can be used also for generic Audio messages with other purpose than signaling ROIs. One example of such Audio messages is given by Audio messages for conveying information/indication of various options the user has in an interactive AR/VR/MR environment (e.g., "jump over the box to your left for entering room X"). Additionally, the VR example will be used, but the mechanisms described in this document apply to any media consumption environment.

2. Terminology and Definitions

The following terminology is used in the technical field:

- **Audio Elements**: Audio signals that can be represented for example as Audio objects, Audio channels, scene based Audio (Higher Order Ambisonics - HOA), or combination of all.
- **Region-of-Interest** (ROI): One region of the video content (or of the environment displayed or simulated) that is of interest to the user at one moment in time. This can be commonly a region on a sphere for example, or a polygonal selection from a 2D map. The ROI identifies a specific region for a particular purpose, defining the borders of an object under consideration.
- **User position information**: location information (e.g., x, y, z coordinates), orientation information (yow, pitch, roll), direction and speed of movement, etc.
- **Viewport:** Part of the spherical Video that is currently displayed and viewed by the user.
- **Viewpoint**: the center point of the Viewport.
- **360-degree video** (also known as immersive video or spherical video): represents in the context of this document a video content that contains more than one view (i.e., viewport) in one direction at the same moment in time. Such content can be created, for example, using an omnidirectional camera or a collection of cameras. During playback the viewer has control of the viewing direction.
- **Adaptation Sets** contain a media stream or set of media streams. In the simplest case, one Adaptation Set contains all audio and video for the content, but to reduce bandwidth, each stream can be split into a different Adaptation Set. A common case is to have one video Adaptation Set, and multiple audio Adaptation Sets (one for each supported language). Adaptation Sets can also contain subtitles or arbitrary metadata.

- The Earcon metadata can be created in such a way that the Earcon is signaled to be active
- Some devices will know only how to parse the metadata and reproduce the Earcon
- Some newer devices that additionally have a better ROI processor can decide to deactivate it in case is not needed

- More information and an additional figure about the adaptation sets.

Therefore, in a VR/AR environment the user can usually visualize full 360-degree content using for example a Head Mounted Display (HMD) and listen to it over headphones. The user can usually move in the VR/AR space or at least change the viewing direction - the so-called “viewport“ for video. Compared with classic content consumption, for VR the content creators cannot any-longer control what the user visualizes at various moments in time - the current viewport. The user has the freedom to choose different viewports at each instance of time, out of the allowed or available viewports. In order to indicate to the user the Region Of Interest (ROI), audible sounds can be used, either natural or synthetic sounds, by playing them back at the position of the ROI. These audio messages are known as “Earcons“. This invention proposes a solution for efficient delivery of such messages and proposes an optimized receiver behaviour for making use of the Earcons without affecting the user experience and the content consumption. This leads to an increased Quality of Experience. This can be achieved by using dedicated metadata and metadata manipulation mechanisms on systems level for enabling or disabling the Earcons in the final scene.

The metadata processor 132 may be configured to receive and/or process and/or manipulate metadata 141 so as to cause, at the decision that the information message is to be reproduced, the reproduction of the Audio information message according to the metadata 141.Audio signals (e.g., those for representing the scene) may be understood as being part of the audio scene (e.g., an audio scene downloaded from a remote server). Audio signals may be in general semantically meaningful for the audio scene and all audio signals present together construct the audio scene. Audio signals may be encoded together in one audio bitstream. Audio signals may be created by the content creator and/or may be associated to the particular scene and/or may be independent from the ROI.

The audio information message (e.g., earcon) may be understood as not semantically meaningful to the audio scene. It may be understood as an independent sound that can be generated artificially, such as recorded sound, a recorder voice of a person, etc. It can be also device-dependent (a system-sound generated at the press of a button on the remote

- The Audio data of each Stream 116, 316 and the modified Audio Metadata and Earcon Metadata are then provided to a Stream Muxer or multiplexer which can generate based on this, one Audio Stream 414 with one set of Metadata (modified Audio metadata 238 and modified Earcon metadata 234).

- This Stream 414 may be decoded by a single Media Audio Decoder 112 based on the user position information 122.

- Additionally, in a Streaming environment, based on the same information, the ROI Processor 120 can decide to request the Earcon(s) Stream(s) 140 in advance (e.g., when the user looks into the wrong direction with few seconds before the ROI is enabled).

6.6 Example of Fig. 5

Fig. 5 shows a system 500 comprising, at the client side 204, a system 502 (client system) which may embody, for example, the system 100 or 200. Here, reference is made to Earcons, even it is possible to generalize the concept to any Audio information messages.

The system 502 may comprise the ROI processor 120, the metadata processor 132, a Stream multiplexer or muxer 412.

In this example, an Earcon Stream is not provided by a remote entity (at the client side), but is generated by the synthetic Audio generator 246 (which may also have the capability of storing a Stream, for re-using subsequently, or to use a stored compress/uncompressed version of a natural sound). The Earcon metadata 141 are notwithstanding provided by the remote entity, e.g., in an Audio Stream 116 (which is not an Earcon Stream). Therefore, the synthetic Audio generator 246 may be activated to create an Audio Stream 140 on the basis of the attributes of the Earcon metadata 141. For example, the attributes may refer to a type of the synthesized voice (natural sound, synthetic sound, spoken text, and so on) and/or text labels (the Earcon can be generated by creating synthetic sound based on the text in the metadata). In examples, after that the Earcon Stream has been created, the same may be stored for being re-used in the future. Alternatively, the synthetic sound can be a generic sound stored permanently in the device.

A Stream muxer or multiplexer 412 may be used to merge packets of the Audio Stream 116 (and in case also of other Streams, such as the auxiliary Audio Stream 316) with the packets of the Earcon Stream generated by the generator 246. After that, an Audio Stream 414 may be obtained which is associated to modified Audio metadata 238 and modified Earcon

- and, if requested by the ROI Processor 120, modify the spatial position and gain information contained in the Earcon metadata 141 accordingly.

- The Metadata Processor 132 may parse also the Audio Metadata (e.g., 236, 417) of all Audio Streams (116, 316) and manipulate the Audio Specific Information in such a way that the Earcon can be used as part of the Audio scene (e.g., if the Audio scene has a 5.1 channel bed and 4 objects, the Earcon Audio element is added to the scene as the fifth object. All metadata fields are updated accordingly).

- The modified Earcon Metadata and the information from the ROI Processor 120 are provided to the Synthetic Audio Generator 246. The Synthetic Audio Generator 246 may create based on the information received a synthetic sound (e.g., based on the spatial position of the Earcon a voice signal is generated spelling the location). Also, the Earcon Metadata 141 is associated with the generated Audio data into a new Stream 414.

- Similarly, as before, the Audio data of each Stream (116, 316) and the modified Audio Metadata and Earcon Metadata are then provided to a Stream Muxer which can generate based on this one Audio Stream with one set of Metadata (Audio and Earcon).

- This Stream 414 is decoded by a single Media Audio Decoder 112 based on the user position information

- Alternatively or additionally, the Audio Data of the Earcon can be cached in the Client (e.g., from previous Earcon usages)

- Alternatively, the output of Synthetic Audio Generator 246 can be uncompressed Audio and it can be mixed into the final rendered scene

- Additionally, in a Streaming environment, based on the same information, the ROI Processor 120 can decide to request the Earcon(s) Stream(s) in advance (e.g., when the user looks into the wrong direction with few seconds before the ROI is enabled).

Claims

1. A content consumption device system for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree Video environment configured to:

receive at least one Video Stream associated to an Audio and Video scene to be reproduced; and

receive at least one first Audio Stream associated to the Audio and Video scene to be reproduced,

receive Audio information message metadata associated with at least one Audio information message from the at least one first Audio Stream,

wherein the content consumption device system comprises:

at least one media Video decoder configured to decode at least one Video signal from the at least one Video Stream for the representation of the Audio and Video scene to a user; and

at least one media Audio decoder configured to decode at least one Audio signal from the at least one first Audio Stream for the representation of the Audio and Video scene to the user;

a metadata processor;

a region of interest, ROI, processor, configured to:

receive a current user’s viewport and/or head orientation and/or movement data;

receive viewport metadata associated with the at least one Video signal from the at least one Video Stream, the viewport metadata defining at least one ROI;

decide, based at least on:

the user’s current viewport and/or head orientation and/or movement data; and

viewport metadata,

whether an Audio information message associated to the at least one ROI is to be reproduced, wherein the Audio information message is independent of the at least one Video signal and the at least one Audio signal; and

request, at the ROI processor's decision that the Audio information message is to be reproduced, the modification of the Audio information message metadata to the metadata processor; and

wherein the metadata processor is configured to receive an Audio information message metadata, to receive a request to modify the Audio information message metadata from the ROI processor, and to modify the Audio information message metadata to a

modified Audio information message metadata according to the request from the ROI processor,

wherein the ROI processor is further configured to cause the reproduction of the Audio information message is according to the modified Audio information message metadata.

2. The content consumption device system of claim 1, wherein the Audio information message is an earcon.

3. The content consumption device system of any of the preceding claims, wherein the ROI processor is configured to perform the decision also based on Audio information message metadata.

4. The content consumption device system of any of the preceding claims, wherein the ROI processor is further configured to perform the decision so as to:

in case the at least one ROI being outside the user's current viewport, cause the reproduction of an Audio information message associated to the at least one ROI, in addition to the reproduction of the at least one Audio signal; and

in case the at least one ROI being within the user's current viewport, deactivate the reproduction of the Audio information message associated to the at least one ROI.

5. The content consumption device system of any of the preceding claims, further configured to:

receive at least one additional Audio Stream in which the at least one Audio information message is encoded,

wherein the content consumption device system further comprises:

at least one muxer or multiplexer to merge, under the control of the metadata processor or another processor, packets of the at least one additional Audio Stream with packets of the at least one first Audio Stream in one Stream, based on the decision provided by the ROI processor that the at least one Audio information message is to be reproduced, to cause the reproduction of the Audio information message in addition to the Audio scene.

6. The content consumption device system of any of claims 1-4, further configured to:

receive at least one Audio metadata describing the at least one Audio signal encoded in the at least one first Audio Stream;

at the ROI processor's decision that the Audio information message is to be reproduced, modify, by the metadata processor, the Audio information message metadata to enable the reproduction of the Audio information message, in addition to the reproduction of the at least one Audio signal.

7. The content consumption device system of any of claims 1-4, further configured to:

receive at least one Audio metadata describing the at least one Audio signal encoded in the at least one first Audio Stream;

at the ROI processor's decision that the Audio information message is to be reproduced, modify, by the metadata processor, the Audio information message metadata to enable the reproduction of an Audio information message in association with the at least one ROI, in addition to the reproduction of the at least one Audio signal; and

modify, by the metadata processor, the Audio metadata describing the at least one Audio signal to allow a merge of the at least one first Audio Stream and the at least one additional Audio Stream.

8. The content consumption device system of any of claims 1-4, further configured to:

receive at least one Audio metadata describing the at least one Audio signal encoded in the at least one first Audio Stream;

at the ROI processor's decision that the Audio information message is to be reproduced, provide, by the metadata processor, the Audio information message metadata to a synthetic Audio generator to create a synthetic Audio Stream, so as to associate the Audio information message metadata to the synthetic Audio Stream, and to provide the synthetic Audio Stream and the Audio information message metadata to a multiplexer or muxer to allow a merge the at least one first Audio Stream and the synthetic Audio Stream.

9. The content consumption device system of claim 5, further configured to obtain the Audio information message metadata from the at least one additional Audio Stream in which the Audio information message is encoded.

10. The content consumption device system of any of the preceding claims, further configured to store, for future use, the Audio information message metadata and/or the Audio information message Stream.

11. The content consumption device system of any of the preceding claims, further comprising:

a synthetic Audio generator configured to synthesize an Audio information message on the basis of Audio information message metadata associated to the at least one ROI.

12. The content consumption device system of any of claims 1-4, wherein the metadata processor is further configured to control a muxer or multiplexer to merge, on the basis of the Audio metadata and the Audio information message metadata, packets of the Audio information message Stream with packets of the at least one first Audio Stream in one Stream to obtain an addition of the Audio information message to the at least one first Audio Stream.

13. The content consumption device system of any of the preceding claims, wherein the Audio information message metadata is encoded in a configuration frame or in a data frame including at least gain data associated to the loudness level of the Audio information message to be reproduced.

14. The content consumption device system of any of the preceding claims, wherein the Audio information message metadata is encoded in a configuration frame or in a data frame including at least one of:

positional data,
language of the Audio information message,
data text length,
data text of the associated text label, and
description of the Audio information message.

15. The content consumption device system of any of the preceding claims, wherein the metadata processor is further configured to modify Audio information message metadata to activate the Audio information message and/or set/change its position.

16. The content consumption device system of any of the preceding claims, wherein the metadata processor is further configured to perform at least one of the following operations:

extract Audio information message metadata from a Stream;
embed metadata back in a Stream;
feed the Stream to an additional media decoder;
extract Audio metadata from the least one first Audio Stream;
extract Audio information message metadata from an additional Stream;

modify Audio metadata of the least one first Audio Stream so as to take into consideration the existence of the Audio information message and allow merging;

feed a Stream to the multiplexer or muxer to multiplex or mux them based on the information received from the ROI processor.

17. The content consumption device system of any of the preceding claims, wherein the ROI processor is further configured to perform a local search for an additional Audio Stream in which the Audio information message is encoded and/or Audio information message metadata and, in case of non-retrieval, request the additional Audio Stream and/or Audio information message metadata to a remote entity.

18. The content consumption device system of claim 11, wherein the ROI processor is further configured to perform a local search for an additional Audio Stream and/or an Audio information message metadata and, in case of non-retrieval, cause the synthetic Audio generator to generate the Audio information message Stream and/or Audio information message metadata.

19. The content consumption device system of claim 1, further configured to:

receive at least one additional Audio Stream in which at least one Audio information message associated to the at least one ROI is included,

wherein the at least one media Audio decoder is further configured to decode the at least one additional Audio Stream if the ROI processor decides that an Audio information message associated to the at least one ROI is to be reproduced.

20. The content consumption device system of claim 19, wherein:

the at least one first media Audio decoder is configured for decoding the at least one Audio signal from at least one first Audio Stream;

the content consumption device system further includes

at least one additional Audio decoder configured for decoding the at least one Audio information message from an additional Audio Stream; and

at least one mixer and/or renderer for mixing and/or superimposing the Audio information message from the at least one additional Audio Stream with the at least one Audio signal from the at least one first Audio Stream.

21. The content consumption device system of any of the preceding claims, further configured to keep track of metrics defining the accumulated number of reproductions of the

Audio information message, so as to disable the Audio information message's reproduction if the metrics is over a predetermined threshold.

22. The content consumption device system of any of the preceding claims, wherein the ROI processor's decision is based on a prediction of user's current viewport and/or position and/or head orientation and/or movement data in relationship to the position of the ROI.

23. The content consumption device system of any of the preceding claims, further configured, at the ROI processor's decision that the information message is to be reproduced, to request an Audio message information Stream from a remote entity.

24. The content consumption device system of any of the preceding claims, further configured to establish whether to reproduce two Audio information messages at the same time or whether to select a higher-priority Audio information message to be reproduced with priority with respect to a lower-priority Audio information message.

25. The content consumption device system of any of the preceding claims, further configured to identify an Audio information message among a plurality of Audio information messages encoded in one additional Audio Stream on the basis of the address and/or position of the Audio information messages in an Audio Stream.

26. The content consumption device system of any of the preceding claims, wherein the Audio Streams are formatted in the MPEG-H 3D Audio Stream format.

27. The content consumption device system of any of the preceding claims, further configured to:

receive data about availability of a plurality of Adaptation Sets, the available Adaptation Sets including at least one Audio scene Adaptation Set for the at least one first Audio Stream and at least one Audio message Adaptation Set for the at least one additional Audio Stream containing at least one Audio information message;

create, based on the ROI processor's decision, selection data identifying which of the Adaptation Sets are to be retrieved, the available Adaptation Sets including at least one Audio scene Adaptation Set and/or at least one Audio message Adaptation Set; and

request and/or retrieve the data for the Adaptation Sets identified by the selection data,

wherein each adaptation set groups different encodings for different bitrates.

28. The content consumption device system of claim 27, wherein at least one if its elements comprises a Dynamic Adaptive Streaming over HTTP, DASH, client and/or is configured to retrieve the data for each of the adaptation set using the ISO Base Media File Format, ISO BMFF, or MPEG-2 Transport Stream, MPEG-2 TS.

29. The content consumption device system of any of the preceding claims, wherein the ROI processor is further configured to check correspondences between the ROI and the current viewport and/or position and/or head orientation and/or movement data so as to check whether the ROI is represented in the current viewport, and, in case the ROI is outside the current viewport, to audibly signal the presence of the ROI to the user.

30. The content consumption device system of any of the preceding claims, wherein the ROI processor is further configured to check correspondences between the ROI and the current viewport and/or position and/or head orientation and/or movement data so as to check whether the ROI is represented in the current viewport, and, in case the ROI is within the current viewport and/or position and/or head orientation and/or movement data, to refrain from audibly signal the presence of the ROI to the user.

31. The content consumption device system of any of the preceding claims, further configured to receive, from a remote entity, the at least one Video stream associated to the Video scene and the at least one first Audio Stream associated to the Audio scene, wherein the Audio scene is associated to the Video scene.

32. The content consumption device system of any of the preceding claims, wherein the ROI processor is further configured to choose, among a plurality of Audio information messages to be reproduced, the reproduction of one first Audio information message before a second Audio information message.

33. The content consumption device system of any of the preceding claims, further comprising a cache memory to store an Audio information message received from a remote entity or generated synthetically, to reuse the Audio information message at different instances of time.

34. The content consumption device system of any of the preceding claims, wherein the at least one Video stream and/or the at least one first Audio stream are part of the current Video scene and Audio scene, respectively, and are independent of the user's current

viewport and/or head orientation and/or movement data in the current Video and Audio scene.

35. The content consumption device system of any of the preceding claims, further configured to request the at least one first Audio stream and/or at least one Video stream to a remote entity in association to the Audio stream and/or Video stream, respectively, and to reproduce the at least one Audio information message on the basis of the user's current viewport and/or head orientation and/or movement data.

36. The content consumption device system of any of claims 1-34, further configured to request the at least one first Audio stream and/or at least one Video stream to a remote entity in association to the Audio stream and/or Video stream, respectively, and to request, to the remote entity, the at least one Audio information message on the basis of the user's current viewport and/or head orientation and/or movement data.

37. The content consumption device system of any of claims 1-34, further configured to request the at least one first Audio stream and/or at least one Video stream to a remote entity in association to the Audio stream and/or Video stream, respectively, and to synthesize the at least one Audio information message on the basis of the user's current viewport and/or head orientation and/or movement data.

38. The content consumption device system of any of the preceding claims, further configured to check at least one of additional criteria for the reproduction of the Audio information message, the criteria further including a user's selection and/or a user's setting.

39. The content consumption device system of any of claims 1-37, configured to check at least one of additional criteria for the reproduction of the Audio information message, the criteria further being based on inhibiting the repetition of an Audio information message if already reproduced, the criteria using a timer for avoiding too quick repetitions.

40. The content consumption device system of any of claims 1-37, configured to check at least one of additional criteria for the reproduction of the Audio information message, the criteria further including a flag in an audio message metadata obtained from a remote entity, wherein the flag defines whether the Audio information message is active.

41. A system comprising a client configured as the content consumption device system of any of the preceding claims and a remote entity configured as a server for delivering the at least one Video Stream and the at least one first Audio Stream.

42. The system of claim 41, wherein the remote entity is configured to search, in a database, intranet, internet, and/or geographical network, the at least one additional Audio Stream and/or Audio information message metadata and, in case of retrieval, delivery the at least one additional Audio Stream and/or the Audio information message metadata.

43. The system of claim 42, wherein the remote entity is configured to synthesize the at least one additional Audio Stream and/or generate the Audio information message metadata.

44. A method for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360 degree Video in a content consumption device system, the method comprising:

receiving, by the content consumption device system, at least one Video stream to be reproduced;

receiving, by the content consumption device system, at least one Audio stream to be reproduced;

receiving, by the content consumption device system, Audio information message metadata associated with at least one Audio information message from the at least one first Audio Stream;

decoding, by at least one media Video decoder, at least one Video signal from the at least one Video stream ;

decoding, by at least one media Audio decoder, at least one Audio signal from the at least one Audio stream;

receiving, by a region of interest, ROI, processor, a current user's viewport and/or head orientation and/or movement data;

receiving, by the ROI processor, viewport metadata associated with the at least one Video signal from the at least one Video Stream, the viewport metadata defining at least one ROI;

by the ROI processor, deciding, based on:

the user's current viewport and/or head orientation and/or movement data and viewport metadata,

whether an Audio information message associated to the at least one ROI is to be reproduced, wherein the Audio information message is independent on the at least one Video signal and the at least one Audio signal;

by the ROI processor, requesting, at the ROI processor's decision that the Audio information message is to be reproduced, the modification of the Audio information message metadata to the metadata processor;

receiving, by the metadata processor, an Audio information message metadata,

receiving, by the metadata processor, a request to modify the Audio information message metadata from the ROI processor,

modifying, by the metadata processor, the Audio information message metadata to a modified Audio information message metadata according to the request from the ROI processor,

by the ROI processor, causing, the reproduction of the Audio information message according to the modified Audio information message metadata.

45. The method of claim 44, further comprising:

reproducing the Audio and Video scene; and

in case the at least one ROI is outside the user’s current viewport, causing the reproduction of an Audio information message associated to the at least one ROI, in addition to the reproduction of the at least one Audio signal; and/or

in case the at least one ROI is within the user’s current viewport, deactivating the reproduction of the Audio information message associated to the at least one ROI.

46. A non-transitable storage unit comprising instructions which, when executed by a processor, cause the processor to perform a method according to claim 44 or 45.