Claims

1. A system for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment configured to receive video and audio streams to be reproduced in a media consumption device,

wherein the system comprises:

at least one media video decoder configured to decode video signals from video streams for the representation of VR, AR, MR or 360-degree video environments to a user, and

at least one audio decoder configured to decode audio signals from audio streams for the representation of audio scenes,

wherein the system is configured to request first audio streams and second audio streams and/or one audio element of an audio stream and/or one adaptation set to a server on the basis of at least the user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data, and

wherein the first audio streams are requested and/or received at a higher bitrate and/or quality level than the bitrate and/or quality level of the second audio streams,

wherein a plurality of N audio elements are defined, and, in case the user's distances to the positions or areas of these audio elements are larger than a predetermined threshold, the N audio elements are processed to obtain a smaller number M of audio elements associated to a position or area close to the position or area of the N audio elements, so as

to provide the system with at least one audio stream associated to the N audio elements, in case the user's distances to the positions or areas of the N audio elements are smaller than a predetermined threshold, or

to provide the system with at least one audio stream associated to the M audio elements, in case the user's distances to the positions or areas of the N audio elements are larger than a predetermined threshold.

2. A system for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment configured to receive video and audio streams to be reproduced in a media consumption device,

wherein the system comprises:

at least one media video decoder configured to decode video signals from video streams for the representation of VR, AR, MR or 360-degree video environments to a user, and

at least one audio decoder configured to decode audio signals from audio streams for the representation of audio scenes,

wherein the system is configured to request first audio streams and second audio streams and/or one audio element of an audio stream and/or one adaptation set to a server on the basis of at least the user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data, and

wherein the first audio streams are requested and/or received at a higher bitrate and/or quality level than the bitrate and/or quality level of the second audio streams,

wherein at least one video environment is associated to at least one plurality of N audio elements, each audio element being associated to a position and/or area in the video environment,

wherein the at least at least one plurality of N audio elements is provided in at least one representation at high bitrate, and

wherein the at least at least one plurality of N audio elements is provided in at least one representation at low bitrate and/or quality level, where the at least one representation is obtained by processing the N audio elements to obtain a smaller number M of audio elements associated to a position or area close to the position or area of the N audio elements.

3. The system of claim 1 or 2, configured to provide the server with the user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data so as to obtain the at least one audio stream and/or one audio element of an audio stream and/or one adaptation set from the server.

4. The system of claim 1 or 2, wherein at least one audio scene is associated to at least one audio element, each audio element being associated to a position and/or area in the video environment where the audio element is audible, so

that different audio streams are provided for different user's positions and/or viewports and/or head orientations and/or movement data and/or interaction metadata and/or virtual positional data in an audio scene.

5. The system of claim 1 or 2, configured to decide whether at least one audio element of an audio stream and/or one adaptation set is to be reproduced for the current user's viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual position in an audio scene, and

wherein the system is configured to request and/or to receive the at least one audio element at the current user's virtual position.

6. The system of claim 1 or 2, configured to predictively decide whether at least one audio element of an audio stream and/or one adaptation set will become audible based on at least the user´s current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data,

wherein the system is configured to request and/or to receive the at least one audio element and/or audio stream and/or adaptation set at a particular user's virtual position before the predicted user's movement and/or interaction in an audio scene, and

wherein the system is configured to reproduce the at least one audio element and/or audio stream, when received, at the particular user's virtual position after the user's movement and/or interaction in an audio scene.

7. The system of claim 1 or 2, configured to request and/or to receive the at least one audio element at a lower bitrate, at the user's virtual position before a user's interaction, the interaction resulting from either change in positional data in the same audio scene or entering a next scene separated from the current scene,

wherein the system is configured to request and/or to receive the at least one audio element at a higher bitrate, at the user's virtual position after the user's interaction in an audio scene.

8. The system of claim 1 or 2, wherein at least one audio element associated to at least one audio scene is associated to a position and/or area in the video environment associated to an audio scene,

wherein the system is configured to request and/or receive streams at higher bitrate for audio elements closer to the user than for audio elements more distant from the user.

9. The system of claim 1 or 2, wherein at least one audio element is associated to at least one audio scene, the at last one audio element being associated to a position and/or area in the video environment associated to an audio scene,

wherein the system is configured to request different streams at different bitrates for audio elements based on their relevance and/or auditability level at each user's virtual position in an audio scene,

wherein the system is configured to request an audio stream at higher bitrate for audio elements which are more audible at the current user's virtual position, and/or

an audio stream at lower bitrate for audio elements which are less audible at the current user's virtual position.

10. The system of claim 1 or 2, wherein at least one audio element is associated to an audio scene, each audio element being associated to a position and/or area in the video environment associated to an audio scene,

wherein the system is configured to periodically send to the server the user´s current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data, so that:

for a first position, a stream at higher bitrate is provided, from the server, and

for a second position, a stream at lower bitrate is provided, from the server,

wherein the first position is closer to the at least one audio element than the second position.

11. The system of claim 1 or 2, wherein a plurality of audio scenes are defined for multiple video environments such as adjacent and/or neighbouring video environments,

so that first streams are provided associated to a first, current audio scene and, in case of user’s transition to a second, further audio scene, to provide both the

audio streams associated to the first audio scene and the second streams associated to the second audio scene.

12. The system of claim 1 or 2, wherein a plurality of audio scenes are defined for a first and a second video environments, the first and second video environments being adjacent and/or neighbouring video environments,

wherein first streams associated to the first audio scene are provided, from the server, for the reproduction of the first audio scene in case of the user's position or virtual position being in a first video environment associated to the first audio scene,

second streams associated to the second audio scene are provided, from the server, for the reproduction of the second audio scene in case of the user's position or virtual position being in a second video environment associated to the second audio scene, and

both first streams associated to the first audio scene and second streams associated to the second audio scene are provided in case of the user's position or virtual position being in a transitional position between the first audio scene and the second audio scene.

13. The system of claim 1 or 2, wherein a plurality of audio scenes are defined for a first and a second video environments, which are adjacent and/or neighbouring environments,

wherein the system is configured to request and/or receive first streams associated to a first audio scene associated to the first environment, for the reproduction of the first audio scene in case of the user's virtual position being in the first environment,

wherein the system is configured to request and/or receive second streams associated to the second audio scene associated to the second environment, for the reproduction of the second audio scene in case of the user's virtual position being in the second environment, and

wherein the system is configured to request and/or receive both first streams associated to the first audio scene and second streams associated to the second audio scene in case of the user's virtual position being in a transitional position between the first environment and the second environment.

14. The system of any of claims 11-13, wherein

the first streams associated to the first audio scene are obtained at a higher bitrate when the user is in the first environment associated to the first audio scene,

while the second streams associated to the second audio scene associated to the second environment are obtained at a lower bitrate when the user is in the beginning of a transitional position from the first audio scene to the second audio scene, and

the first streams associated to the first audio scene are obtained at a lower bitrate and the second streams associated to the second audio scene are obtained at a higher bitrate when the user is in the end of a transitional position from the first audio scene to the second audio scene,

wherein the lower bitrate is lower than the higher bitrate.

15. The system of claim 1 or 2, wherein a plurality of audio scenes is defined for multiple environments such as adjacent and/or neighbouring environments,

so that the system is configured to obtain the audio streams associated to a first current audio scene associated to a first, current environment, and,

in case the distance of the user's position or virtual position from a boundary of an audio scene is below a predetermined threshold, the system further obtains audio streams associated to a second, adjacent and/or neighbouring video environment associated to the second audio scene.

16. The system of claim 1 or 2, wherein a plurality of audio scenes is defined for multiple video environments,

so that the system requests and/or obtains the audio streams associated to a current audio scene at a higher bitrate and the audio streams associated to the second audio scene at a lower bitrate,

wherein the lower bitrate is lower than the higher bitrate.

17. The system of claims 1 and 2, wherein, in case the user's distances are lower than a predetermined distance threshold, or the audibility levels are lower than a predetermined distance threshold, different audio streams are obtained for the different audio elements.

18. The system of claim 1 or 2, configured to request and/or obtain the audio streams on the basis of the user's orientation and/or user's direction of movement and/or user's interactions in an audio scene.

19. The system of claim 1 or 2, wherein the viewport is associated to the position and/or virtual position and/or movement data and/or head orientation.

20. The system of claim 1 or 2 or 19, wherein different audio elements are provided at different viewports, wherein the system is configured to request and/or receive, in case one first audio element falls within a viewport, the first audio element at a higher bitrate than a second audio element which does not fall within the viewport.

21. The system of claim 1 or 2, configured so as to request and/or receive first audio streams and second audio streams, wherein the first audio elements in the first audio streams are more relevant and/or more audible than the second audio elements in the second audio streams,

wherein the first audio streams are requested and/or received at a higher bitrate than the bitrate of the second audio streams.

22. The system of claim 1 or 2, wherein at least two visual environment scenes are defined, wherein at least one first and second audio elements are associated to a first audio scene associated to a first video environment, and at least one third audio element is associated to a second audio scene associated to a second video environment,

wherein the system is configured to obtain interaction metadata describing that the at least one second audio element is additionally associated with the second video environment,

wherein the system is configured to request and/or receive the at least one first and second audio elements in case the user's virtual position is in the first video environment,

wherein the system is configured to request and/or receive the at least one second and third audio elements in case the user's virtual position is in the second video environment, and

wherein the system is configured to request and/or receive the at least one first and second and third audio elements in case the user's virtual position is in transition between the first video environment and the second video environment.

23. The system of claim 22, wherein the at least one first audio element is provided in at least one audio stream and/or adaptation set, and the at least one second audio element is provided in at least one second audio stream and/or adaptation set, and the at least one third audio element is provided in at least one third audio stream and/or adaptation set, and wherein the at least one first video environment is described by interaction metadata as an audio scene which requires the at least one first and second audio streams and/or adaptation sets, and wherein the second video environment is described by interaction metadata as an audio scene which requires the at least one third audio stream and/or adaptation set and the at least one second audio stream and/or adaptation sets associated with the at least one first video environment,

wherein the system comprises a metadata processor configured to manipulate the interaction metadata, to merge the second audio stream belonging to the first video environment and the third audio stream associated with the second video environment into a new single stream, in case the user's virtual position is in the second video environment.

24. The system of claim 1 or 2, wherein the system comprises a metadata processor configured to manipulate metadata in at least one audio stream prior to the at least one audio decoder based on user’s current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data.

25. The system of claim 24, wherein the metadata processor is configured to enable and/or disable at least one audio element in at least one audio stream prior to

the at least one audio decoder based on user’s current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data, wherein

the metadata processor is configured to disable at least one audio element in at least one audio stream prior to the at least one audio decoder in case the system decides that the audio element is not to be reproduced anymore as a consequence of a current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data, and wherein

the metadata processor is configured to enable at least one audio element in at least one audio stream prior to the at least one audio decoder, in case the system decides the audio element is to be reproduced as a consequence of a user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data.

26. The system of claim 1 or 2, configured to disable the decoding of audio elements element selected the basis of the user’s current viewport and/or head orientation and/or movement data and/or metadata and/or virtual position.

27. The system of claim 1 or 2, configured to merge at least one first audio stream associated to a current audio scene to at least one stream associated to a neighbouring, adjacent and/or future audio scene.

28. The system of claim 1 or 2, configured to obtain and/or collect statistical or aggregated data on the user’s current viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data, so as to transmit the request to the server associated to the statistical or aggregated data.

29. The system of claim 1 or 2, configured to deactivate the decoding and/or reproduction of at least one stream on the basis of metadata associated to the at least one stream and on the basis of the user’s current viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data.

30. The system of claim 1 or 2 further configured to:

manipulate metadata associated with a group of selected audio streams based on at least the user's current or estimated viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data, so as to:

select and/or activate audio elements composing an audio scene to be reproduced; and/or

merge all selected audio streams into a single audio stream.

31. The system of claim 1 or 2, wherein information is provided from the server, for each audio element or audio object, wherein the information includes descriptive information about the locations in which an audio scene or the audio elements are active.

32. The system of claim 1 or 2, configured to choose between reproducing one audio scene and composing or mixing or muxing or superposing or combining at least two audio scenes on the basis of the current or future or viewport and/or head orientation and/or movement data and/or metadata and/or virtual position and/or a user's selection, the two audio scenes being associated to different neighbouring and/or adjacent environments.

33. The system of claim 1 or 2, configured to create or use at least the adaptation sets so that:

A number of Adaptation Sets are associated with one Audio Scene; and/or

Additional information is provided that relates each Adaptation Set to one Viewpoint, or one Audio Scene; and/or

Additional information is provided that include

- Information about the boundaries of one Audio Scene and/or
- Information about the relation between one Adaptation Set and one Audio Scene (e.g., Audio Scene is encoded in three streams that are encapsulated into three Adaptation Sets) and/or
- Information about the connection between the boundaries of an audio scene and the multiple Adaptation Sets.

34. The system according to any of the preceding claims, configured to:

receive a stream for an audio scene associated to a neighbouring or adjacent environment;

start decoding and/or reproducing the audio stream for the neighbouring or adjacent environment at the detection of the transition of a boundary between two environments.

35. A system comprising the system of any of the preceding claims configured for operating as a client and a server configured for delivering video and/audio streams to be reproduced in a media consumption device.

36. The system of claim 1 or 2, wherein the system is further configured to:

request and/or receive at least one first adaptation set comprising at least one audio stream associated with at least one first audio scene;

request and/or receive at least one second adaptation set comprising at least one second audio stream associated with at least two audio scenes, including the at least one first audio scene; and

merge the at least one first audio stream and the at least one second audio stream into a new audio stream to be decoded, based on metadata available regarding user's current viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data and/or information describing an association of the at least one first adaptation set to the at least one first audio scene and/or an association of the at least one second adaptation set to the at least one first audio scene.

37. The system of claim 1 or 2 configured to

receive information about user's current viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data and/or any information characterizing changes triggered by the user's actions; and

receive information about the availability of adaptation sets and information describing an association of at least one adaptation set to at least one audio scene and/or Viewpoint and/or viewport and/or position and/or virtual position and/or movement data and/or orientation.

38. The system of claim 1 or 2, configured to

decide if at least one audio element from at least one audio scene embedded in at least one audio stream and at least one additional audio element from at least one additional audio scene embedded in at least one additional audio stream are to be reproduced; and

cause, in case of a positive decision, an operation of merging or composing or muxing or superposing or combining at the least one additional stream of the additional audio scene to the at least one stream of the at least one audio scene.

39. The system of claim 1 or 2, configured to

manipulate audio metadata associated with selected audio streams, based on at least the user's current viewport and/or head orientation and/or movement data and/or metadata and/or virtual positional data, in order to:

select and/or enable and/or activate the audio elements composing an audio scene decided to be reproduced; and

enable the merge of all selected audio streams into a single audio stream.

40. A server for delivering audio and video streams to a client for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment, the video and audio streams to be reproduced in a media consumption device,

wherein the server comprises an encoder to encode and/or a storage to store video streams to describe a video environment, the video environment being associated to an audio scene;

wherein the server further comprises an encoder to encode and/or a storage to store a plurality of audio streams and/or audio elements and/or adaptation sets to be delivered to the client, the audio streams and/or audio elements and/or adaptation sets being associated to at least one audio scene,

wherein the server is configured to:

select and deliver a video stream on the basis of a request from the client, the video stream being associated to an environment;

select an audio stream and/or audio element and/or adaptation set on the basis of a request from the client, the request being associated to at least

the user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data and to an audio scene associated to the environment; and

deliver the audio stream to the client,

wherein the first audio streams are requested and/or received at a higher bitrate than the bitrate of the second audio streams,

wherein a plurality of N audio elements are defined, and, in case the user's distances to the positions or areas of these audio elements is larger than a predetermined threshold, the N audio elements are processed to obtain a smaller number M of audio elements associated to positions or areas close to the position or area of the N audio elements, so as

to provide the system with at least one audio stream associated to the N audio elements, in case the user's distances to the positions or areas of the N audio elements are smaller than a predetermined threshold, or

to provide the system with at least one audio stream associated to the M audio elements, in case the user's distances to the positions or areas of the N audio elements are larger than a predetermined threshold.

41. A server for delivering audio and video streams to a client for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment, the video and audio streams to be reproduced in a media consumption device,

wherein the server comprises an encoder to encode and/or a storage to store video streams to describe a video environment, the video environment being associated to an audio scene;

wherein the server further comprises an encoder to encode and/or a storage to store a plurality of audio streams and/or audio elements and/or adaptation sets to be delivered to the client, the audio streams and/or audio elements and/or adaptation sets being associated to at least one audio scene,

wherein the server is configured to:

select and deliver a video stream on the basis of a request from the client, the video stream being associated to an environment;

select an audio stream and/or audio element and/or adaptation set on the basis of a request from the client, the request being associated to at least

the user's current viewport and/or head orientation and/or movement data and/or interaction metadata and/or virtual positional data and to an audio scene associated to the environment; and

deliver the audio stream to the client,

wherein the first audio streams are requested and/or received at a higher bitrate than the bitrate of the second audio streams,

wherein at least one video environment is associated to at least one plurality of N audio elements, each audio element being associated to a position and/or area in the video environment,

wherein the at least at least one plurality of N audio elements is provided in at least one representation at high bitrate, and

wherein the at least at least one plurality of N audio elements is provided in at least one representation at low bitrate, where the at least one representation is obtained by processing the N audio elements to obtain a smaller number M of audio elements associated to a position or area close to the position or area of the N audio elements.

42. The server of claim 40 or 41, wherein the audio streams are encapsulated into adaptation sets, each adaptation set including a plurality of streams associated to different representations, at different bitrate, of a same audio content,

wherein the selected adaptation set is selected on the basis of the request from the client.

43. A method for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment configured to receive video and/audio streams to be reproduced in a media consumption device, comprising:

decoding video signals from video streams for the representation of VR, AR, MR or 360-degree video environments to a user,

decoding audio signals from audio streams for the representation of audio scenes,

requesting to, and/or obtaining from, a server at least one audio stream on the basis of the user's current viewport and/or positional data and/or head orientation

and/or movement data and/or metadata and/or virtual positional data and/or metadata, and

wherein the first audio streams are requested and/or received at a higher bitrate than the bitrate of the second audio streams,

wherein a plurality of N audio elements are defined, and, in case the user's distances to the positions or areas of these audio elements is larger than a predetermined threshold, the N audio elements are processed to obtain a smaller number M of audio elements associated to a position or area close to the positions or areas of the N audio elements, so as

to provide the system with at least one audio stream associated to the N audio elements, in case the user's distances to the positions or areas of the N audio elements are smaller than a predetermined threshold, or

to provide the system with at least one audio stream associated to the M audio elements, in case the user's distances to the positions or areas of the N audio elements is larger than a predetermined threshold.

44. A method for a virtual reality, VR, augmented reality, AR, mixed reality, MR, or 360-degree video environment configured to receive video and/audio streams to be reproduced in a media consumption device, comprising:

decoding video signals from video streams for the representation of VR, AR, MR or 360-degree video environments to a user,

decoding audio signals from audio streams for the representation of audio scenes,

requesting to, and/or obtaining from, a server at least one audio stream on the basis of the user's current viewport and/or positional data and/or head orientation and/or movement data and/or metadata and/or virtual positional data and/or metadata, and

wherein the first audio streams are requested and/or received at a higher bitrate than the bitrate of the second audio streams,

wherein at least one video environment is associated to at least one plurality of N audio elements, each audio element being associated to a position and/or area in the video environment,

wherein the at least at least one plurality of N audio elements is provided in at least one representation at high bitrate, and

wherein the at least at least one plurality of N audio elements is provided in at least one representation at low bitrate, where the at least one representation is obtained by processing the N audio elements to obtain a smaller number M of audio elements associated to a position or area close to the position or area of the N audio elements.

45. A computer program comprising instructions which, when executed by a processor, cause the processor to perform the method of claim 43 or 44.