captured audio …

Some MPEG-I 6DoF Use Cases
Virtual Concerts

Experience a virtual concert in 6-DoF and move through the venue

Some MPEG-I 6DoF Use Cases
Virtual Art, Virtual Exhibitions

Source: google.com



Source: google.com

Some MPEG-I 6DoF Use Cases
Social VR, Joint Experience

MPEG-I 6DoF Audio
System Architecture

An MPEG-I 6DoF VR/AR Audio system comprises

- Compressed representation of waveforms used in the VR/AR content (channel, object, HOA signals)
- Compressed representation of metadata that describes the properties of the sound sources, acoustic environment, …
- Dedicated 6DoF rendering for headphones and loudspeakers

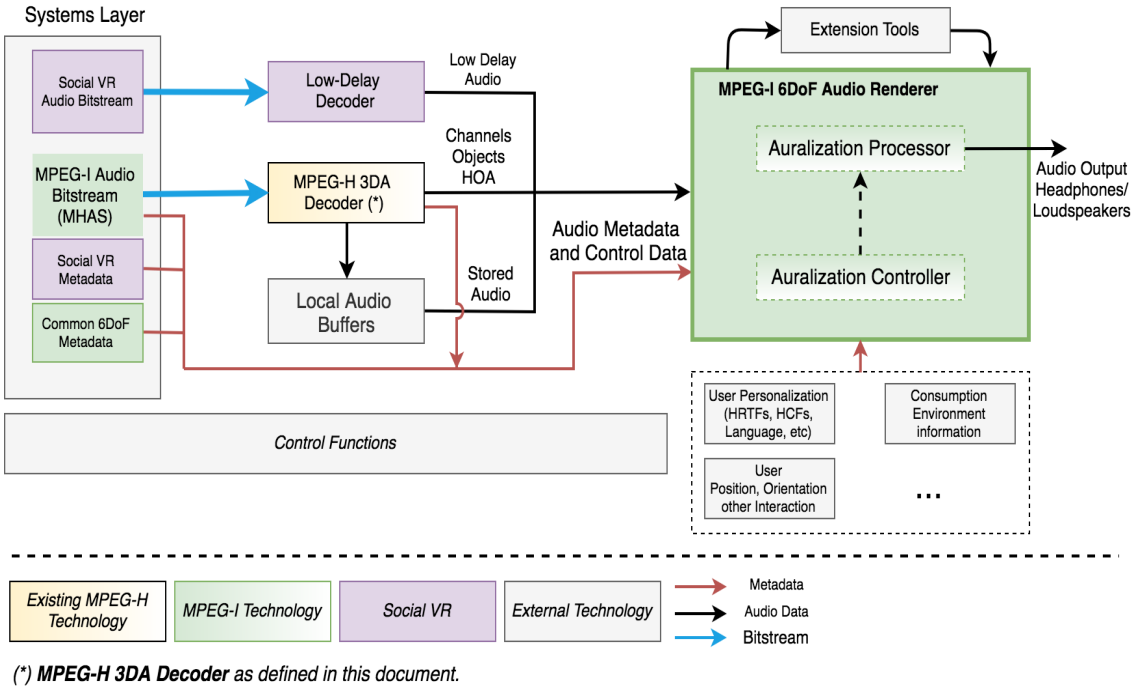
Basic decisions:

- Waveform carriage will employ MPEG-H 3D Audio codec

  ⇒ *Some forward/backward compatibility with MPEG-H Content*

- Additional metadata and rendering to be developed during work item

MPEG-I Audio Renderer Architecture (from N18158)



(*) **MPEG-H 3DA Decoder** as defined in this document.

MPEG-I 6DoF Audio
Setting Up The Environment

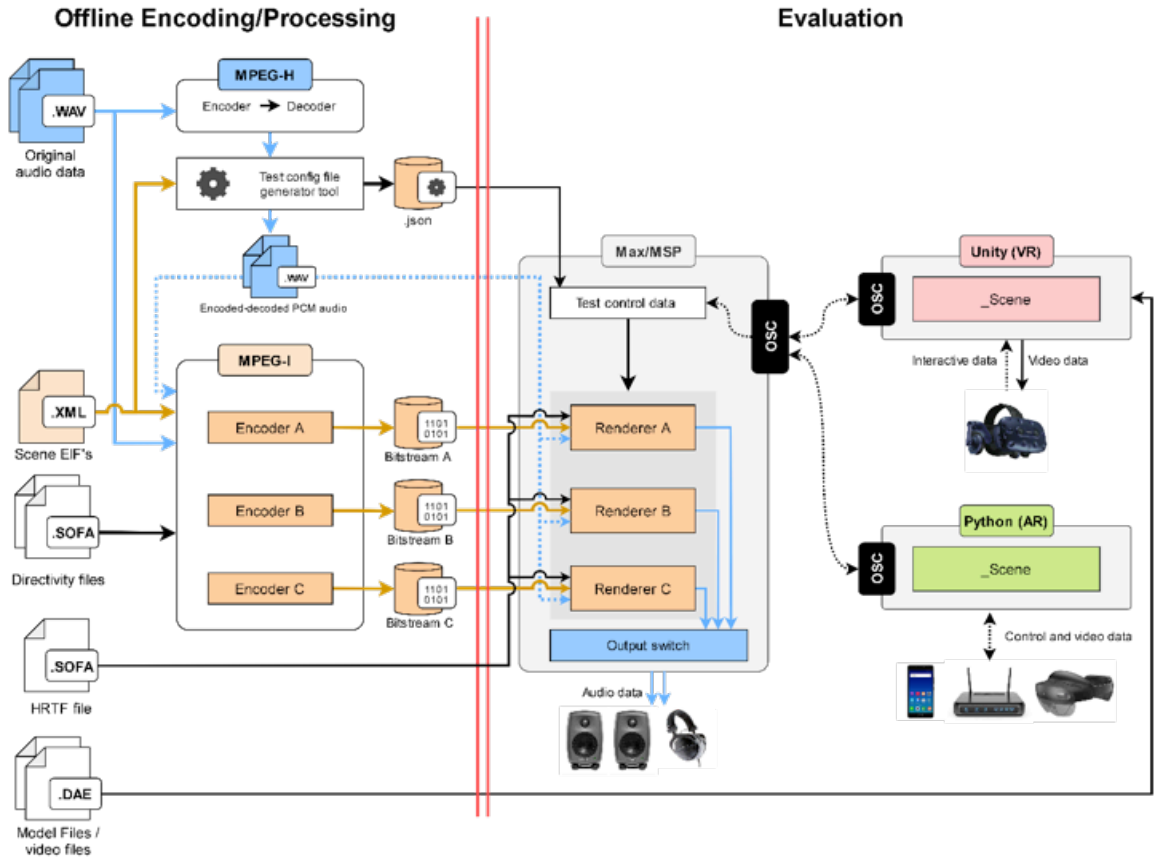
Audio Evaluation Platform (AEP):

- Real-time A/V 6DoF environment with unhindered body motion
- Hardware: PC + VR/AR Hardware (HMD incl. tracker and controllers)
  VR: HTC Vive Pro, AR: MS HoloLens(2)
- Visual host/rendering by Unity (i.e CG-based)
- Audio host: Max/MSP + different audio renderers to be evaluated
  (plugged into Max/MSP)

Content Description & Test Material:

- Defined simple XML-based uncompressed 6DoF scene description format as an "Encoder Input Format" (EIF)
- Collection of rich test material expressed in EIF – testing all required rendering aspects (source size & directivity, occlusion, diffraction, room acoustics, …)

MPEG-I 6DoF Audio Evaluation Platform - Overview



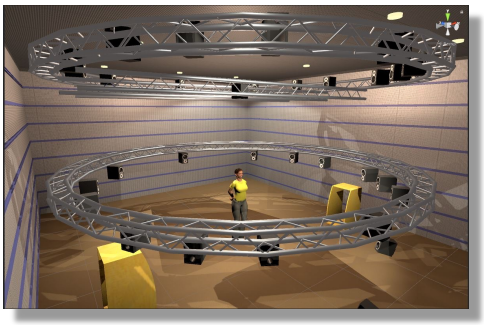
EIF Content Description Example: Audio Object

- Trumpet
  - Position (x, y, z)
  - Orientation (y, p, r)
  - Directivity
  - Gain
  - mode="Continuous"

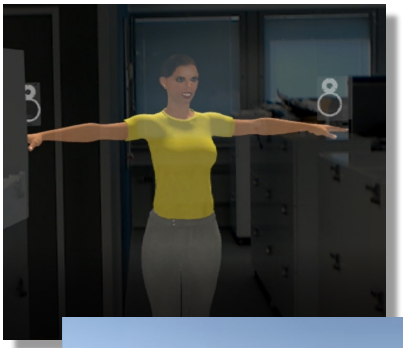
```
<AudioScene>
  <AudioStream id="signal:trumpet"
               file="armstrong.wav"
               mode="continuous" />
  <SourceDirectivity id="dir:trumpet"
                     file="trumpet.sofa" />
  <ObjectSource id="src:trumpet"
                position="2 1.7 -1.25"
                orientation="30 -12 0"
                signal="signal:trumpet"
                directivity="dir:trumpet"
                gainDb="-2"
                active="true" />
</AudioScene>
```

Creation of Test Material – Some Examples

"Singer In The Lab" (VR)



"Singer In Your Lab" (AR)



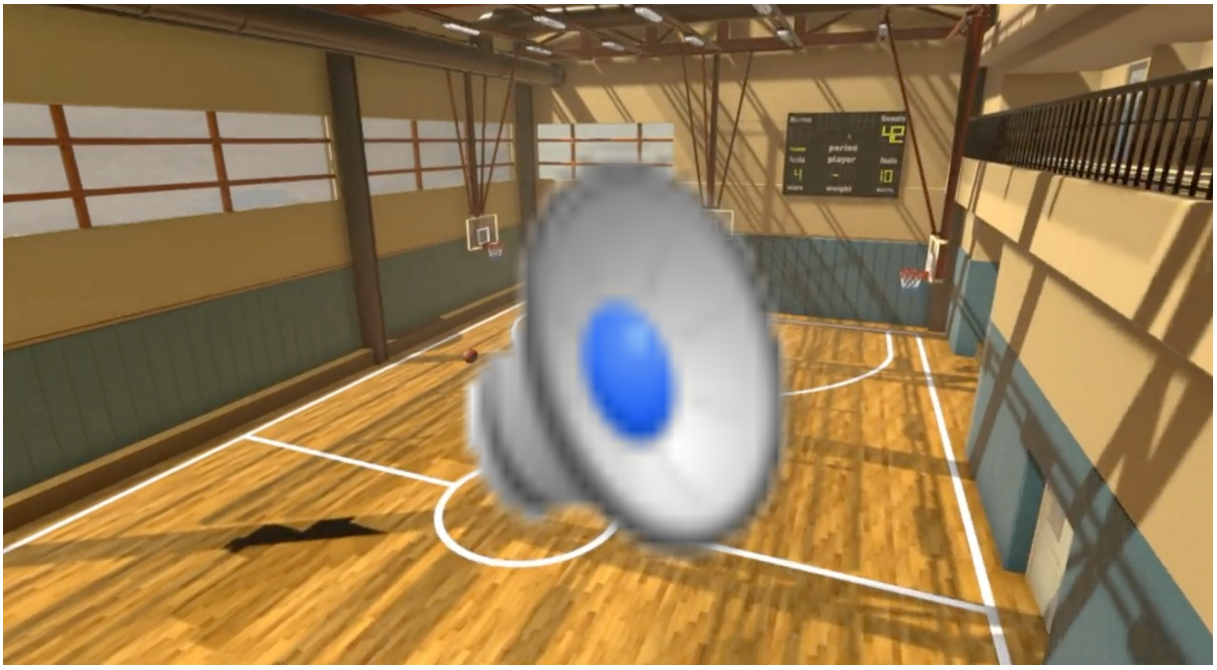
"Basket Ball" (VR)

Fountain Music VR (VR)

VR Test Scene: 'Downtown Bus'

Reflections, Moving Sources and Occluders

VR Test Scene: ‘Virtual Basket Ball’ – User Interaction



AR Test Scene: ‘AR Portal’

Coupling of Acoustic Spaces, Occlusion/Diffraction etc.

MPEG-I IMMERSIVE AUDIO

END OF PART 1 (“THE PROJECT”)

→ NOW ON TO QUALITY EVALUATION

Workshop AES NYC 2022

Part2: Quality Evaluation for Virtual and Augmented Reality

Thomas Sporer

Fraunhofer IDMT

Fraunhofer IDMT

1

Testing of Audio Quality in MPEG before MPEG-I

- **Recommendation ITU-R BS.1116:**
  Triple Stimulus with hidden reference
  - Near instantaneous switching between the three stimuli
  - Impairment Scale from 1.0 to 5.0
  - Comparison with an open reference (shall be scored as 5.0)
  - No direct comparison between different systems
- **Recommendation ITU-R BS.1534:**
  Multi-Stimulus with hidden reference and anchors (MUSHRA)
  - Near instantaneous switching between all stimuli
  - Quality Scale from 0 to 100
  - Comparison with an open reference (shall be scored at 100)
  - Direct comparison of all stimuli with each other (ranking)

Fraunhofer IDMT

2

Challenges in MPEG-I Testing

- Many scenes and many proposals
  - Huge number of comparisons
- Content is produced
  - **No reference** available
- Listener is free to move everywhere
  - No repeatable tracks
  - Pre-recorded tracks would prevent 6DoF interactivity and immersion
- Room acoustics in scenes specified as parameters
  - Rendered stimuli depend on what happened before
- Near instantaneous switching is very demanding
  - Loudness calibration
  - Computational load – all renderers in a test must run in parallel

Fraunhofer IDMT

3

Approaches for reference-less testing in ITU-R

Recommendation ITU-R BS.1284 (2019):
"General methods for subjective assessment of sound quality"

- Comparison of pairs of stimuli
  - 7-point comparison scale
  - No reference given
  - Direct comparison of all stimuli with each other (ranking)

Recommendation ITU-R BS.2132 (2019):

- "MUSHRA without a reference and without anchors"
- "Attribute Ratings"
  - One attribute at a time
  - Typical attributes: "scene depth", "envelopment", "localization accuracy", "brightness", "distortion"

Fraunhofer IDMT

4

Candidates in MPEG-I – AB-Testing

AB-Testing

- Direct comparison of **two** stimuli with each other
- Counting how often renderer X was preferred to renderer Y
  - Comparison reduced to a forced choice
- Thurstone V to boot strap a scale from paired comparison matrix
  - Problem: Comparison scale contains zero
    - Solution: Ties are counted as 0.5 for both renders
- Boot strapped scale used to compare renderers
- Incomplete balanced block design to reduce test time

Fraunhofer IDMT

5

Candidates in MPEG-I - MuSCR

MuSCR – Multi Stimulus Category Rating

- Direct comparison of **more than two** stimuli with each other
- Statistics to compare renderers:
  - Average/standard deviation - ANOVA
  - Median and boxplots
- Limited number of stimuli in parallel to simplify task for listener and test environment (computational load)
- Incomplete balanced block design to reduce test time

Fraunhofer IDMT

6

Candidates in MPEG-I

- MAACR - Multi Attribute Absolute Category rating
  - **One stimulus presented at a time**
  - No direct comparison of stimuli
  - **Four** attributes scored in parallel:

  "basic audio quality", "plausibility", "externalization", "consistency"
  - Scale from 0 to 3
  - Statistics based on a "rejection criterion" and counting occurrence:
    - A renderer with too many "0" or "1" is rejected
    - Different ways to combine four dimensions to a figure of merit
- Incomplete balanced block design to reduce test time

Fraunhofer IDMT

7

Advantages and Disadvantages

- AB-Testing
  - Easy task for listeners
  - Moderate computational complexity (two stimuli)
  - No absolute quality value
- MuSCR
  - Known task for listeners
  - High computational complexity (several stimuli)
  - Scale has meaning to listeners (but this is site dependent)
- MAACR
  - Complicated task for listeners
  - Lowest computational complexity
  - Multi-dimensional criterion for rejection
  - Unclear total figure of merit

Fraunhofer IDMT

8

The MPEG-I CfP Listening Tests

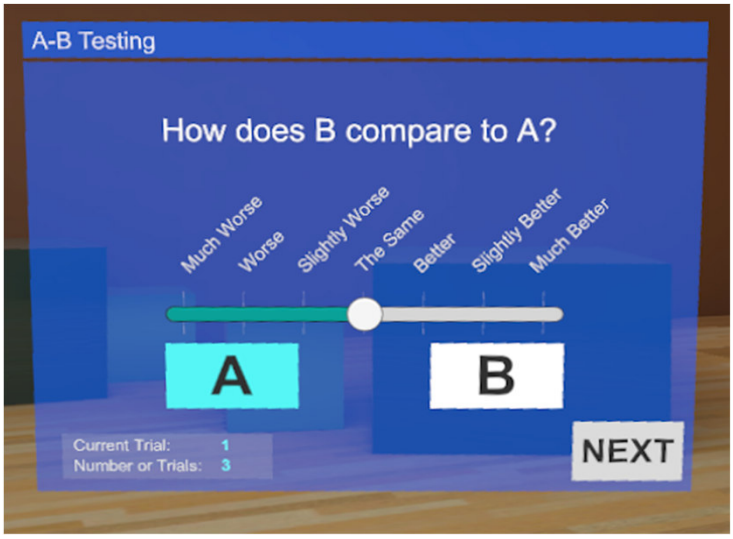
- AB-Test was selected for CfP testing
- Visualization of the user interface (for scoring on-demand)
- "Scene-Tasks": Instructions for listeners what to focus on for each scene
- Test 1: VR had 14 scenes and 14 renderers
- 91 paired comparisons for each scene – total of 1274 scores
- Each listener listened to all 14 renderers and all 14 scenes, but not to all combinations
- Observation from test: some combinations of renderers where too complex for the platform

- MuSCR and MAACR will be used in **Core Experiments (CE)** and **Verification Tests**

Fraunhofer IDMT

9

AB Testing – Test Panel overlayed to scene



Fraunhofer IDMT

10

Challenges in MPEG-I

- Many scenes and many proposals
  - Huge number of comparisons → incomplete block design
- Content is produced
  - **No reference** available → AB testing
- Listener is free to move everywhere
  - No repeatable tracks
  - Pre-recorded tracks would impair immersion → no bug, feature
- Room acoustics in scenes specified as parameters
  - Rendered stimuli depend on what happened before → parallel rendering
- Near instantaneous switching is very demanding
  - Loudness calibration → calibrated at reference point
  - Computational load → AB testing with only two renderers

Fraunhofer IDMT

11

MPEG-I Immersive Audio:

Part 2 - Where Do We Stand Now?

Prof. Dr.-Ing. Jürgen Herre

International Audio Laboratories Erlangen

Erlangen, Germany

Standardization Process
The "Hot Phase"

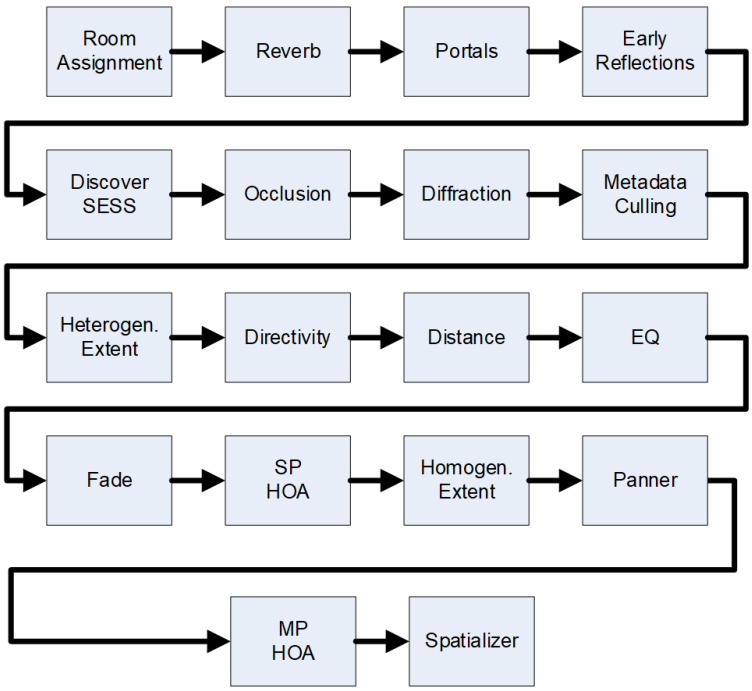
- "Call for Proposals" (CfP) issued in April 2021
- 8 technology proposals submitted on November 10, 2021
- Competitive evaluation by large-scale subjective testing
  (VR & AR tests with headphone reproduction, 12 test sites worldwide)
- Selection of best performing baseline technology in January 2022:
  - Winner is joint submission of Fraunhofer IIS, Nokia & Ericsson
  - Some low bitrate winning technology ('category winner') from Dolby, Philips, Qualcomm

  → 'Reference Model' (baseline for all further technical development)

Standardization Process
The First Reference Model

Core Part of RM:

"Rendering Pipeline"
with subsequent stages



Summary & Outlook

- First well-performing & feature-rich reference model
- Improvement work on some missing aspects until FDIS in 2023, e.g.:
  - Loudspeaker rendering
  - Client-server based streaming operation with a back-channel
  - "Social VR" (incl. real-time communication aspects)
  - …

Ultimately, the work item establishes a ***first long-time stable format*** for ***compressed representation of audio for 6DoF VR / AR content*** based on ***MPEG-H 3D Audio*** that can be used for consumer applications like broadcasting, streaming, social VR by 2023 …

A Final Word: Acknowledgements

The presented technical work is the result of a large-scale effort of the teams at Ericsson, Fraunhofer IIS / International Audio Laboratories Erlangen and Nokia.

Additional technology contributions come from the teams at Dolby Laboratories, Philips, Qualcomm and other MPEG Audio participants

Special acknowledgement goes to Dr. Schuyler Quackenbush for his diligent leadership of the standardization process and to the entire ISO/MPEG Audio group (ISO/IEC JTC1/SC29/WG6).

Thank You Very Much
For Your Attention!

Any Questions?
Time for Q&A

© AudioLabs, 2022
J. Herre

"What is going on at MPEG Audio?"
AES Workshop 10-2022

AUDIO LABS