1. **Problem statement**

Mission: Identify key elements and attributes of an immersive multi-channel audio format that would enable the end user to experience a navigable audio scene, i.e. choose their location and orientation at rendering time.

Do any existing audio content formats meet this need? If not, what are the gaps?

**Cinematic vs. interactive**

This study is concerned with pre-determined “cinematic” experiences (i.e. recordings) as opposed to interactive experiences that are computer generated at playback time.

Examples of cinematic audio: music, soundtracks, cinematic VR.

Examples of interactive audio: games, simulation, interactive VR, AR/MR. These experiences readily allow for
6DOF: both position and orientation of the user.

**6DOF vs 3DOF**

Current solutions for cinematic experiences (both professional installation/theater and consumer) describe the immersive audio scene at one point. In this document, we refer to such solution as "3DOF" (only allowing rotations of the user, looking around while staying at a pre-determined position in the virtual scene. Current cinematic VR delivery formats fall in this category.

In contrast, a solution allowing translation of the user in three dimensions, in addition to the 3 axes of rotation, is qualified as "6DOF" (i.e. supporting six degrees of freedom).

**Functional block diagram**

The end-to-end solution involves three high-level functions

- **Capture**: collect/analyze/combine raw audio elements of the content and encode them into a content format.

- **Transmit/store**: all audio and data necessary for a renderer to reproduce the scene in a file format that can be used for content interchange and consumer delivery
  - It is customary to provide a different format for media interchange (professional uncompressed format) vs. consumer delivery (low bit rate compression, streamable).

- **Render**: generate audio signals feeding the reproduction transducers, taking as input
  - incoming content provided in the above format
  - user's real-time position and orientation within the scene
  - playback configuration (headphones, arbitrary array of loudspeakers)
    - note that format is agnostic to the render device or playback transducer configuration.

![Functional block diagram](image)

**2. Review of Use Scenarios**

**Usage Scenarios**
The new format enables immersive and interactive audio experiences.
At least three categories have been identified

- Live
- Recorded
- Blended

Examples

**Sporting Event**

The end user could be a “drone” that can change their experience to any location on, above, or around the field. The user or broadcaster can change the mix to include or exclude the announcer team, the players, the referees, the attendees, etc.

**Holodeck**

The end user can explore and move around in virtual or augmented space which may include a blend of captured and synthetic elements.

**Industrial AR**

Some Industrial usages may include

- A factory worker has the ability to be guided to a specific part in a warehouse using visual and audio cues
- Audible warning systems can alert to dangers coming from specific locations around or behind the person
- 6DOF environment models can be used to study/improve efficiencies and provide training
- Another benefit might include the ability to run through a fire evacuation scenario or some other health or safety application

**Multiple POV movies**

The end user can enjoy a 6DOF movie where they can move around within the scene to choose their experience point.

- Note that this cedes control from the director, so there may be limitations that would be placed on where the user can move to

**User Generated Content**

The end user can recreate a living memory and explore it from different angles so that they can build a more emotional and realistic connection with the content. This may be most applicable for celebrations and social gatherings.

3. Review of relevant existing formats and gaps therein

In order to achieve the 6DOF user scenarios mentioned in the previous section, various data needs to be represented in a format for rendering a realistic, immersive audio scene. Not all information is equally important for rendering, but the lack of information can be limiting for more realistic renderers, so it is deemed that more information is better.

A set of key features were identified for rendering a 6DOF scene independent of pre-existing spatial audio formats to prevent the discovery of required features being overly influenced by prior art.
A format for a premium 6DOF audio experience should include the following features/elements:

- one or more position-related beds, such as Ambisonic multi-channel signals
- one or more objects, each functioning as a perceived sound source / emitter having:
  - mono waveform (clean signal) and sample rate, signal format etc...
  - position
  - orientation
  - frequency dependent directivity
  - intensity/gain
  - size (i.e. spatial extent)
  - diegetic or non-diegetic state (non-diegetic => head-related, optional)
  - rendering priority (optional)
- one or more content-aware environments, each having (note: this information could be shared with or inherited from other included scene description elements):
  - geometry
  - information necessary to render acoustic obstruction effects
  - absorption rate
  - reverberation properties (reverb time)

To avoid the difficult task of inventing a new format, pre-existing formats were evaluated to understand the current capabilities and gaps in comparison to the list of attributes for a new 6DOF format. The ideal situation would be to build on an existing format that contains most or all of the elements needed to support a 6DOF experience.

<table>
<thead>
<tr>
<th>Feature/Element</th>
<th>MPEG-4 v2/v3</th>
<th>AABIFS</th>
<th>MPEG-H ADM</th>
<th>Dolby Atmos (w VR extension)</th>
<th>DTS-X, MDA</th>
<th>Auro VR</th>
<th>Google VR</th>
<th>Facebook VR</th>
</tr>
</thead>
<tbody>
<tr>
<td>one or more position-related beds, such as Ambisonic signals</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>only 1 bed?</td>
<td>1 bed</td>
<td>1 bed</td>
</tr>
<tr>
<td>one or more objects as a perceived sound emitter, each having:</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>waveform (clean signal) and sample rate, signal format etc...</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>position</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>orientation</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>?</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>directivity (vs frequency)</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>?</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>intensity/gain</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>size (spatial extent)</td>
<td>Y</td>
<td>N</td>
<td>?</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>object priority (in media format) **</td>
<td>Y</td>
<td>Y</td>
<td>?</td>
<td>Y</td>
<td>?</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>non-diegetic components</td>
<td>?</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>?</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>one or more environments /</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
### Notes on requirements and assumptions

**About the format**

- Must include the definition of the coordinate system, necessary for expressing the listener position and orientation to the renderer which needs to be aware of the reference position/orientation, and distance units.
- Streamable bitstream
- Should we limit the scope to representing diegetic scenes?
- Should we limit the scope to a single acoustic space/room?
- Should the format restrict where the user is allowed to go? Continuous or discrete path?
- Should the format include restrictions on object manipulations?

**About the rendering engine: functionality requirements**

- Positional audio rendering (over headphones or flexible loudspeaker configuration)  
  —> sounds come from where you see them
- Source directivity and orientation
- Obstruction, occlusion effects
- Multi-room artificial reverberation, dynamic early reflections  
  —> recordings convey environment acoustics
- Ambisonic decoding

### 4. Conclusions, recommendations, next steps

**Summary:**

- This seems to be a problem worth solving.
- Live recording of 6DOF audio material is a technically challenging problem, requiring technology not yet on the market.
- An initial recommended approach is to record multiple location-tagged scenes (e.g. Ambisonic microphone array recordings), sampling the sound field at multiple possible listening positions.
- MPEG-4 Audio BIFS seems to include all the key necessary elements, but did not explicitly consider capturing multiple location-tagged audio scenes.
- The more recently developed object-based formats, such as ADM, MPEG-H, Dolby Atmos or DTS-X, could be extended to support the necessary additional features, specifically
  - Objects as sound sources/emitters (having directivity and orientation)
  - Room/environment characteristics
Recommendations, action items:

- Research methods for object-based far-field recording (example: SALSA Sound)
- Research the following problems:
  - Interpolating between several Ambisonic recordings
  - Converting Ambisonic recordings to object-based representations
- Engage with immersive audio standardization workgroups to consider extending recently developed formats, such as ADM.
- Agree on terminology to refer to real-time computer generated experiences such as games, which are different from what is addressed here. ("Interactive" is confusing.)

5. References

http://resource.isvr.soton.ac.uk/FDAG/VAP/index.htm
https://en.wikipedia.org/wiki/Ambisonics
https://en.wikipedia.org/wiki/Dolby_Atmos
https://en.wikipedia.org/wiki/MPEG-H_3D_Audio
https://en.wikipedia.org/wiki/Auro-3D
https://dysonics.com/
Bruel & Kjaer type 4100 HATS (Head And Torso Simulator)
MPEG-4 Audio BIFS
ITU BS.2076: Audio Definition Model (ADM)
SALSA Sound: Spatially Automated Live Sports Audio

Copyright 2000-2017, Fat Labs, Inc., ALL RIGHTS RESERVED
www.projectbarbq.com