“Immersive Audio with MPEG-H and MPEG-I”

Jürgen Herre

International Audio Laboratories Erlangen

Erlangen, Germany

Overview

- During the recent years, Immersive Audio has become both technologically mature and available
- Long way from stereo to surround to 3D audio, both loudspeaker and headphone reproduction are important

  - Part 1: MPEG-H Audio - A Brief Overview
    Extremely versatile codec for next-generation audio (NGA) systems

  - Part 2: MPEG-I Audio
    Immersive Audio for VR/AR in 3DoF and 6DoF
    - Based on MPEG-H
    - Standardization for 6DoF currently ongoing
    - Requirements, Architecture, Ongoing Process

AUDIO LABS

Part 1:

"MPEG-H 3D Audio Coding – A Brief Overview"

Recent Standard for Universal Spatial / 3D Audio Coding

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

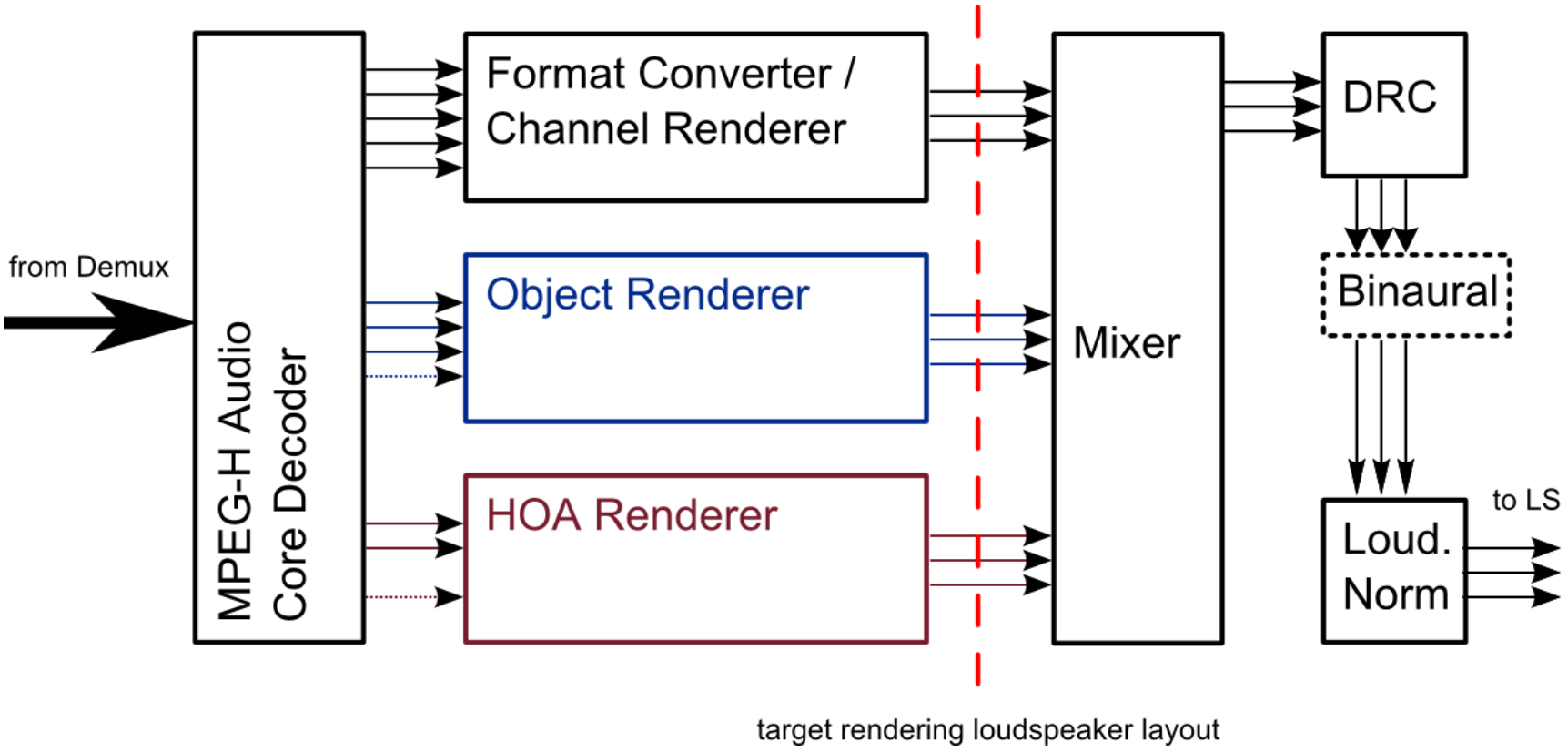
Fraunhofer IIS

Goal: Faithful Reproduction of 3D Spatial Audio

- 3D Loudspeaker setups incl. ‘height’ / ‘lower’ speakers
  - Examples: 5.1+2 … 7.1+4 … 22.2
  - ‘3D’ clearly increases spatial realism / envelopment
  - Plethora of loudspeaker setups, lack of compatibility
- Different production / representation paradigms:
  - Channels – *tied to a specific loudspeaker layout*
  - Objects (waveforms + metadata) – *loudspeaker layout agnostic*
  - Ambisonics … HOA – *loudspeaker layout agnostic*
- Efficient representation / distribution of 3D audio content? Compression needed (e.g. for wireless applications)

The MPEG-H 3D Audio Decoder Model

Decoder Architecture



MPEG-H Audio Summary

- MPEG standard for ***universal*** and ***efficient*** representation of immersive / 3D audio content
  - Highly flexible in input paradigm (channels, objects, HOA)
  - Highly flexible w.r.t. output / rendering (22.2 … stereo / binaural)
  - Arbitrary combinations of channels, objects, HOA possible
  - Interactivity, personal sound experience
  - Low-complexity profile for broadcast applications
  - Baseline profile for generic 3D applications (no HOA)
- Became ‘International Standard’ in February 2015/2017
- Very comprehensive standard for coded representation of 3D Audio; deployed since 2017 (e.g. South Korea)

Part 2:

“MPEG-I Immersive Audio for”

Ongoing Standardization on Audio for Virtual Reality (VR) and Augmented Reality (AR)

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

MPEG-I Audio
Future 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***

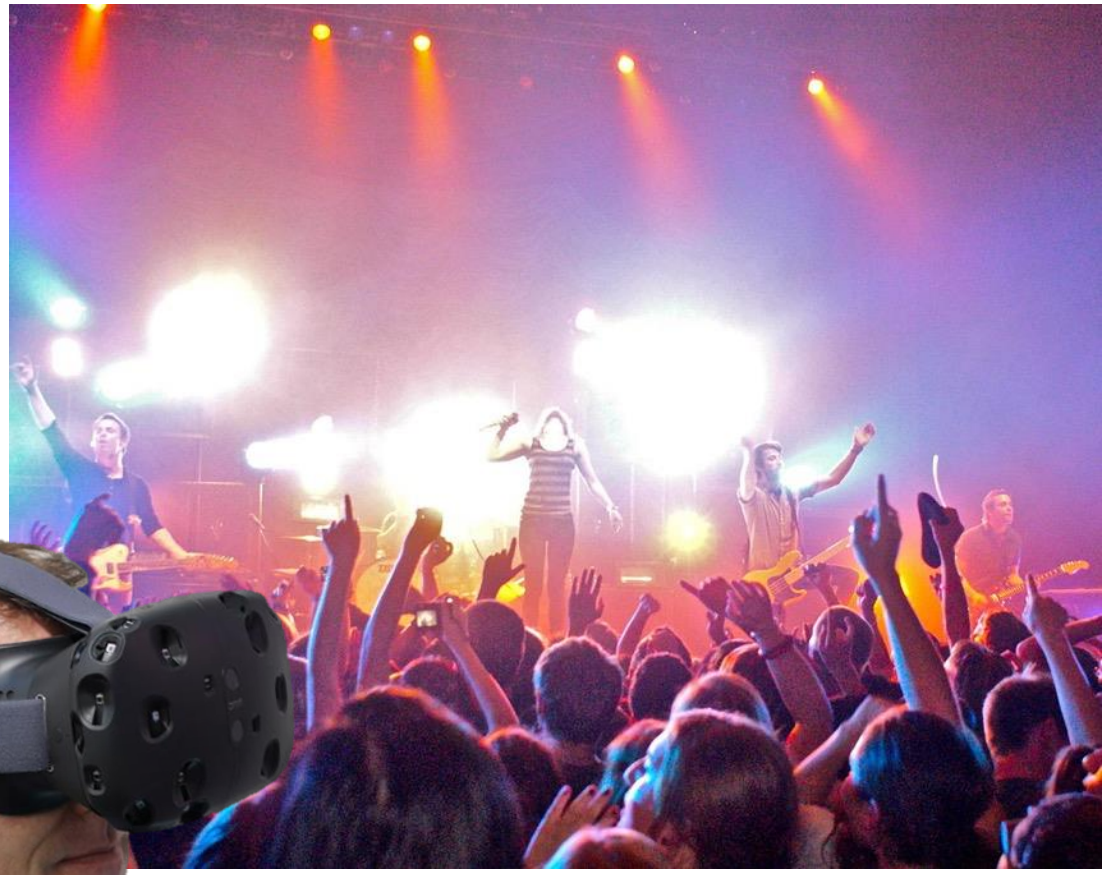
Ongoing Work Item: MPEG-I 6DoF Audio
Some Requirements

- Intended 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

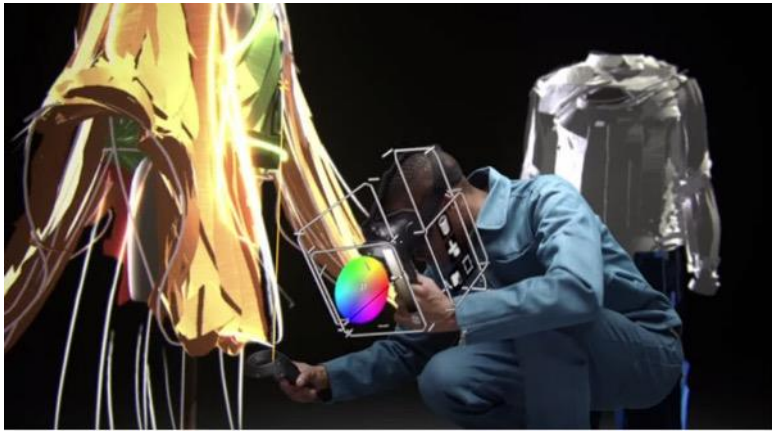


Experience a virtual concert in 6-DoF and move through the venue

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 will comprise

- 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

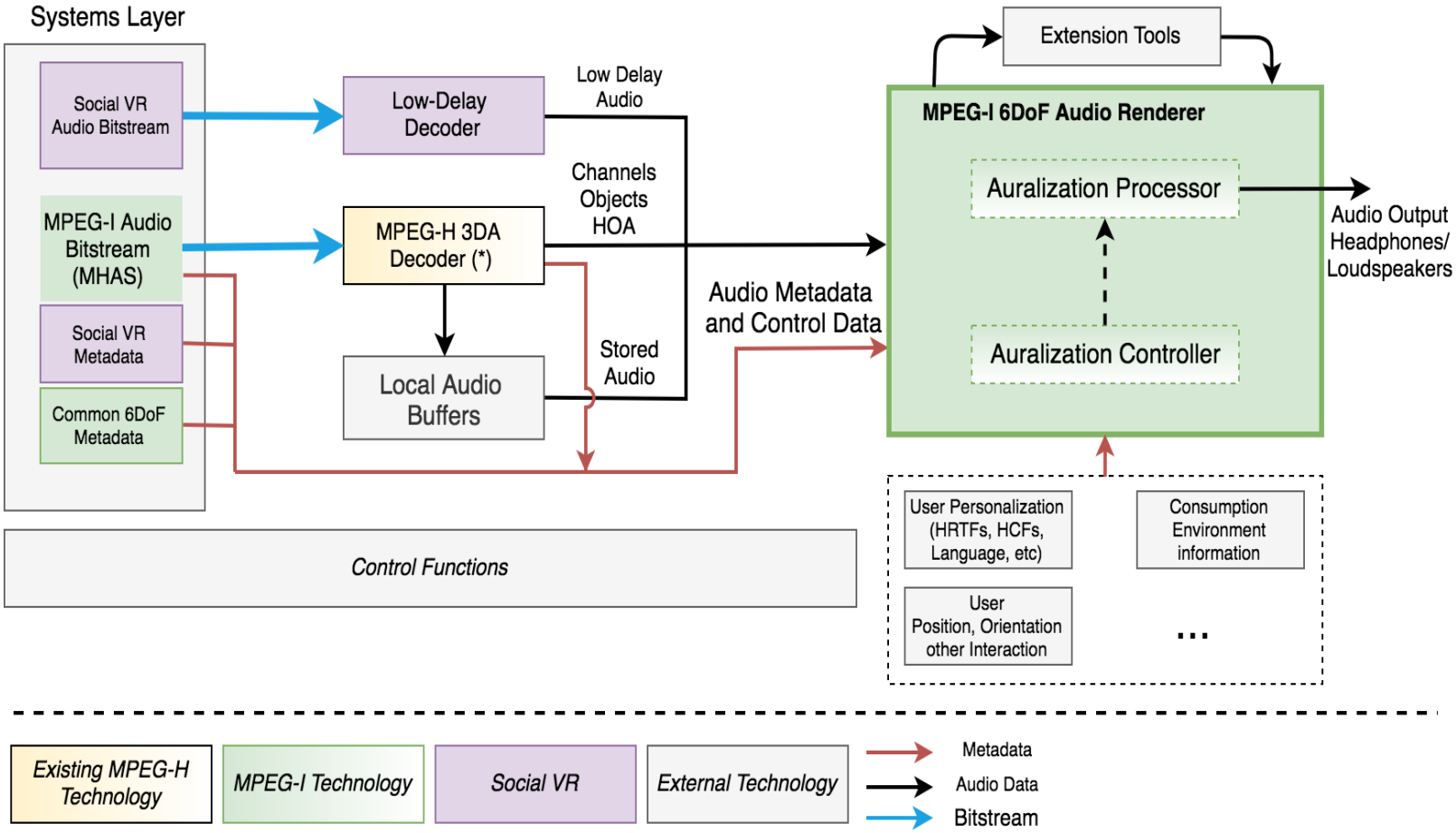
Basic decisions:

- 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

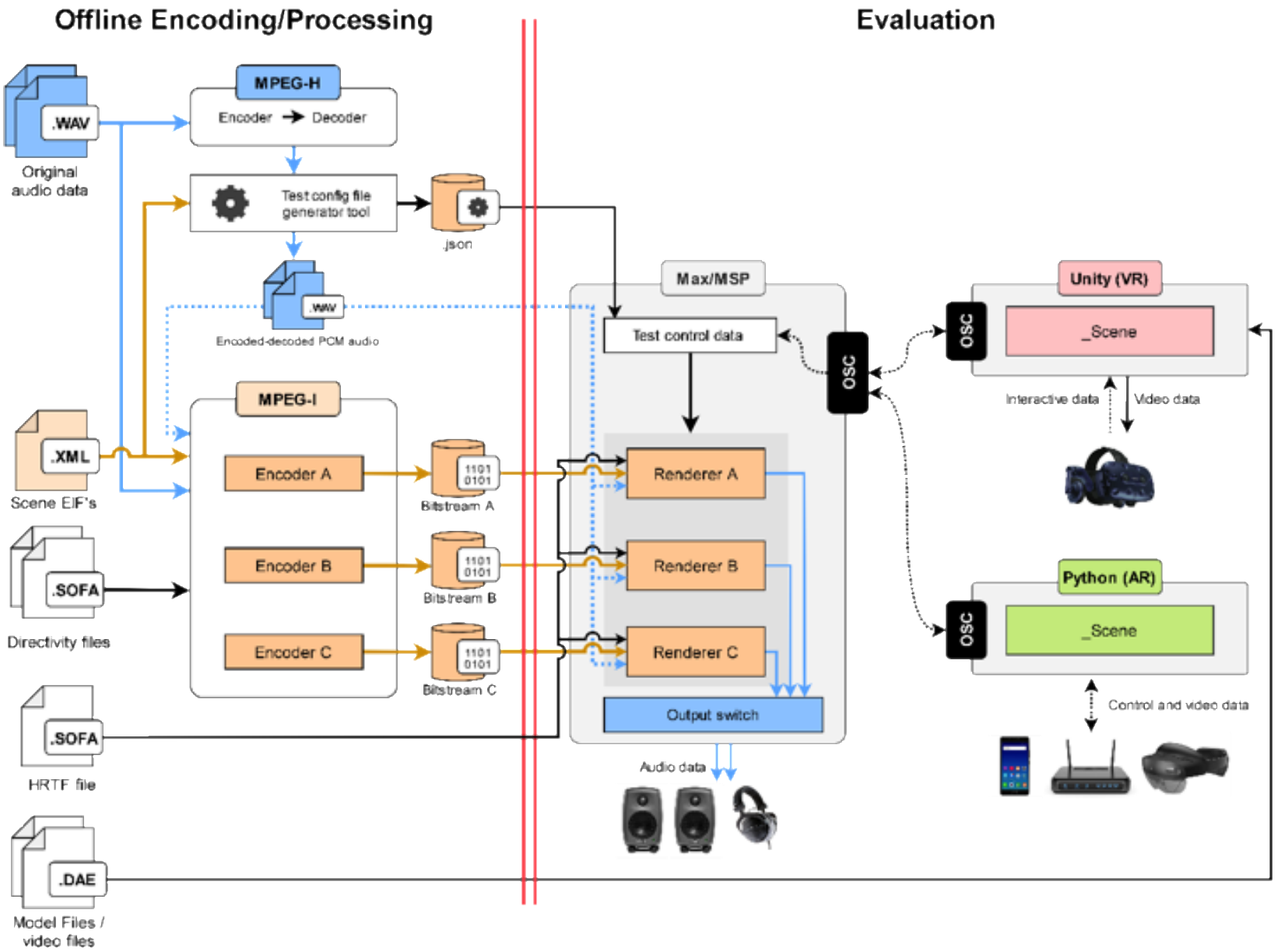
Evaluation Platform:

- 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



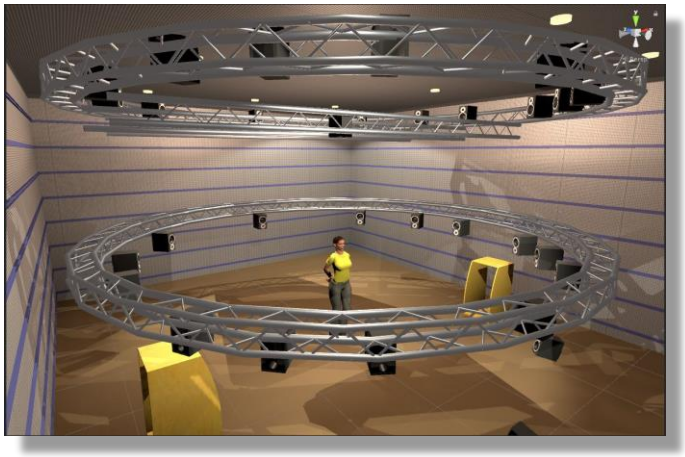
Example: Audio Object

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

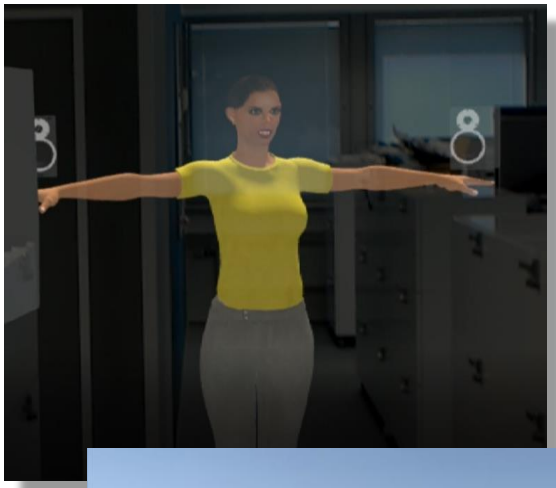
```
<AudioScene>
  <AudioStream id="signal:trumpet"
               file="armstrong.wav"
               mode="continuous" />
  <SourceDirectivity id="dir:trumpet"
                     file="trumpet.sofa" />
  <ObjectSource id="src:trumpet"
                position="2 1.7 -1.25"
                orientation="30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



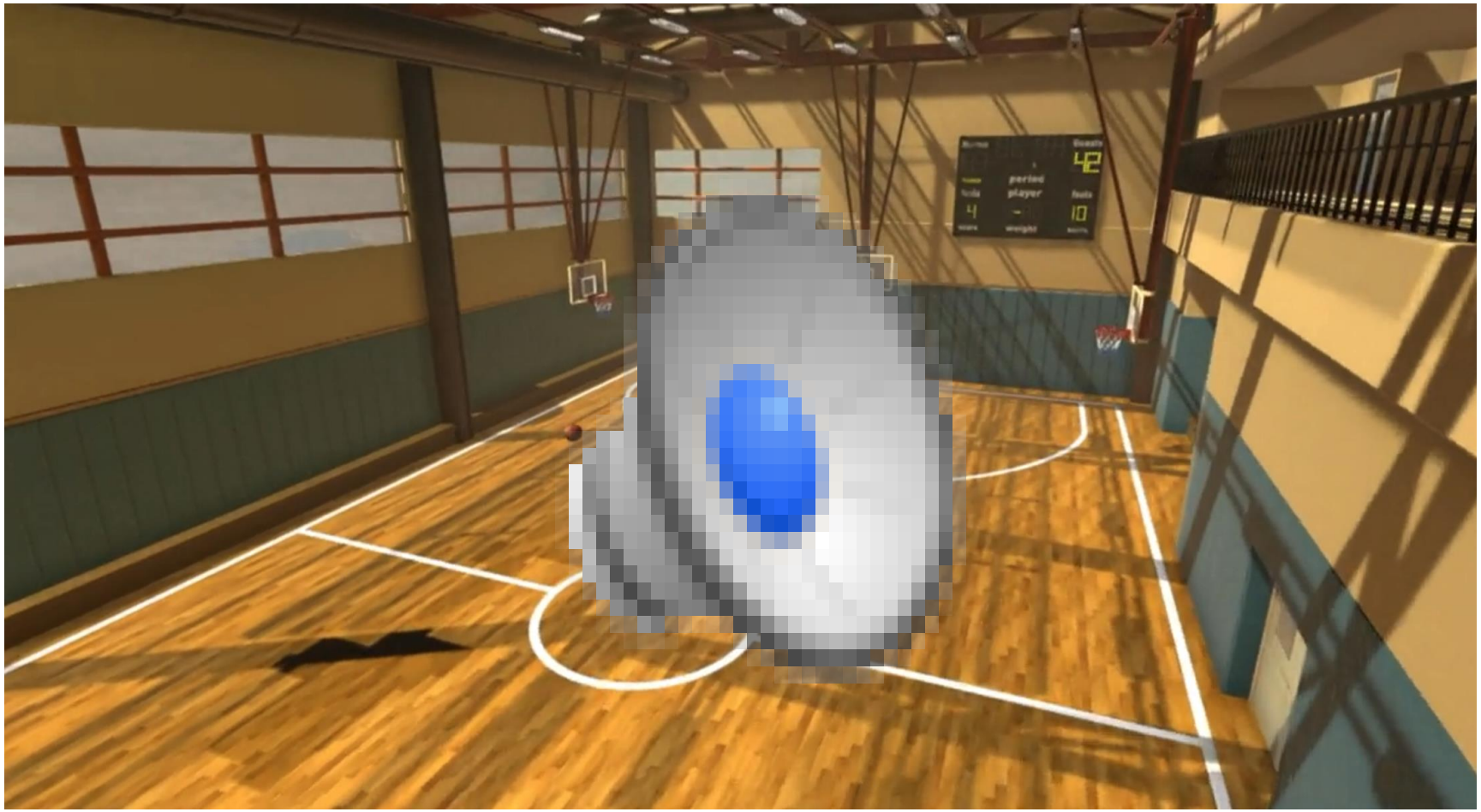
“Basket Ball” (VR)

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’

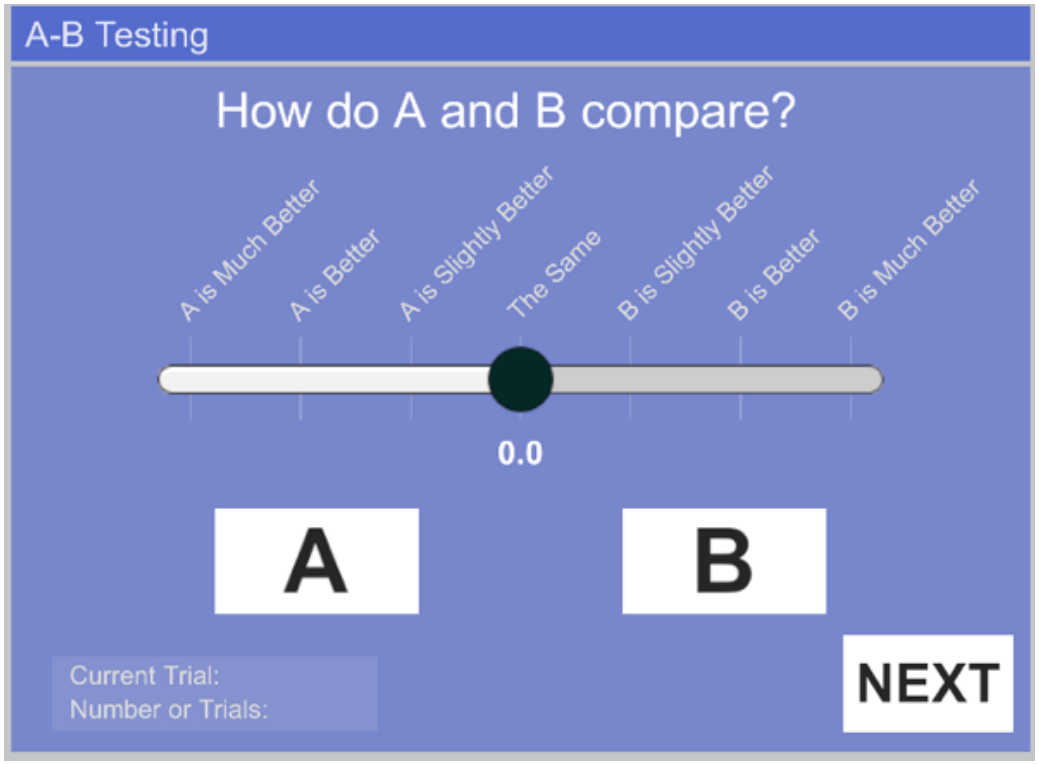
Coupling of Acoustic Spaces, Occlusion/Diffraction etc.

Subjective Quality Assessment:
Extrapolating Established Methods

Virtual MUSHRA-style panel pops up in test scene (“MUSHRA-VR”)

Subjective Quality Assessment (2)

- Chosen for first test round: *A/B Comparison Test* methodology
- Requires only 2 renderers running in real time simultaneously



Current Snapshot of MPEG-I 6DoF Audio
The “Hot Phase”

- 8 technology proposals submitted on November 10, 2021
- Competitive evaluation by large-scale subjective testing
  (VR & AR with headphone reproduction, 12 test sites worldwide)
- Selection of baseline technology based on test results in January 2021,
  then subsequent improvement until FDIS in 2023

Ultimately, the work item will establish 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 …

Thank You Very Much For Your Kind Attention!

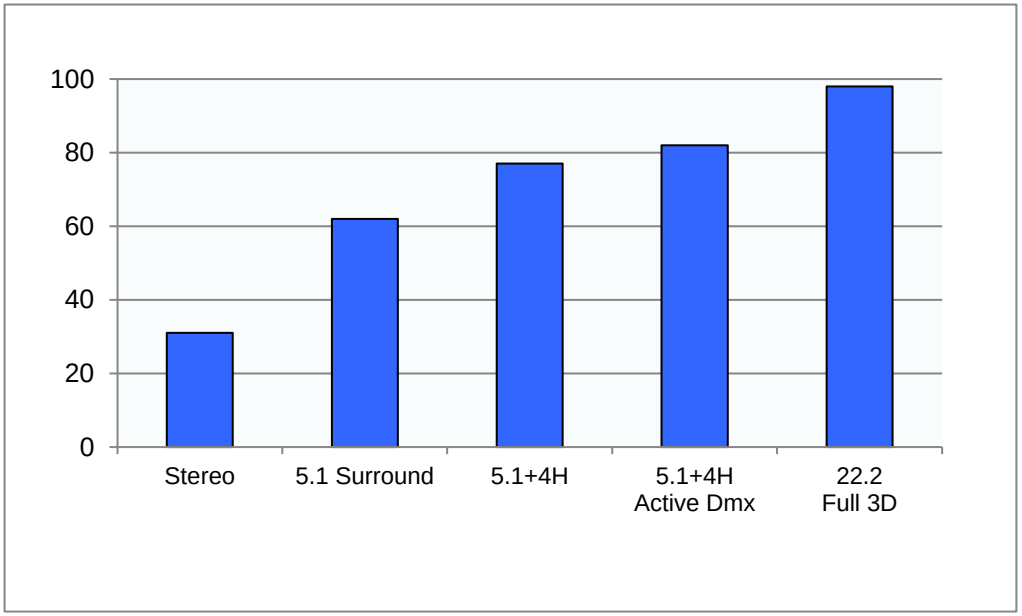
Background: MPEG

- 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)

- MPEG-H Audio has been developed as an extremely versatile codec for next-generation audio (NGA) systems

MPEG-H 3D Audio Philosophy
Embracing All Production Paradigms

- Channel-based input (C)
  - Traditional approach: Transmit signals for loudspeakers at precisely specified locations relative to listener (e.g. 2.0, 5.1, … 22.2 …)
  - Clear improvement from stereo to surround and '3D':



[Silzle et al. 2011]

(BAQ when 22.2 is the reference)

MPEG-H 3D Audio Philosophy
Embracing All Production Paradigms (2)

- Object-based input (O)
  - Increasingly popular: Transmit ‘object’ signals to be rendered at target locations specified by associated metadata
  - Time-varying target locations (e.g. plane fly-over)
  - Personalized/interactive experience (e.g. adjust object characteristics)
  - *Speaker layout agnostic*, rendering to target setup

- Higher Order Ambisonics (HOA)
  - Transmit ‘coefficient’ signals corresponding to a spherical expansion of the sound field in a point. No direct relation to C or O.
  - *Speaker layout agnostic*, rendering to target setup

MPEG-I 6DoF Audio
Relation to MPEG-H 3D Audio

MPEG-I 6DoF Audio …

- extrapolates MPEG-H technology into the VR/AR world
- will accept MPEG-H content for use in VR/AR applications (➔ content authoring)
- will be able to decode/render MPEG-H content
- content can be fed back into MPEG-H decoders

Some Literature On Subjective Testing

- T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets: "Online vs. Offline Multiple Stimulus Audio Quality Evaluation for Virtual Reality", 145th AES Convention, Paper 10131, New York 2018
- T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets: "Evaluation of Binaural Renderers in Virtual Reality Environments: Platform and Examples", 145th AES Convention, e-Brief 454, New York 2018
- O. Rummukainen, T. Robotham, S. Schlecht, A. Plinge, J. Herre, and E.A.P. Habets: "Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking", Proc. of the Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA, August 2018 (Best Peer-Reviewed Paper Award)