3GPP TSG SA4 110-e ***S4-201195***

Electronic, 19th–28th August 2020

**Agenda item** 11.5 ITT4RT

**Source:** KPN N.V.

**Title:** Spatial Audio related Metadata in ITT4RT

**Document for** Discussion & Agreement

# 1 Introduction

In ITT4RT we identified several requirements for audio (conversation)[1], e.g.:

* Support of “two-way audio”
* “remote single user wearing an HMD participates to a conference … receives stereo or immersive voice/audio and immersive video captured by an omnidirectional camera”

In this contribution we address two cases related to audio metadata that can address and improve the above in order to allow spatial rendering of audio data. These two cases were also identified in the study [2] {5.7.3(5)} in the following way:

1. “Spatial Audio orientation of a speaker”
2. “360 image metadata for associated audio”, i.e. in this contribution we focus on the audio properties of the (360) room.

# 2 Spatial Audio orientation of a speaker (talker)

This section outlines 2 use cases to explains the concept of directional rendering of a remote talker (remote user in VR).

**2.1 Use Case – Remote VR user with directional audio in conference room**

Our general use case for ITT4RT assumes a conference room with a 360-degree camera and that allows a remote user to join the conference room in VR. Let’s further assume a central audio speaker system with multiple loudspeakers attached to the 360-degree camera in the center of the room. Knowing the position of the user (same as the 360-degree camera) we can now simply render directional audio with the rotation properties of the user (Pitch, Yaw, Roll). This is the loudspeaker can direct the audio of the remote user based on the rotation and thus increase the naturalness of conversation between the users in the conference room and the remote VR user. Figure 1 shows this concept.

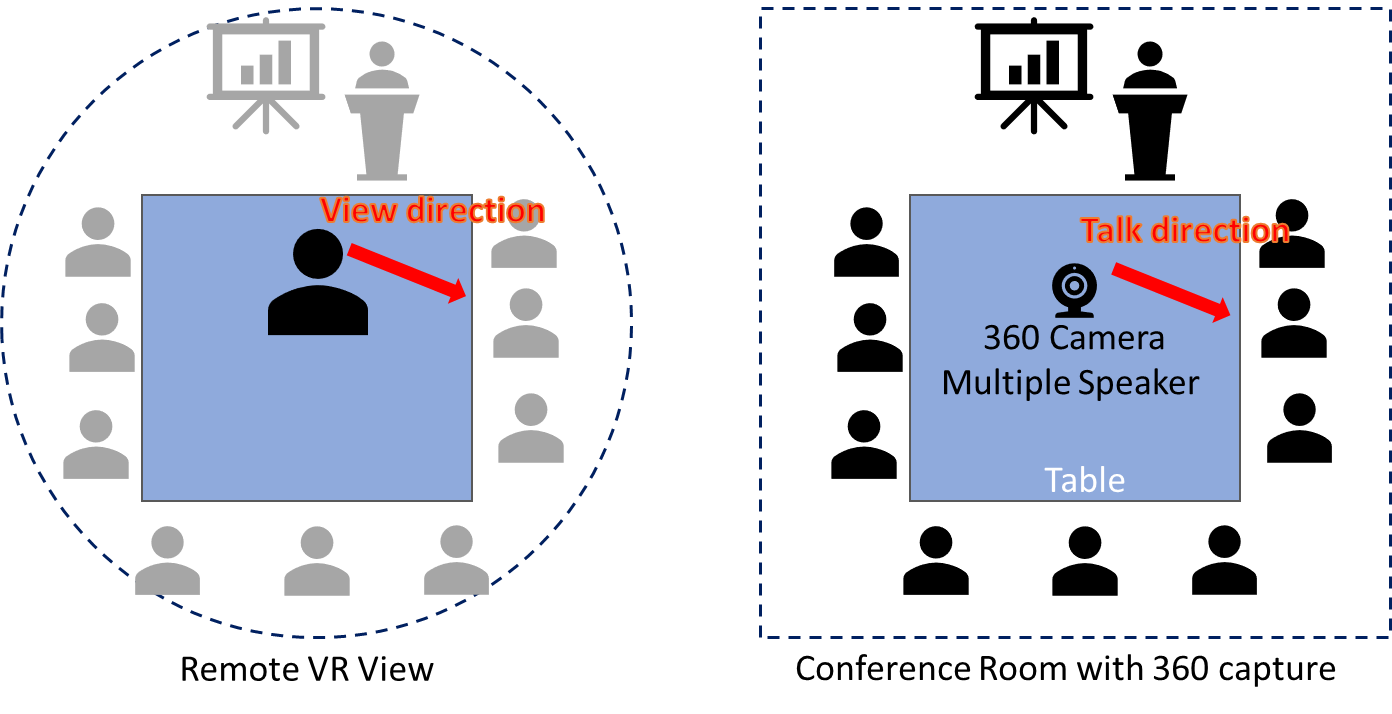
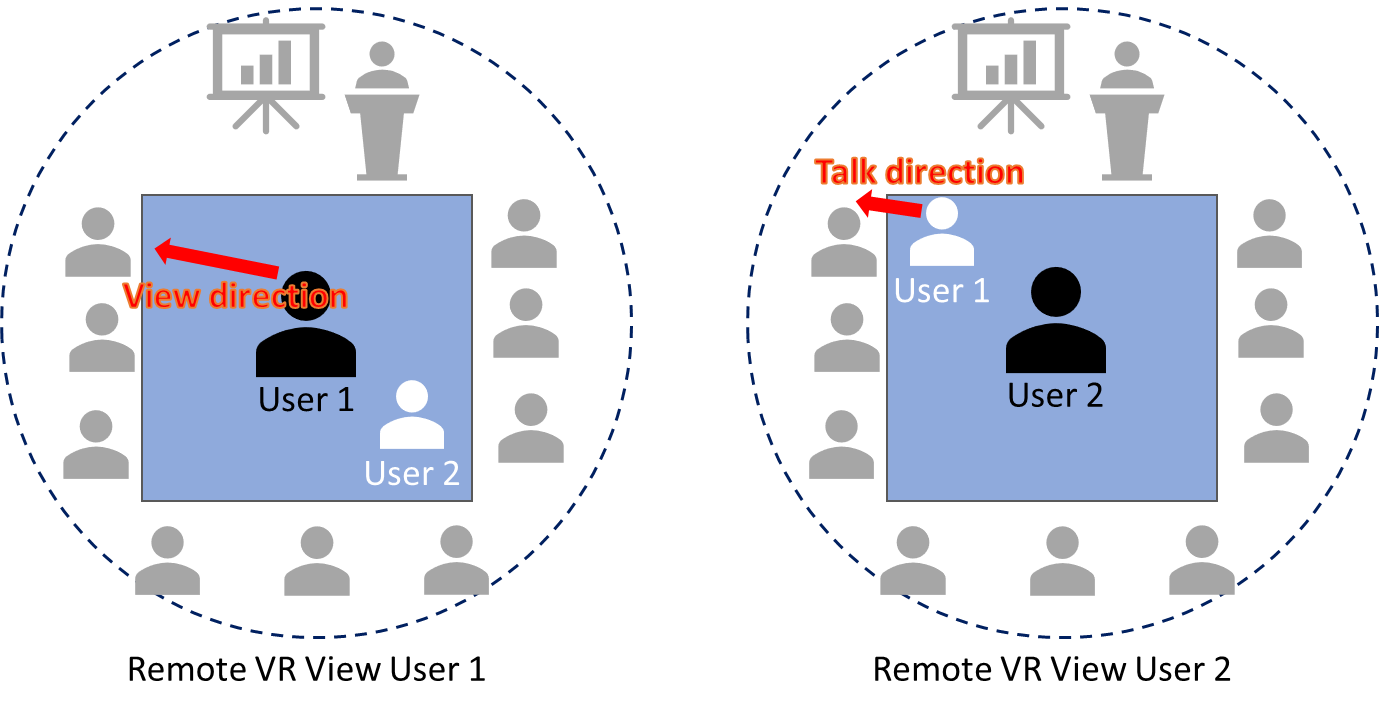


Figure 1: Remote VR user joins conference room while the speech of the remote user is made audible in the conference room cording to his/her viewing direction

**2.2 Use Case – Remote talker in VR**

Adding to use case 1, we can render directional audio between two users in VR. This is, the audio of user 1 (see figure 2) is directed according to his/her view direction and follows the acoustical properties of the room in the virtual environment. This also requires that the virtual audio characteristics of the room (see chapter 3) are also know to the VR rendering engine.



*Figure 2: Two Remote VR users joins a conference room the speech of user 1 is rendered in the virtual environment of user 2 according to the viewing direction of user 1*

**2.3 Solution**

Based on the previous discussion (telco 110-e MTSI 24.08. 11pm CEST) the above use cases are currently not covered with the current (RTCP based) viewport signaling. Thus, allowing the above use cases would require new metadata and corresponding signaling (e.g. over RTCP feedback (FB) message). However, essentially the signaling for the view direction for speech direction would follow the same RTCP feedback procedure as the viewport signaling while including the metadata properties as follows.

The minimal metadata information we foresee to allow the above use cases are:

* Position = (X,Y,Z) = (-0.2,0,0.2) {Placement of the user in the environment}
* Orientation = (Pitch, Yaw, Roll) = (0,270,0) {View / Talk rotation of the user}

# 3 Audio properties of the (360) room

For any VR experience we can assume that the audio of an environment should be consistent with the visuals of an environment, both spatially and temporally.

In ITT4RT this can be challenging as certain audio can be added to the VR environment separately to a 360 video/audio feed. As an example, this could be:

* Audio of individual users of the conference room
* Audio from overlay content
* Audio of remote users

The problem is that the “separate” audio will not sound like it is produced in the original environment (i.e. the conference room captured with a 360 camera): This is there are no proper reflection and there is no proper reverb. One simple way to fix this for ITT4RT is to add metadata describing the audio room properties and to adjust the playout of individual audio to the room properties.

It is currently not clear how this is covered with our current metadata in ITT4RT (i.e. GLTF as room description or OMAF for media description). Further specifying all objects and all characteristics of a room might be too complex for scope of ITT4RT.

**3.1 Possible Solution**

One simple solution is to describe the room as metadata in a simple box model, spatial properties of the room (i.e. dimension, size) and spatial characteristics (indicative for the reverb and audio reflective behaviour). Example of this can be seen in [3], defining the following items (where as in ITT4RT the user could be seen as the centre of this room):

* dimensions
  + width in meters, e.g. width = 4
  + height in meters, e.g. height = 2.5
  + depth in meters, e.g. depth = 5
* materials
  + left as a string, e.g. left = brick-painted
  + right as a string, e.g. right = curtain-heavy
  + front as a string, e.g. front = brick-bare
  + back as a string, e.g. back = glass-thick
  + up as a string, e.g. up = wood-ceiling
  + down as a string, e.g. down = linoleum-on-concrete

This simple box model metadata could be covered and be part of the “SDP configurations for negotiating of immersive video and voice/audio capabilities”

# 4 Conclusion / Proposal

With this contribution we propose the following.

*4.1 For discussion*

**Adding use cases (2.1 and 2.2) to PD:**

* Discuss the relevancy of the proposed use cases
* Clarify process to enable spatial rendering of talker in ITT4RT

**Audio properties of the room:**

* Adding requirements (new bullets in section 4 [1])
  + It is recommended to signal a room configuration for the 360 – degree room when joining an ITT4RT session. This room configuration should include the audio properties of the virtual room as metadata to allow spatial audio rendering of individual audio objects in the virtual scene (e.g. overlays, and user related representations and audio).
* Addressing this requirement as part of the “SDP configurations for negotiating of immersive video and voice/audio capabilities”

# 5 References

1. 3GPP SA4 ITT4RT Permanent Document v0.8.0, <http://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_110-e/Docs/S4-201024.zip>
2. 3GPP TR 26.928: “Extended Reality in 5G”, <https://www.3gpp.org/ftp/Specs/archive/26_series/26.928/>
3. <https://developers.google.com/resonance-audio/develop/overview>