



(12)

EUROPEAN PATENT APPLICATION

- (43)

Date of publication:
09.04.2025 Bulletin 2025/15
- (51)

International Patent Classification (IPC):
H04S 7/00 (2006.01)
- (21)

Application number: 23201344.1
- (52)

Cooperative Patent Classification (CPC):
H04S 7/303; H04S 7/304; H04S 3/008;
H04S 2400/11; H04S 2420/01; H04S 2420/11
- (22)

Date of filing: 03.10.2023

- (84)

Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC ME MK MT NL
NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA
Designated Validation States:
KH MA MD TN
- (71)

Applicant: Koninklijke Philips N.V.
5656 AG Eindhoven (NL)
- (72)

Inventors:
• JELFS, Sam Martin
Eindhoven (NL)
- (74)

Representative: Philips Intellectual Property &
Standards
High Tech Campus 52
5656 AG Eindhoven (NL)
- SZCZERBA, Marek Zbigniew
Eindhoven (NL)

• KOPPENS, Jeroen Gerardus Henricus
Eindhoven (NL)

• DE BONT, Fransiscus Marinus Jozephus
Eindhoven (NL)

• OOMEN, Arnoldus Werner Johannes
Eindhoven (NL)

(54)

GENERATING OF AN AUDIO DATA SIGNAL

- (57)

An audio apparatus comprises a receiver (301) receiving audio elements including a number of audio objects linked with a position in an audio scene. A listener pose receiver (303) receives an indication of a listener pose and a designator (305) designates audio objects as close or remote audio objects depending on a comparison of a distance measure indicative of a distance between a pose of the first audio object and the listener pose to a threshold. An audio mix generator (307) generates
- an Ambisonic audio mix from a first plurality of the audio elements that includes one or more audio objects designated as remote audio object. A data generator (311) generates an audio data signal comprising the Ambisonic audio mix and which further includes audio objects that are designated as close audio objects. The audio apparatus may be an edge device operating with an audio end device to provide split rendering of audio across the devices.

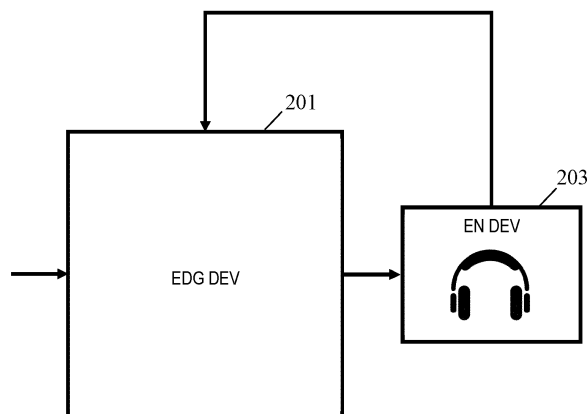


FIG. 2

Description

FIELD OF THE INVENTION

[0001] The invention relates to generating, and in some cases rendering, an audio data signal, and in particular, but not exclusively, to generating such signals to support e.g., split rendering applications.

BACKGROUND OF THE INVENTION

[0002] The variety and range of experiences based on audiovisual content have increased substantially in recent years with new services and ways of utilizing and consuming such content continuously being developed and introduced. In particular, many spatial and interactive services, applications and experiences are being developed to give users a more involved and immersive experience.

[0003] Examples of such applications are Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) applications (commonly referred to as eXtended Reality XR applications) which are rapidly becoming mainstream, with a number of solutions being aimed at the consumer market. A number of standards are also under development by a number of standardization bodies. Such standardization activities are actively developing standards for the various aspects of VR/AR/MR/XR systems including e.g., streaming, broadcasting, rendering, etc.

[0004] VR applications tend to provide user experiences corresponding to the user being in a different world/ environment/ scene whereas AR (including Mixed Reality MR) applications tend to provide user experiences corresponding to the user being in the current environment but with additional information or virtual objects or information being added. Thus, VR applications tend to provide a fully immersive synthetically generated world/ scene whereas AR applications tend to provide a partially synthetic world/ scene which is overlaid on the real scene in which the user is physically present. However, the terms are often used interchangeably and have a high degree of overlap. In the following, the term Virtual Reality/ VR will be used to denote both Virtual Reality and Augmented Reality.

[0005] VR applications typically provide a virtual reality experience to a user allowing the user to (relatively) freely move around in a virtual environment and to dynamically change their pose (pose and/or orientation). Typically, such virtual reality applications are based on a three-dimensional model of the scene with the model being dynamically evaluated to provide the specific requested view. This approach is well known from e.g., game applications, such as in the category of first person shooters, for computers and consoles.

[0006] In addition to the visual rendering, most VR (and more generally XR) applications further provide a corresponding audio experience. In many applications, the

audio preferably provides a spatial audio experience where audio sources are perceived to arrive from positions that correspond to the positions of the corresponding objects in the visual scene (including both objects that are currently visible and objects that are not currently or partly visible (e.g., behind the user)). Thus, the audio and video scenes are preferably perceived to be consistent and provide a full spatial experience.

[0007] For audio, headphone reproduction using binaural audio rendering technology is widely used. In many scenarios, headphone reproduction enables a highly immersive, personalized user experience. Using headtracking, the rendering can be made responsive to the user's head movements, which highly increases the sense of immersion.

[0008] In order to generate suitable audio, the rendering devices are provided with spatial audio data representing an audio scene. However, audio sources can be represented in many different ways including as audio channel signals, audio objects, diffuse non-spatially specific audio sources (e.g., background noise), Ambisonic audio, etc. Indeed, it is becoming increasingly prevalent to provide audio scene descriptions using a plurality of different audio representations to reflect the different types of audio sources that may be represented. However, such approaches may increase complexity and resource requirements and often lead to processing-intensive rendering algorithms.

[0009] Such issues are problematic and in particular in conflict with a desire to enable devices that can render audiovisual content with very low complexity and resources. In particular, it is increasingly desirable to provide end user devices that are low cost, small size, low weight, low complexity, and with low computational resource. For example, body worn devices with such properties are becoming increasingly widespread. In particular, in order to support various XR applications, it is desirable to be able to provide small and lightweight XR headsets, such as specifically relatively small and lightweight XR glasses.

[0010] In order to enable or facilitate such devices, it has been proposed to use an approach of split rendering where (the final) part of the rendering process is performed by an end device whereas other more computationally demanding parts of the rendering are performed by another device, referred to as an edge device. The edge device performs a potentially substantial part of the rendering process and may generate audio signals that are then transmitted to the end device for a final adaptation.

[0011] The edge device is typically a substantially more complex and resourceful device than the end device and accordingly can implement more complex rendering algorithms and functions. For example, the edge device may be a mobile phone, game console, computer, remote server etc. and the end device may be a user worn rendering and audio reproduction device such as an XR headset/glasses.

[0012] As an example, it has been proposed for an edge device to render received audio data to generate rendered audio signals that are then transmitted to the end device which may perform additional simple operations on these audio signals (e.g. simple panning) to adapt to the user movement.

[0013] However, whereas such approaches may provide desirable applications and operations in many scenarios, current approaches tend to be suboptimum. In particular, in many situations, current approaches may provide a suboptimum audio quality, a suboptimal response to user movement, a suboptimal resource usage and distribution, etc.

[0014] Hence, an improved approach for distribution and/or rendering and/or processing of audio signals, in particular for a Virtual/ Augmented/ Mixed/ eXtended Reality experience/ application, would be advantageous. In particular, an approach that allows improved operation, increased flexibility, reduced complexity, facilitated implementation, an improved user experience, improved audio quality, improved adaptation to audio reproduction functions, facilitated and/or improved adaptation to changes in listener position/orientation (e.g., a virtual listener position/orientation), improved resource demand/processing distribution for split rendering approaches, an improved eXtended Reality experience, and/or improved performance and/or operation would be advantageous.

SUMMARY OF THE INVENTION

[0015] Accordingly, the invention seeks to preferably mitigate, alleviate or eliminate one or more of the above mentioned disadvantages singly or in any combination.

[0016] According to an aspect of the invention there is provided an audio apparatus comprising: a receiver arranged to receive a plurality of audio elements representing a three dimensional audio scene, the audio elements including a number of audio objects, each audio object being linked with a position in the three dimensional audio scene; a listener pose receiver arranged to receive an indication of a listener pose; a designator arranged to designate at least a first audio object of the number of audio objects as a close audio object or as a remote audio object depending on a comparison of a distance measure indicative of a distance between a pose of the first audio object and the listener pose to a threshold; an audio mix generator arranged to generate an Ambisonic audio mix from a first plurality of the audio elements, the first plurality of audio elements comprising the first audio object if this is designated as a remote audio object; a data generator arranged to generate an audio data signal comprising the Ambisonic audio mix; and wherein the data generator is arranged to include the first audio object in the audio data signal if the first audio object is designated as a close audio object.

[0017] The approach may allow an improved output audio signal to be generated. The approach may in many

embodiments and scenarios provide an improved audio quality and may e.g., allow a reduced complexity for a rendering device rendering audio based on the audio data signal.

[0018] The approach may in particular allow improved split rendering where the rendering of an audio scene may be split over a plurality of devices. The audio apparatus may provide an audio data signal which allows improved distribution of functionality, complexity, and computational resource usage. The audio apparatus may generate an audio data signal which provides an improved trade-off between the additional requirements for rendering of individual audio objects and rendering multiple sources comprised in a single Ambisonic audio mix.

[0019] A pose may be a position and/or orientation. The listener pose may be a listener position, orientation, or position and orientation.

[0020] The designator may be arranged to designate at least a first audio object of the number of audio objects as a close audio object or as a remote audio object depending on whether (or not) a distance measure indicative of a distance between a pose of the first audio object and the listener pose exceeds a threshold. The designator may be arranged to designate at least a first audio object of the number of audio objects as a close audio object if a distance measure indicative of a distance between a pose of the first audio object and the listener pose is below a threshold and as a remote audio object if the distance measure exceeds the threshold (and typically depending on the individual embodiment as a remote or close object if the distance measure is the same as the threshold).

[0021] The listener pose may be indicative of a pose in the audio scene. The audio scene may be a three dimensional audio scene.

[0022] The audio mix generator may be arranged to generate the Ambisonic audio mix to include the first audio object only if this is designated as a remote audio object. The data generator may be arranged to not include the first audio object in the audio data signal if the first audio object is designated as a remote audio object.

[0023] In accordance with an optional feature of the invention, the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a loudness measure for the first audio object.

[0024] In accordance with an optional feature of the invention, the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a direction from the listener pose to a position of the first audio object.

[0025] In accordance with an optional feature of the invention, the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a trajectory of a position of the first audio object in the audio scene.

[0026] In accordance with an optional feature of the

invention, the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on an order of the Ambisonic audio mix.

[0027] In accordance with an optional feature of the invention, the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a distance between the position of the first audio object and a position of a second audio object.

[0028] In accordance with an optional feature of the invention, the data generator is arranged to transmit the audio data signal to a remote apparatus over a communication link; and the designator is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a data transfer property of the communication link.

[0029] These features may provide improved and/or facilitated operation and/or performance in many embodiments. The features may result in the generation of an audio data signal which provides an improved trade-off between the additional requirements for rendering of individual audio objects and rendering multiple sources comprised in a single Ambisonic audio mix.

[0030] In accordance with an optional feature of the invention, the audio mix generator is arranged to vary a data rate for the first audio object when transitioning between being designated as a close object and being designated as a remote object.

[0031] This may provide improved and/or facilitated operation and/or performance in many embodiments.

[0032] In accordance with an optional feature of the invention, the listener pose receiver arranged to receive a plurality of listener poses, the designator is arranged to determine the distance measure in dependence on the plurality of listener poses, and the data generator is arranged to transmit the audio data signal to a plurality of remote devices.

[0033] This may provide improved and/or facilitated operation and/or performance in many embodiments.

[0034] In accordance with an optional feature of the invention, there is provided an audio system comprising an audio apparatus as described above and a rendering device comprising: a receiver arranged to receive the audio data signal from the audio apparatus; a renderer arranged to generate a binaural audio signal representing the audio scene, the binaural audio signal comprising contributions from the first audio mix and the first audio object if present in the audio data signal.

[0035] In accordance with an optional feature of the invention, the rendering device is a headset comprising audio transducers arranged to reproduce the binaural rendering signal.

[0036] In accordance with an optional feature of the invention, the designator is arranged to designate the first audio object as a close audio object or as a remote audio object in dependence on a property of the rendering device.

[0037] This may provide improved and/or facilitated operation and/or performance in many embodiments.

[0038] In accordance with an aspect of the invention, there is provided a method of operation for an audio apparatus, the method comprising: receiving a plurality of audio elements representing a three dimensional audio scene, the audio elements including a number of audio objects, each audio object being linked with a position in the three dimensional audio scene; receiving an indication of a listener pose; designating at least a first audio object of the number of audio objects as a close audio object or as a remote audio object depending on a comparison of a distance measure indicative of a distance between a pose of the first audio object and the listener pose to a threshold; generating an Ambisonic audio mix from a first plurality of the audio elements, the first plurality of audio elements comprising the first audio object if this is designated as a remote audio object; and generating an audio data signal comprising the Ambisonic audio mix; and including the first audio object in the audio data signal if the first audio object is designated as a close audio object.

[0039] In accordance with an optional feature of the invention, the audio apparatus performing the method of claim 13 and transmitting the audio data signal to a rendering device; and the rendering device performs the steps of: receiving the audio data signal from the audio apparatus; and generating a binaural audio signal representing the audio scene, the binaural audio signal comprising contributions from the first audio mix and the first audio object if present in the audio data signal.

[0040] These and other aspects, features and advantages of the invention will be apparent from and elucidated with reference to the embodiment(s) described hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0041] Embodiments of the invention will be described, by way of example only, with reference to the drawings, in which

FIG. 1 illustrates an example of a client server based eXtended Reality system;

FIG. 2 illustrates an example of elements of an audio rendering arrangement in accordance with some embodiments of the invention;

FIG. 3 illustrates an example of elements of an audio apparatus in accordance with some embodiments of the invention;

FIG. 4 illustrates an example of elements of an audio rendering apparatus in accordance with some embodiments of the invention;

FIG. 5 illustrates an example of audio objects in an audio scene;

FIG. 6 illustrates an example of audio objects in an audio scene; and

FIG. 7 illustrates some elements of a possible ar-

rangement of a processor for implementing elements of an apparatus in accordance with some embodiments of the invention.

DETAILED DESCRIPTION OF SOME EMBODIMENTS OF THE INVENTION

[0042] The following description will focus on eXtended Reality applications where audio is rendered reflecting a user pose in an audio scene to provide an immersive user experience. Typically, the audio rendering may be accompanied by a rendering of images such that a complete audiovisual experience is provided to the user. However, it will be appreciated that the described approaches may be used in many other applications.

[0043] EXtended Reality (including Virtual Augmented and Mixed Reality) experiences allowing a user to move around in a virtual or augmented world are becoming increasingly popular and services are being developed to improve such applications. In many such approaches, visual and audio data may dynamically be generated to reflect a user's (or viewer's) current pose.

[0044] In the field, the terms placement and pose are used as a common term for position and/or orientation / direction. The combination of the position and direction/ orientation of e.g., an object, a camera, a head, or a view may be referred to as a pose or placement. Thus, a placement or pose indication may comprise up to six values/ components/ degrees of freedom with each value/ component typically describing an individual property of the position/ location or the orientation/ direction of the corresponding object. Of course, in many situations, a placement or pose may be represented by fewer components, for example if one or more components is considered fixed or irrelevant (e.g., if all objects are considered to be at the same height and have a horizontal orientation, four components may provide a full representation of the pose of an object). In the following, the term pose is used to refer to a position and/or orientation which may be represented by one to six values (corresponding to the maximum possible degrees of freedom).

[0045] Many XR applications are based on a pose having the maximum degrees of freedom, i.e., three degrees of freedom for the position and three degrees of freedom for the orientation resulting in a total of six degrees of freedom. A pose may thus be represented by a set or vector of six values representing the six degrees of freedom and thus a pose vector may provide a three-dimensional position and/or a three-dimensional direction indication. However, it will be appreciated that in other embodiments, the pose may be represented by fewer values (Rotation may also be in the form of a quaternion or rotation matrix, so more than six values are possible).

[0046] A system or entity based on providing the maximum degree of freedom for the viewer is typically referred to as having 6 Degrees of Freedom (6DoF). Many systems and entities provide only an orientation or posi-

tion and these are typically known as having 3 Degrees of Freedom (3DoF - which is typically used to represent approaches using a pose with a variable orientation and a fixed position).

[0047] Typically, the Virtual Reality application generates a three-dimensional output in the form of separate view images for the left and the right eyes. These may then be presented to the user by suitable means, such as typically individual left and right eye displays of a VR headset. In other embodiments, one or more view images may e.g., be presented on an autostereoscopic display, or indeed in some embodiments only a single two-dimensional image may be generated (e.g., using a conventional two-dimensional display).

[0048] Similarly, for a given viewer/ user/ listener pose, an audio representation of the scene may be provided. The audio scene is typically rendered to provide a spatial experience where audio sources are perceived to originate from desired positions. As audio sources may be static in the scene, changes in the listener pose will result in a change in the relative position of the audio source with respect to the user's pose. Accordingly, the spatial perception of the audio source may change to reflect the new position relative to the user. The audio rendering may accordingly be adapted depending on the listener pose.

[0049] The listener pose input may be determined in different ways in different applications. In many embodiments, the physical movement of a user may be tracked directly. For example, a camera surveying a user area may detect and track the user's head (or even eyes (eye-tracking)). In many embodiments, the user may wear a VR headset which can be tracked by external and/or internal means. For example, the headset may comprise accelerometers and gyroscopes providing information on the movement and rotation of the headset and thus the head. In some examples, the VR headset may transmit signals or comprise (e.g., visual) identifiers that enable an external sensor to determine the position and orientation of the VR headset.

[0050] In many systems, the VR/ scene data, and in particular the audio data representing an audio scene, may be provided from a remote device or server. For example, a remote server may generate audio data representing an audio scene and may transmit audio signals corresponding to audio components/ objects/ channels, or other audio elements corresponding to different audio sources in the audio scene together with position information indicative of the position of these (which may e.g., dynamically change for moving objects). The audio signals/elements may include elements associated with specific positions but may also include elements for more distributed or diffuse audio sources. For example, audio elements may be provided representing generic (non-localized) background sound, ambient sound, diffuse reverberation etc.

[0051] The local VR device may then render the audio elements appropriately, and specifically by applying ap-

propriate binaural processing reflecting the relative position of the audio sources for the audio components.

[0052] Similarly, a remote device may generate visual/-video data representing a visual audio scene and may transmit visual scene components/ objects/ signals or other visual elements corresponding to different objects in the visual scene together with position information indicative of the position of these (which may e.g., dynamically change for moving objects). The visual items may include elements associated with specific positions but may also include video items for more distributed sources.

[0053] In some embodiments, the visual items may be provided as individual and separate items, such as e.g., descriptions of individual scene objects (e.g., dimensions, texture, opaqueness, reflectivity etc.). Alternatively or additionally, visual items may be represented as part of an overall model of the scene e.g., including descriptions of different objects and their relationship to each other.

[0054] For a VR service, a central server may accordingly in some embodiments generate audiovisual data representing a three dimensional scene, and may specifically represent the audio by a number of audio signals representing audio sources in the scene which can then be rendered by the local client/ device.

[0055] FIG. 1 illustrates an example of a VR/XR system in which a central server 101 liaises with a number of remote clients 103 e.g., via a network 105, such as e.g., the Internet. The central server 101 may be arranged to simultaneously support a potentially large number of remote clients 103.

[0056] Such an approach may in many scenarios provide an improved trade-off e.g., between complexity and resource demands for different devices, communication requirements etc. For example, the scene data may be transmitted only once or relatively infrequently with the local rendering device (the remote client 103) receiving a viewer pose and locally processing the scene data to render audio and/or video to reflect changes in the viewer pose. This approach may provide for an efficient system and attractive user experience. It may for example substantially reduce the required communication bandwidth while providing a low latency real time experience while allowing the scene data to be centrally stored, generated, and maintained. It may for example be suitable for applications where a VR experience is provided to a plurality of remote devices.

[0057] In some cases, the remote clients may include a plurality of devices that are arranged to interwork to provide the rendering of the audiovisual data. In particular, as illustrated in FIG. 2, a rendering arrangement (which specifically may correspond to a remote client) may include a first device 201, also referred to as an edge device 201, which receives audiovisual data describing a scene. In order to produce an audio representation, the first/edge device 201 accordingly receives audio data that describes an audio scene.

[0058] The edge device 201 is arranged to process the received audio data to generate intermediate audio data that is transmitted to an end device 203 which is arranged to process the intermediate audio data to generate rendered output audio signals. These output signals may specifically be a binaural signal for the left and right ear of a listener respectively. The approach thus uses a split rendering approach where the rendering of the audio (scene) is split over more than one device. The end device 203 may in many embodiments include audio reproduction means, such as specifically audio transducers (and often with one audio transducer for the right ear of the user and one audio transducer for the left ear of the user).

[0059] The edge device 201 may specifically be a mobile phone, computer, game console, laptop, tablet, etc. and the end device 203 may typically be a user worn reproduction device, such as XR glasses and/or headphones.

[0060] FIG. 3 illustrates an example of some elements of an audio apparatus that is arranged to generate an audio data signal from audio data describing an audio scene, and typically for a three dimensional audio scene. The apparatus may specifically be the edge device 201 of FIG. 2 and will be described with reference thereto. A critical issue for an approach, such as that of FIG. 2, is that of how to distribute the functionality and processing across the different devices and of what data to transmit from the edge device to the end device. Such considerations include not only considerations of the computational loads at the different devices but also consideration of other parameters such as the impact of the communication between the devices including the impact of communication errors and delay. For example, the responsiveness of the rendered audio to fast changes in the listener pose is typically highly dependent on the communication delay. Accordingly, it is desirable for the speed and responsiveness that the processing is predominantly performed at the end device. However, from a resource, size, complexity etc., perspective, it is typically desirable for the processing to be performed predominantly at the edge device. Accordingly, a trade-off between conflicting requirements tends to be critical for the performance of the approach.

[0061] The edge device 201 of FIG. 3 comprises a receiver which is arranged to receive audio data that describes a three dimensional audio scene. The audio data includes a plurality of audio elements with each audio element providing a representation of audio/sound in the audio scene. An audio element may be an audio signal representing an audio source, diffuse noise, ambient sound, a plurality of sources etc. In many cases, the audio elements may include a range of different types of audio elements, including audio objects, audio channel signals, Ambisonic audio signals/mixes, etc.

[0062] In some cases, one or more audio elements may include one or more audio channel signals. Such channel signals may for example be provided for pre-

determined and/or nominal positions, such as nominal speaker positions. Each channel may in such cases have information about a specific part of the mix, relative to the specified speaker configuration.

[0063] The audio elements specifically include a number of audio objects. Each audio object provides a description/representation of audio, and typically a description/representation of audio from an audio source. Each audio object is linked with a position in the audio scene, and thus the audio object provides audio data and spatial data. The position/orientation may e.g., be absolute to the scene (Global), relative to another object within the scene, or relative to the listener. Typically, an audio object provides audio data and position data for an audio source in the audio scene. Object-based immersive audio may add audio streams and metadata into a decoder, giving instructions on how each stream should be placed in the 3D sound field.

[0064] In many scenarios, the audio elements may further comprise one or more Ambisonic signals/mixes.

[0065] Ambisonics and in particular Higher Order Ambisonics (HOA) is a method of recording and reproducing a 3-dimensional sound field (scene-based audio). Unlike the common channel-oriented transmission methods, the system focuses on the reproduction of the entire sound field at the listening position and does not require a predetermined number of speaker positions for sound reproduction. The relevant loudspeaker signals are calculated for each loudspeaker position used by mathematical derivations of the transmitted values for sound pressure and speed. In the basic version, known as the First-order Ambisonics, Ambisonics can be understood as a three-dimensional extension of M/S (mid/side) stereo, adding additional difference channels for height and depth. The resulting signal set is called B-format. The sound information is transmitted in four channels W, X, Y and Z. The component W contains only the sound pressure, which is typically recorded with a non-directional or omni-directional microphone. The signals X, Y and Z are the directional components in the corresponding spatial axes. They may be recorded with microphones whose figure-of-eight is aligned with the corresponding axis. A simple Ambisonic approach is the A-Format which uses 4 cardioid microphones, which can then be converted to a typically more practical B-format.

[0066] The purpose of the process is to reconstruct the recorded sound pressure and the associated sound direction vector from these signals at the listener's listening position.

[0067] For audio objects that are part of the Ambisonics mix, the Ambisonics order determines the amount of spatial smearing of point sources. The higher the order, the lower the spatial smearing and the higher the spatial directionality.

[0068] The edge apparatus 301 further comprises a listener pose receiver 303 arranged to receive an indication of a listener pose. The listener pose receiver 303 is arranged to determine a listener pose from data which

may be provided to the edge device 201 from the end device 203. Thus, the edge device 201 and end device 203 may establish a communication link allowing listener pose data to be transmitted from the end device 203 to the edge device 201. The processing of the edge device 201 may thus be dependent on the pose of the end device 203, e.g., a VR headset/glasses may track user movement and generate sensor data that is transmitted to the edge device 201 and the edge device 201 may adapt processing in dependence on the user movement.

[0069] The listener pose may specifically be determined in response to sensor input, e.g., from suitable sensors being part of a headset. It will be appreciated that many suitable algorithms will be known to the skilled person and for brevity this will not be described in more detail herein.

[0070] In some embodiments, the determination of a listener pose in the audio scene may be performed in the end device 203 which may then transmit the determined listener pose to the edge device. In some embodiments, other data, such as sensor data, may be transmitted to the edge device 201 which may on the basis thereof be arranged to determine the listener pose.

[0071] The listener pose is typically a pose in the audio scene from which the audio presented to the user is to be perceived from, i.e., it represents the listener/user's position in the audio scene. In the approach of FIG. 2, the edge device 201 and end device 203 cooperate to process the received audio data to generate output audio signals, such as specifically a binaural audio output signal, for the user to provide a spatial audio experience/perception of the audio scene from the listener pose. Information of the listener pose is provided to the edge device 201 such that the processing of both the edge device 201 and the end device 203 can adapt to the current listener pose.

[0072] A critical issue in such scenarios is which processing to perform at respectively the edge device 201 and at the end device 203 and which specific information to transmit from the edge device 201 to the end device 203. Typically, the edge device 201 has substantially more computational resource than the end device 203 and therefore it is desirable to perform much of the processing, and in particular complex resource demanding processing, at the edge device 201. However, the communication of the data on the listener pose, and the communication of the corresponding audio data to the end device 203 introduces a communication delay which is typically significant and which can result in a reduced user experience. It introduces a round trip delay which can lead to a perceptually significant delay between the user's movements and the corresponding perceived audio. Therefore, it is often desirable for the end device 203 to be able to perform some, preferably low complexity, processing to locally adapt the rendered audio to changes in the listener pose while still maintaining the bulk of the processing at the edge device 201. However, in order to achieve an efficient trade-off, the distribution of

processing and the data communicated from the edge device 201 to the end device 203 is critical.

[0073] FIG. 4 illustrates an example of elements of the end device 203 of FIG. 2. The end device 203 comprises a receiver 401 which receives an audio data signal from the edge device 201. As will be described in more detail in the following, the audio data signal comprises audio signals generated by the edge device 201 to represent the audio scene. The receiver 401 is coupled to a renderer 403 which is arranged to render an audio signal to a listener from the received audio signals. The renderer 403 specifically generates a binaural audio signal representing the audio scene from the received audio data. As will be described in the following, the audio data signals include one or more Ambisonic audio mixes and typically one or more audio objects, and the renderer 403 is arranged to generate the binaural audio signal from these signals.

[0074] The binaural output signal is in the example fed to an output circuit 405 which outputs the binaural signal to headphones that may typically be part of a VR headset. The output circuit 405 may for example include Digital-to-Analog Converters, amplifier functions etc. as will be well known to the skilled person.

[0075] The renderer 403 is coupled to a listener pose processor 409 which is arranged to determine a listener pose. In the example, the listener pose is determined based on sensor input from the headphones/VR headset 407. The listener pose is provided to the renderer 403 which is arranged to generate the binaural signal to represent the audio scene from the listener pose, and specifically to dynamically affect this to follow changes in the listener pose. The listener pose may accordingly correspond to the position of the user/listener in the scene.

[0076] The listener pose processor 409 is further coupled to a transmitter 411 which is arranged to transmit the listener pose to the edge device 201.

[0077] In the example, the edge device 201 is arranged to generate an audio data signal and transmit it to the end device 203 where the audio data signal includes an Ambisonic audio mix as well as a number of audio objects. The edge device 201 is arranged to determine whether some audio elements are included in the audio data signal as part of an Ambisonic audio mix or as audio objects dependent on the listener pose.

[0078] The edge device 201 comprises a designator 305 which is arranged to designate at least one of the received audio objects as a close audio object or as a remote audio object depending on the listener pose. Specifically, for a given audio object, the designator 305 may determine a distance measure indicative of the distance between a pose of the first audio object and the listener pose. The distance measure is then compared to a threshold and if the distance measure exceeds the threshold (indicating that the distance is larger than a given value), the audio object is designated as a remote object and otherwise it is designated as a

close object. It will be appreciated that in some embodiments, the audio object may include designation into other possible categories including subcategories of the designation of the audio object as a close or remote object.

[0079] The distance may be any suitable distance measure which indicates a distance between the listener pose and the audio object pose in the audio scene, such as for example a Euclidian distance, a sum of absolute coordinate differences, etc.

[0080] The distance measure may have an increasing value for an increasing distance between the pose of the first audio object and the listener pose and the description will focus on such an example (e.g. when referring to comparisons to suitable thresholds).

[0081] The edge device 201 further comprises an audio mix generator 307 which is arranged to generate an Ambisonic audio mix from a plurality of the audio elements. The audio mix generator 307 is arranged to include or not the audio object in the Ambisonic audio mix depending on whether it is designated as a close audio object or as a remote object. In particular, if the audio object is designated as remote audio object, it is included in the Ambisonic audio mix but if it is designated as a close audio object it is not included in the Ambisonic audio mix.

[0082] The audio mix generator 307 may in many embodiments be arranged to include multiple, and in many cases all, audio objects that are designated as remote audio objects. It may in many embodiments be arranged to include no audio objects that are designated as close audio objects.

[0083] Thus, the audio mix generator 307 is arranged to generate an Ambisonic audio mix that includes audio objects that are designated as remote audio objects.

[0084] Further, the audio mix generator 307 is in many embodiments arranged to include other types of audio elements into the Ambisonic audio mix, such as for example received Ambisonic audio mixes, and indeed in some embodiments the audio mix generator 307 is arranged to add the audio objects designated as remote audio objects to an existing (received) Ambisonic audio mix. The Ambisonic audio mix may in some embodiments also be generated to include e.g., channel based audio signals, etc.

[0085] In some embodiments, the audio mix generator 307 may be coupled to a first audio signal generator 309 which may be arranged to generate a first audio signal for transmission to the end device 203. The first audio signal represents the Ambisonic audio mix and in many cases the generated Ambisonic audio mix may be transmitted directly without modification. However, in other embodiments the Ambisonic audio mix may be processed to generate a different representation, such as for example to provide a binaural representation.

[0086] The Ambisonic audio mix/first audio signal is fed to a data signal generator 311 which is arranged to generate an audio data signal comprising the Ambisonic

audio mix (either represented directly as the Ambisonic audio mix or possibly as represented by a set of binaural signals etc).

[0087] The audio objects that are designated as close audio objects are fed to the generator 311 and are included in the audio data signal as audio objects. In some cases, the edge device 201 further comprises a second audio signal generator 313 which may generate a second audio signal providing a suitable representation of the audio objects, such as e.g., by encoding a binaural representation.

[0088] The generator 311 may then generate the audio data signal to include the Ambisonic audio mix as well as any audio objects that are designated as close audio objects. Further, the audio data signal may be generated to not include any audio objects that are designated as remote audio objects, but these audio objects may instead be included in an Ambisonic audio mix.

[0089] The end device 203 may then receive the audio data signal and process the received Ambisonic audio mix and audio objects to generate a binaural output signal for the current listener pose.

[0090] It will be appreciated that in many embodiments, the audio data signal may include other audio elements than the Ambisonic audio mix and the audio objects. For example, it may include channel based audio, diffuse background audio signals, other Ambisonic audio mixes etc. In such cases, the end device 203 may include functionality for rendering such signals and combining them with the audio signals generated from the Ambisonic audio mix and audio objects.

[0091] The approach provides a particularly advantageous distribution of processing and an advantageous selection of audio data to transmit from an edge device 201 to an end device 203 in many embodiments. It may typically allow an advantageous user experience with high audio quality and fast adaptation while requiring relatively little computational resource in the edge device 201.

[0092] Ambisonics is an efficient format to represent a (large) set of audio objects. In particular when these audio objects are diffuse, it suffices to use a low order Ambisonics mix. The attainable spatial resolution of audio objects in an Ambisonics mix is directly coupled to the Ambisonics order. It is noted that a higher attainable spatial resolution will then apply to the complete Ambisonics mix, even in the case there is only a single audio object present in the Ambisonics mix. It is therefore not efficient to represent a limited set of point sources as an Ambisonics mix of sufficient order to attain a certain spatial resolution. Instead, it is more efficient to represent a limited set of point sources as separate audio objects, optionally in combination with an Ambisonics mix.

[0093] A scene rendered from an Ambisonics mix can be considered as "existing" on a sphere. Consequently, when the user moves closer to a specific audio object on that sphere, the audio objects will remain on the sphere, albeit rendered with a different gain. In other words, when

moving towards the sphere, the sphere basically moves along with the user. Audio objects rendered using an Ambisonics mix are therefore not "approachable". Note that instead of the user approaching a point source audio object, the point source audio object may also 'approach' the user, for example dictated by the trajectory of a specific object. As soon as the audio objects become approachable these will typically not be well represented using an Ambisonics mix, independent of the order.

[0094] An Ambisonics mix may be rendered onto a set of loudspeakers or efficiently converted into a binaural signal for reproduction on a headphone. Inherently, an Ambisonics mix is easily converted to account for (3D) rotations of the user. That is, in the case the user rotates his head by a certain angle in any (3D) direction, the Ambisonics mix may be efficiently updated to account for the rotation. Consequently, the user will experience an updated Ambisonics mix reflecting their head rotation.

[0095] However, very different from rotations, translations (change in position) cannot be efficiently compensated for in the Ambisonics mix. There is no efficient method to accommodate translations, in particular when audio objects become approachable in response to e.g., a user's translation, such as a user moving close towards a point source or an audio object in the Ambisonics mix approaching the user. As illustrated in FIG. 5, for audio objects that are closer to the user (object 1), a translation (x) of the user will have a larger effect on the perceived angle to the audio object than for audio objects that are further away from the user (object 2). For object 3, that is located more to the side, the translation (x) of the user only has a small effect on the perceived angle to the audio object. However, the translation may render the audio object approachable since the distance between object 3 and the (translated) user falls below a threshold. For the new position of the user a new shape (circle in this example) can be defined for approachable objects. Object 1, although having a larger effect on its perceived angle, may move out of the (new) circle in this example.

[0096] Since an Ambisonics mix is an efficient representation for (preferably diffuse) audio objects, one approach could be to disregard the effect of (small) user translations and not update the Ambisonics mix accordingly.

[0097] However, when audio objects are point sources, specifically when these are approachable and/or when there is also a visual component associated with the source, there is a clear benefit of an accurate representation of the audio object. In the case of a corresponding visual component, even a match between the audio and visual component is highly desirable for a proper user experience. This requires a dynamic tradeoff between representing audio objects separately or as part of an Ambisonics mix.

[0098] One approach for split rendering could be to pre-render all content into an Ambisonic mix based on a predicted listener pose and finally rendering from Ambisonics to binaural on the end device. Since the user's

three translational directions of freedom (X, Y, Z) are not represented in an Ambisonics rendering, this may result in an increased round-trip delay (from user translation to receiving the updated pre-rendered Ambisonics representation from the edge device). Even if the round-trip delay is acceptable, specifically for sources that are relatively close to the user, such translation changes can result in audible artefacts as they are more likely to result in significant changes in the source's angle of incidence which is an important perceptual cue for assessing the location of a source.

[0099] Therefore, the Inventors have realized that an approach where a scene may be pre-rendered properly to the Ambisonics format based on the last known listener pose is not suitable for scenarios where the listener is able to get close to sources or significantly change their distance to the sound source quickly.

[0100] Using an approach of pre-rendering multiple binaural signal pairs at the edge device and interpolating the ultimate binaural pair based on those signals at the end device also tends to be suboptimal both in terms of resulting audio quality as well as computational complexity required for pre-and post-rendering. In such cases, the output sound quality depends on the ultimate pose offset. For example, using variants rotated by 15 degrees to interpolate output at the ultimate pose, severe artefacts start to appear when exceeding 20 degrees rotation. The artefacts from interpolation are proportional to the minimum angle between the available poses and the actual pose. Furthermore, the proposed approach will have artefacts if the listener's pose changes in other directions than yaw, such as pitch and roll rotations of the user's head.

[0101] The previously described approach where a dynamic adaptation between representing different audio objects, and thus corresponding audio sources, as parts of an Ambisonic audio mix or as separate audio objects may provide an improved and particularly advantageous approach in many embodiments as it may address many of the above mentioned issues.

[0102] The approach may be illustrated by FIG. 6 which illustrates an example of a scene composed of a number of audio objects (1, 2, 3, 4, 5). Some audio objects are included as part of the Ambisonic audio mix (3, 4, 5) and other audio objects (1, 2) are separately transmitted as separate audio objects for independent binaural rendering. The Ambisonic audio mix and separate audio objects are included in the alternative audio data and transmitted to the end-device. The Ambisonic audio mix may be provided as a binauralized mix generated by the second audio signal generator. The audio objects are rendered (binauralized or rendered to loudspeakers) at the end-device and subsequently combined with the pre-rendered Ambisonic audio mix to produce the rendered output for consumption over one or more transducers, typically a headphone or AR/VR set. Specifically, the Ambisonic audio mix may be binauralized or rendered to loudspeakers for the combination with audio objects.

[0103] In the example, the set of audio objects that are included in the Ambisonic audio mix and are respectively represented individually may be varied dynamically. For example, in FIG. 7, the trajectory may bring audio object/-source 4 closer to the listener pose and specifically it may move within the threshold and be changed from being designated as a remote object to being designated as a close audio object. Accordingly, it may be removed from the Ambisonic audio mix and introduced as a separate audio object in the audio data signal thereby allowing a more accurate and dedicated rendering by the end device 203.

[0104] The change in the relative position and distance may equivalently occur by a movement, and specifically a translation, of the user/listening position. For the audio objects included in the Ambisonics mix (3, 4, 5), translation (X, Y, Z) of the user does not result in the expected directional change of the audio object until after the round-trip delay, i.e., after the user translation has been accommodated in an update of the Ambisonic audio mix (the translation is reported in the listener pose transmitted to the edge device 201 where the Ambisonic audio mix is modified to reflect the new position resulting in the transmitted Ambisonic audio mix corresponding to the new position). As indicated earlier, such a translation and disparity between the listener pose for which the Ambisonic audio mix is generated, and the new listener pose is much less perceptible for audio sources that are further away than for audio sources that are proximal or approachable. Both, the spatial smearing from the Ambisonics representation and the round-trip delay would make it more difficult to accurately localize a close audio source by trying to approach it. The distance threshold for inclusion of audio sources and objects in the Ambisonic audio mix may be determined as a suitable threshold at which point an audio object starts to become approachable or where a worst-case user translation (during the round-trip delay) results in a source position error that is deemed unacceptable.

[0105] The distance from the user to the individual audio objects is affected by movements of the user and the audio objects themselves. For instance, in the case the user physically or virtually moves in the direction of object 4 in FIG. 6, this audio object will transition into the circle representing the distance threshold and other audio objects may transition out of the circle (e.g., audio object 1). Alternatively, a similar effect is obtained in the case the trajectory of object 4 enters the circle. In that case, an audio object may become approachable without any translation of the user. Accordingly, the representation of audio objects as respectively a component in an Ambisonic audio mix or as a separate audio object may be updated dynamically. In many embodiments, such switching may include an element of hysteresis.

[0106] In the approach, as soon as an audio object passes or approaches the threshold, the audio object may transition from the Ambisonic audio mix to an audio object or vice versa. A smaller distance means the object

will transition from Ambisonics into an audio object when approaching. These transitions may be seamless, such that audio objects are cross faded from the Ambisonics mix into a separately transmitted audio object and vice versa.

[0107] During these transitions, because of perceptual masking of the audio object in the Ambisonic audio mix, the resolution of the separately transmitted audio object may be advantageously reduced, i.e., a lower bit-rate may be used for coding the audio object during such a transition.

[0108] At the point that an audio object transitions out or into an Ambisonic audio mix, the portion of the audio object that is represented as a separate audio object may be allocated a lower bitrate. The bitrate may increase as the audio objects transitions further out of the Ambisonic audio mix. Also masking (by the Ambisonic audio mix and/or other separate objects) may be taken into account when determining the bitrate required to represent the audio object that is transitioning from the Ambisonic audio mix.

[0109] Thus, in some embodiments, the audio mix generator is arranged to vary the data rate for the first audio object when transitioning between being designated as a close object and being designated as a remote object, and thus when transitioning between being part of the Ambisonic audio mix and not being part of the Ambisonic audio mix. The transitioning may be gradual.

[0110] The designation of audio objects as close or remote objects may in different embodiments advantageously consider different parameters and properties.

[0111] In many embodiments, the designation of an audio object may be dependent on a loudness measure for the audio object. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a loudness measure for the first audio object. Thus, in some embodiments at least one of the distance measure and the threshold is dependent on a loudness measure for the audio object.

[0112] For example, the louder a given audio object is, the more likely it is to be considered a close audio object than a quieter audio object. Thus, the threshold may be a function of the loudness of the audio object and specifically be a monotonically increasing function of the loudness (or equivalently the distance measure could be made a monotonically decreasing function of the loudness). Thus, in many embodiments, the distance at which an audio object is considered a close audio object may be increased for increasing loudness.

[0113] In many embodiments, such an approach may provide advantageous performance and may allow adaptation to provide a perceptually more consistent experience.

[0114] In some embodiments, the designation may further consider loudness of one or more other audio objects and specifically a relative loudness may be considered. This may for example allow masking effects etc.

from other audio sources to be considered.

[0115] Specifically, depending on the loudness and/or temporal/spatial masking of an audio object, the threshold at which distance measure the audio object moves from being included in the Ambisonic audio mix to be represented as a separate audio object (and vice versa) may be adapted. For example, for a loud (and therefore likely prominent) audio object, the threshold may be changed to increase the likelihood of the audio object being designated a close audio object, whereas for an audio object that is masked by another object, the threshold for the masked (less critical) object may be changed to reduce the likelihood of it being designated a close audio object (thus the threshold may be decreased for an increasing loudness and increased for a decreasing loudness).

[0116] In many embodiments, the designation of an audio object may be dependent on a diffuseness measure for the audio object. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a diffuseness measure for the first audio object. Thus, in some embodiments, at least one of the distance measure and the threshold is dependent on a diffuseness measure for the audio object.

[0117] In particular, since diffuse audio objects have poorer localization than point source audio objects, the threshold may depend on a measure of diffuseness of the audio object. For example, in the case (for a specific angle of incidence) the threshold distance (radius) for a point source audio object amounts to d_p , the threshold distance for a diffuse audio object may be $< d_p$ (e.g., $0.75 \cdot d_p$).

[0118] In many embodiments, the designation of an audio object may be dependent on a direction from the listener pose to the position of the audio object. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a direction from the listener pose to the position of the audio object. Thus, in some embodiments at least one of the distance measure and the threshold is dependent on a direction from the listener pose to the position of the audio object.

[0119] Audio objects that are represented with a frontal angle of incidence with respect to the listener can be better localized compared to audio objects represented with, for example, a rear angle of incidence. Therefore, the threshold, previously represented in the figures by a circle (i.e., distance independent of the direction), may be represented by an asymmetric shape. For example, in the case the threshold distance (radius) from the frontal angle of incidence amounts to d , the threshold distance for audio objects with a rear angle of incidence may be $< d$ (e.g., $0.5 \cdot d$). For all the other directions, the threshold distance may seamlessly transition between the front and rear angles of incidence. Typically, the threshold pattern may correspond to the localization accuracy as a function of the angle of incidence.

[0120] In many embodiments, the designation of an audio object may be dependent on a trajectory of a position of the audio object in the audio scene. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a trajectory for the first audio object. Thus, in some embodiments at least one of the distance measure and the threshold is dependent on a trajectory for the audio object.

[0121] The trajectory may be a trajectory relative to the listener pose and thus may be a relative trajectory caused by the movement of the audio object and/or the listener pose in the audio scene.

[0122] The designator 305 may for example keep track of the position of the audio object to determine whether it is moving toward the listener pose or away from it. It may in particular based on the trajectory estimate whether the audio object is likely to move towards the listener pose such that it is likely to be designated as a close object. If so, the distance threshold may for example be increased to result in an earlier designation of the audio object and an earlier removal from the Ambisonic audio mix and inclusion as a separate audio object.

[0123] The speed of the trajectory of the audio object may also impact the threshold at which an object moves from being designated as a close object to being designated as a remote object, or vice versa. For example, for a substantially stationary object, the threshold may remain unchanged, whereas for a fast moving object, the threshold may be increased such that the object is timely represented as a separate object. Different criteria may be used to determine the object's velocity, either based on metadata or prediction, such as e.g.:

- A listener pose to audio object position velocity vector (anticipating, likelihood of change of angular distance)
- Animated (moving) point source audio object

[0124] In many embodiments, the designation of an audio object may be dependent on an order of the Ambisonic audio mix. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of an order of the Ambisonic audio mix. Thus, in some embodiments at least one of the distance measure and the threshold is dependent on the order of the Ambisonic audio mix.

[0125] In some embodiments, the distance may consider the order of the Ambisonic audio mix and specifically the higher the order of the Ambisonic audio mix, the more likely an audio object may be considered a remote audio object and thus included in the Ambisonic audio mix. The higher the order of the Ambisonic audio mix, the better the Ambisonic audio mix may represent audio objects and accordingly the more appropriate it may be to include an increasing number of audio objects in the Ambisonic audio mix.

[0126] In many embodiments, the designator 305 may

also consider the number of audio objects that are included in the Ambisonic audio mix, and specifically the number of audio objects that are designated as remote audio objects. For example, a maximum or preferred number of audio objects for a given order be determined and the number of remote audio objects that are included in the Ambisonic audio mix may be limited to this maximum number.

[0127] In some embodiments, the order of the Ambisonic audio mix may be adapted depending on the number of audio objects that are designated as remote audio objects (and which are to be included in the Ambisonic audio mix).

[0128] The threshold may include a balance between the Ambisonics order versus the number of audio objects. As indicated earlier, by increasing the Ambisonics order, the discernability of objects increases. For representing a high number of objects, it may therefore be beneficial to increase the Ambisonics order so that fewer objects need to be represented separately.

[0129] In many embodiments, the designation of an audio object may be dependent on an importance or priority indication for the audio object. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of an importance or priority indication for the first audio object. Thus, in some embodiments at least one of the distance measure and the threshold is dependent on an importance or priority indication for the audio object.

[0130] The importance/ priority indication may for example be determined based on a type of audio provided, may e.g., be manually assigned, and/or may e.g., be received with the received data for the input audio elements.

[0131] For example, an object representing a person talking, such as a dialog object, is likely to have a higher importance in the scene than a (background) object making some sound. Audio meta data may be used to classify the importance of an object and thereby influence the threshold at which point an object moves out from the Ambisonic audio mix to be represented as a separate audio object. Examples for aspects that may signal high-importance include:

- In the case an audio object represents a user, the need to render as a separate audio object is higher)
- Whether there is a visual component connected to the audio object
- Audio object size
- A flag or 'importance' measure indicated in the metadata.

[0132] In many embodiments, the designation of an audio object may be dependent on a distance between the position of the audio object and the position of another audio object. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a distance be-

tween the position of the audio object and the position of another audio object. Thus, in some embodiments at least one of either the distance measure or the threshold is dependent on a distance between the position of the audio object and the position of another audio object.

[0133] In some embodiments, the designation may be dependent on the position of one or more other audio objects. For audio objects that are close together (substantially co-located) or have a small angular distance with respect to the user, the ability to discern them as individual objects is reduced. Therefore, instead of representing all these co-located objects in the Ambisonic audio mix, co-located objects may be rendered as one or more combined audio objects. This may improve the approachability of such a group of co-located objects, reduce the bit-rate required to jointly represent these objects, and reduce the computational complexity for rendering them as individual audio objects.

[0134] For audio objects closer to each other or for one or more dominant audio objects in a group of close-by audio objects, it may not be required to transmit all the audio objects in the group or employ the same transmission rate for all these individual audio objects. Audio objects in a group of close-by audio objects of similar dominance, may be grouped into a single audio object.

[0135] From a computational complexity perspective, it may be desirable to limit the number of audio objects that are transmitted separately to a maximum. There could for example be separate maxima for the number of audio objects transmitted separately and for the number of audio objects that are anticipated to approach the threshold.

[0136] Separate audio objects could be represented as pre-rendered binaural signals. This could be done as part of a complexity vs. quality trade-off, depending on the end-device's capabilities.

[0137] For example, for low computational resources available, only a single representation of the audio object may be pre-rendered. For medium computational resources available, additional representations of the audio object may be pre-rendered.

[0138] In many embodiments, the designation of an audio object may be dependent on a property of the rendering/end device. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of a property of the rendering/end device. Thus, in some embodiments at least one of either the distance measure or the threshold is dependent on a property of the rendering/end device.

[0139] The property may specifically be a capability of the end device e.g., rendering algorithms available, processing power, etc.

[0140] For example, during setup or initialization, the end device 203 may transmit a data message to the edge device 201 which indicates a property, such as a capability, of the end device 203. For example, the edge device 201 may indicate a type or processing power of the edge device. The designator 305 may take such

information into account. For example, the maximum number of audio objects that can be processed by the end device 203 may be determined dependent on the processing power of the end device 203. The distance threshold may be set dependent on this, and specifically such that the number of audio objects that are separated represented in the audio data signal does not exceed the determined maximum number.

[0141] The distance threshold may in some embodiments depend on the capabilities of the end-device. For example, for an end-device with limited processing capabilities, it may be beneficial to change the overall threshold so as to reduce the number of audio objects that are represented as separate audio objects. Other trade-offs involving combinations of the above parameters may also be possible.

[0142] In some embodiments, the threshold may be changed dynamically to have a certain number of audio objects represented separately to optimize audio quality within the capabilities of the end-device.

[0143] Thus, the end device 203 processing capabilities may be taken into account. For example, for low cost devices, a limited set of pre-rendered audio objects may be provided.

[0144] In many embodiments, the designation of an audio object may be dependent on a data transfer property of the communication link for transmitting the audio data signal to the edge device 201. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of the data transfer property. Thus, in some embodiments at least one of either the distance measure or the threshold is dependent on the data transfer property. The data transfer property may specifically be a data rate for the communication.

[0145] For example, the edge device 201 may estimate a current bandwidth or throughput for the transmission of the audio data to the end device 203. It may then proceed to determine a maximum number of separate audio objects that can be transmitted (in addition to e.g., one Ambisonic audio mix) and proceed to adapt the distance threshold to ensure that the maximum number is not exceeded.

[0146] As another example, in a case where the bandwidth is constrained and the number of objects designated as close objects is relatively high, the distance threshold may be reduced so that less objects are designated as close objects. In addition, because of the bandwidth freed up by moving objects previously designated as close objects, the Ambisonic (HOA) order may be increased so that the objects that were previously designated as close objects will be 'better' represented in the Ambisonic audio mix.

[0147] In many embodiments, the designation of an audio object may be dependent on a plurality of listener poses, e.g., in a gaming scenario. Specifically, the designator 305 may be arranged to determine the threshold, or equivalently the distance measure, as a function of the

plurality of listener poses. Thus, in some embodiments at least one of either the distance measure or the threshold is dependent on the plurality of listener poses.

[0148] As an example, in some embodiments, the edge device 201 may receive listener poses from two, or more, end devices 203 and it may proceed to generate a single audio data signal that may be transmitted back to a plurality of the end devices 203. In such cases, the designation of the audio objects may consider a plurality of listener poses. For example, a distance threshold may be determined for each listener pose and an audio object may be designated as a close audio object if the distance to any of the listener poses is less than the corresponding threshold. Thus, in such embodiments, an acoustic object may be designated as a close object if it is close to any of the listener poses and may only be included in the Ambisonic audio mix if it is remote from all listener poses.

[0149] The distance threshold (or equivalently the distance measure) for an audio object may be determined as a function of many different parameters including one or more of the following:

- The distance of the audio object relative to the listener pose
- The (3D) angle of the audio object relative to the listener pose (angle of incidence)
- The diffuseness of the audio object (point source versus an object with extent)
- The location of other audio objects
 - Co-location of audio objects (small angular distance) can be rendered as one single audio object versus leaving in spatial audio mix
 - Loudness/temporal/spatial masking
- The speed of the audio object
 - User to Object velocity vector (anticipating, likelihood of change of angular distance)
 - Animated (moving) point source audio object
- The number of input objects
- Audio object meta data including:
 - Audio object type (i.e., in the case an audio object represents a user, the need to render it as a separate audio object is higher)
 - Whether there is a visual component connected to the audio object
 - Audio object size
- Dynamic balance between Ambisonics order versus number of audio objects
- Bitrate allocation to audio object during transition from Ambisonics mix to separate audio object. Using masking
- Capabilities of the end-device (e.g., processing capabilities)

[0150] FIG. 7 is a block diagram illustrating an example processor 700 according to embodiments of the disclosure. Processor 700 may be used to implement one or more processors implementing an apparatus as previously described or elements thereof. Processor 700 may be any suitable processor type including, but not limited to, a microprocessor, a microcontroller, a Digital Signal Processor (DSP), a Field Programmable Array (FPGA) where the FPGA has been programmed to form a processor, a Graphical Processing Unit (GPU), an Application Specific Integrated Circuit (ASIC) where the ASIC has been designed to form a processor, or a combination thereof.

[0151] The processor 700 may include one or more cores 702. The core 702 may include one or more Arithmetic Logic Units (ALU) 704. In some embodiments, the core 702 may include a Floating Point Logic Unit (FPLU) 706 and/or a Digital Signal Processing Unit (DSPU) 708 in addition to or instead of the ALU 704.

[0152] The processor 700 may include one or more registers 712 communicatively coupled to the core 702. The registers 712 may be implemented using dedicated logic gate circuits (e.g., flip-flops) and/or any memory technology. In some embodiments the registers 712 may be implemented using static memory. The register may provide data, instructions and addresses to the core 702.

[0153] In some embodiments, processor 700 may include one or more levels of cache memory 710 communicatively coupled to the core 702. The cache memory 710 may provide computer-readable instructions to the core 702 for execution. The cache memory 710 may provide data for processing by the core 702. In some embodiments, the computer-readable instructions may have been provided to the cache memory 710 by a local memory, for example, local memory attached to the external bus 716. The cache memory 710 may be implemented with any suitable cache memory type, for example, Metal-Oxide Semiconductor (MOS) memory such as Static Random Access Memory (SRAM), Dynamic Random Access Memory (DRAM), and/or any other suitable memory technology.

[0154] The processor 700 may include a controller 714, which may control input to the processor 700 from other processors and/or components included in a system and/or outputs from the processor 700 to other processors and/or components included in the system. Controller 714 may control the data paths in the ALU 704, FPLU 706 and/or DSPU 708. Controller 714 may be implemented as one or more state machines, data paths and/or dedicated control logic. The gates of controller 714 may be implemented as standalone gates, FPGA, ASIC or any other suitable technology.

[0155] The registers 712 and the cache 710 may communicate with controller 714 and core 702 via internal connections 720A, 720B, 720C and 720D. Internal connections may be implemented as a bus, multiplexer, crossbar switch, and/or any other suitable connection technology.

[0156] Inputs and outputs for the processor 700 may be provided via a bus 716, which may include one or more conductive lines. The bus 716 may be communicatively coupled to one or more components of processor 700, for example the controller 714, cache 710, and/or register 712. The bus 716 may be coupled to one or more components of the system.

[0157] The bus 716 may be coupled to one or more external memories. The external memories may include Read Only Memory (ROM) 732. ROM 732 may be a masked ROM, Electronically Programmable Read Only Memory (EPROM) or any other suitable technology. The external memory may include Random Access Memory (RAM) 733. RAM 733 may be a static RAM, battery backed up static RAM, Dynamic RAM (DRAM) or any other suitable technology. The external memory may include Electrically Erasable Programmable Read Only Memory (EEPROM) 735. The external memory may include Flash memory 734. The External memory may include a magnetic storage device such as disc 736. In some embodiments, the external memories may be included in a system.

[0158] It will be appreciated that the above description for clarity has described embodiments of the invention with reference to different functional circuits, units and processors. However, it will be apparent that any suitable distribution of functionality between different functional circuits, units or processors may be used without detracting from the invention. For example, functionality illustrated to be performed by separate processors or controllers may be performed by the same processor or controllers. Hence, references to specific functional units or circuits are only to be seen as references to suitable means for providing the described functionality rather than indicative of a strict logical or physical structure or organization.

[0159] The invention can be implemented in any suitable form including hardware, software, firmware or any combination of these. The invention may optionally be implemented at least partly as computer software running on one or more data processors and/or digital signal processors. The elements and components of an embodiment of the invention may be physically, functionally and logically implemented in any suitable way. Indeed the functionality may be implemented in a single unit, in a plurality of units or as part of other functional units. As such, the invention may be implemented in a single unit or may be physically and functionally distributed between different units, circuits and processors.

[0160] Although the present invention has been described in connection with some embodiments, it is not intended to be limited to the specific form set forth herein. Rather, the scope of the present invention is limited only by the accompanying claims. Additionally, although a feature may appear to be described in connection with particular embodiments, one skilled in the art would recognize that various features of the described embodiments may be combined in accordance with the inven-

tion. In the claims, the term comprising does not exclude the presence of other elements or steps.

[0161] Furthermore, although individually listed, a plurality of means, elements, circuits or method steps may be implemented by e.g., a single circuit, unit or processor. Additionally, although individual features may be included in different claims, these may possibly be advantageously combined, and the inclusion in different claims does not imply that a combination of features is not feasible and/or advantageous. Also the inclusion of a feature in one category of claims does not imply a limitation to this category but rather indicates that the feature is equally applicable to other claim categories as appropriate. Furthermore, the order of features in the claims do not imply any specific order in which the features must be worked and in particular the order of individual steps in a method claim does not imply that the steps must be performed in this order. Rather, the steps may be performed in any suitable order. In addition, singular references do not exclude a plurality. Thus, references to "a", "an", "first", "second" etc. do not preclude a plurality. Reference signs in the claims are provided merely as a clarifying example shall not be construed as limiting the scope of the claims in any way.

Claims

1. An audio apparatus comprising:

a receiver (301) arranged to receive a plurality of audio elements representing an audio scene, the audio elements including a number of audio objects, each audio object being linked with a position in the audio scene;

a listener pose receiver (303) arranged to receive an indication of a listener pose; a designator (305) arranged to designate at least a first audio object of the number of audio objects as a close audio object or as a remote audio object depending on a comparison of a distance measure indicative of a distance between a pose of the first audio object and the listener pose to a threshold;

an audio mix generator (307) arranged to generate an Ambisonic audio mix from a first plurality of the audio elements, the first plurality of audio elements comprising the first audio object if this is designated as a remote audio object;

a data generator (311) arranged to generate an audio data signal comprising the Ambisonic audio mix; and wherein the data generator is arranged to include the first audio object in the audio data signal if the first audio object is designated as a close audio object.

2. The audio apparatus of claim 1 wherein the designator (305) is arranged to designate the first audio

object as a close audio object or as remote audio object in dependence on a loudness measure for the first audio object.

3. The audio apparatus of any previous claim wherein the designator (305) is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a direction from the listener pose to a position of the first audio object. 5
4. The audio apparatus of any previous claim wherein the designator (305) is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a trajectory of a position of the first audio object in the audio scene. 10
5. The audio apparatus of any previous claim wherein the designator (305) is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on an order of the Ambisonic audio mix. 15
6. The audio apparatus of any previous claim wherein the designator (305) is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a distance between the position of the first audio object and a position of a second audio object. 25
7. The audio apparatus of any previous claim wherein the data generator (311) is arranged to transmit the audio data signal to a remote apparatus over a communication link; and the designator (305) is arranged to designate the first audio object as a close audio object or as remote audio object in dependence on a data transfer property of the communication link. 30
8. The audio apparatus of any previous claim wherein the audio mix generator (307) is arranged to vary a data rate for the first audio object when transitioning between being designated as a close object and being designated as a remote object. 35
9. The audio apparatus of any previous claim wherein the listener pose receiver (303) arranged to receive a plurality of listener poses, the designator (305) is arranged to determine the distance measure in dependence on the plurality of listener poses, and the data generator (311) is arranged to transmit the audio data signal to a plurality of remote devices. 40
10. An audio system comprising an audio apparatus (201) in accordance with any previous claims and a rendering device (203); wherein the data generator (311) is arranged to transmit the audio data signal to the rendering device (203); and the rendering device (203) comprises: 45

a receiver (401) arranged to receive the audio data signal from the audio apparatus (201); and a renderer (403) arranged to generate a binaural audio signal representing the audio scene, the binaural audio signal comprising contributions from the first audio mix and the first audio object if present in the audio data signal.

11. The audio system of claim 10 wherein the rendering device (203) is a headset comprising audio transducers arranged to reproduce the binaural rendering signal. 50
12. The audio system of claim 10 or 11 wherein the designator (305) is arranged to designate the first audio object as a close audio object or as a remote audio object in dependence on a property of the rendering device. 55
13. A method of operation for an audio apparatus, the method comprising:

receiving a plurality of audio elements representing an audio scene, the audio elements including a number of audio objects, each audio object being linked with a position in the audio scene;
 receiving an indication of a listener pose;
 designating at least a first audio object of the number of audio objects as a close audio object or as a remote audio object depending on a comparison of a distance measure indicative of a distance between a pose of the first audio object and the listener pose to a threshold;
 generating an Ambisonic audio mix from a first plurality of the audio elements, the first plurality of audio elements comprising the first audio object if this is designated as a remote audio object; and
 generating an audio data signal comprising the Ambisonic audio mix; and including the first audio object in the audio data signal if the first audio object is designated as a close audio object.

14. A method of operation for an audio apparatus (201), the audio apparatus performing the method of claim 13 and transmitting the audio data signal to a rendering device (203); and the rendering device (203) performs the steps of: 50

receiving the audio data signal from the audio apparatus (201); and
 generating a binaural audio signal representing the audio scene, the binaural audio signal comprising contributions from the first audio mix and the first audio object if present in the audio data signal. 55

15. A computer program product comprising computer program code means adapted to perform all the steps of claim 13 or 14 when said program is run on a computer.

5

10

15

20

25

30

35

40

45

50

55

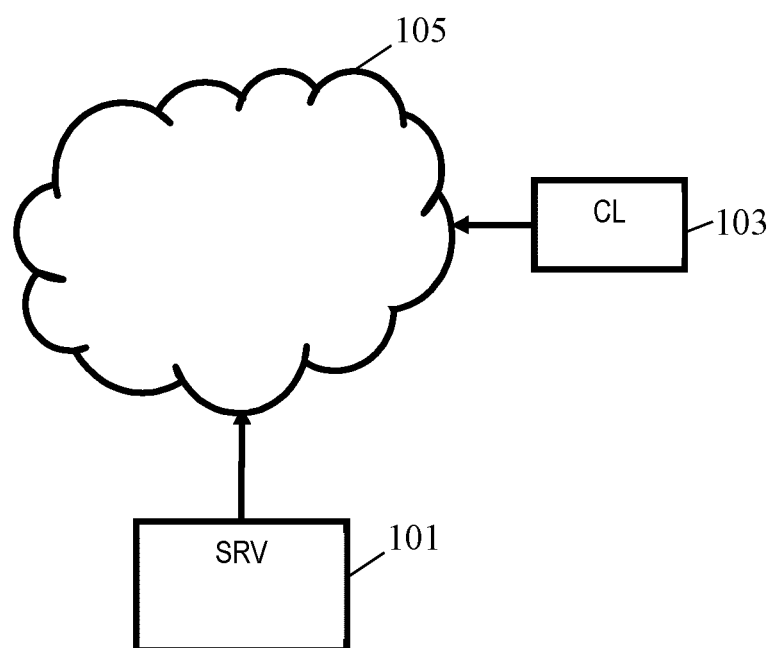


FIG. 1

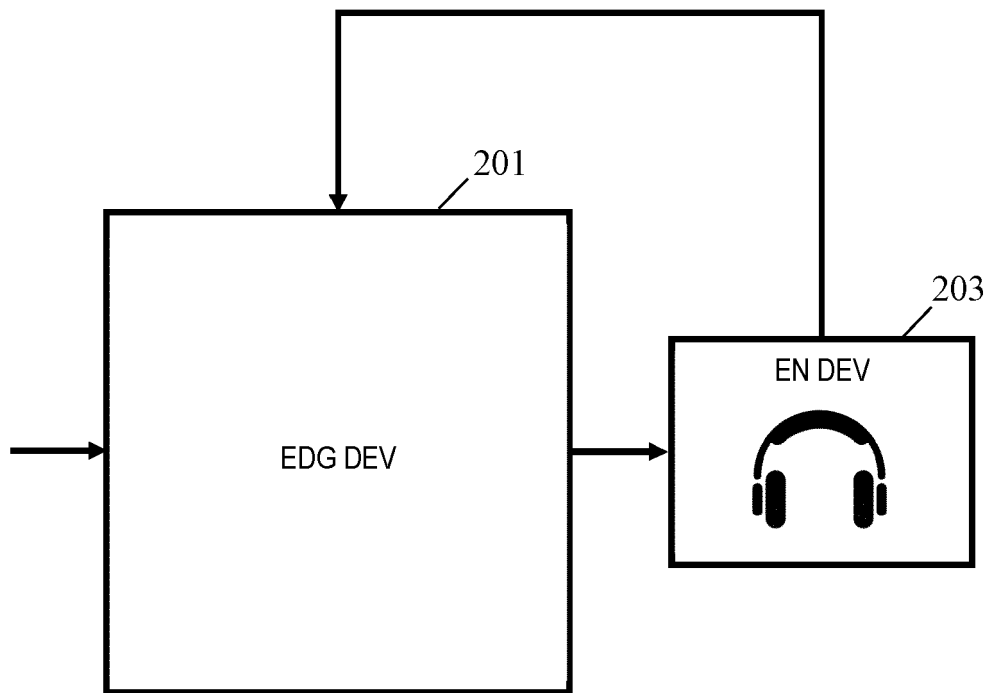


FIG. 2

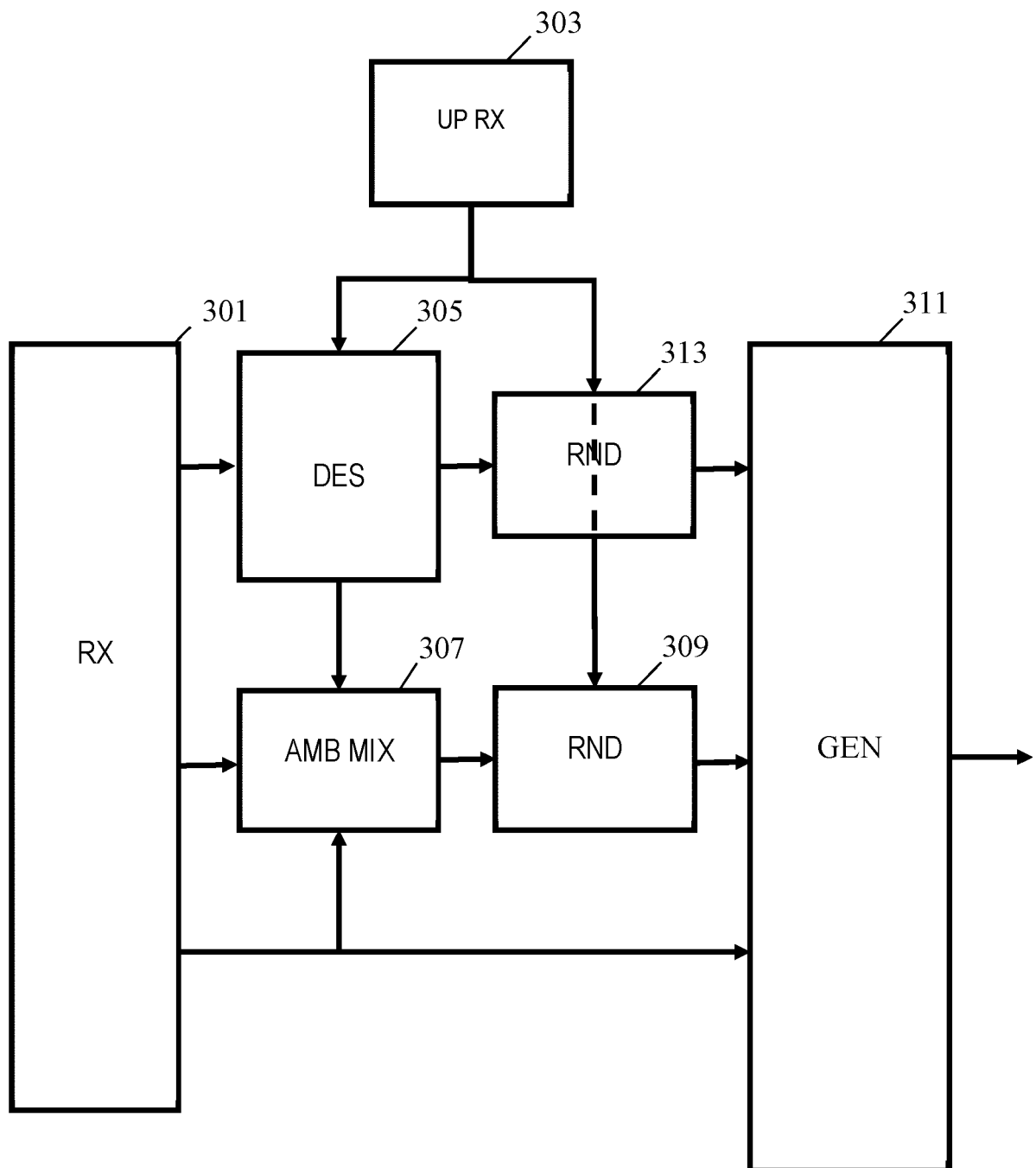


FIG. 3

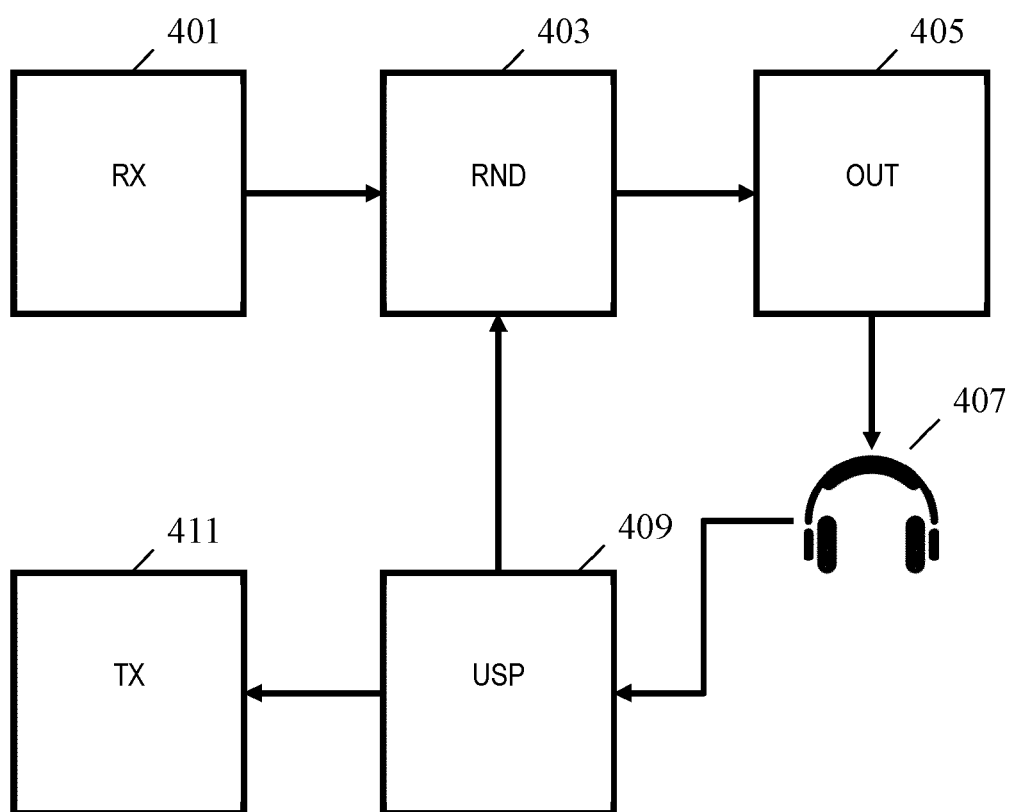


FIG. 4

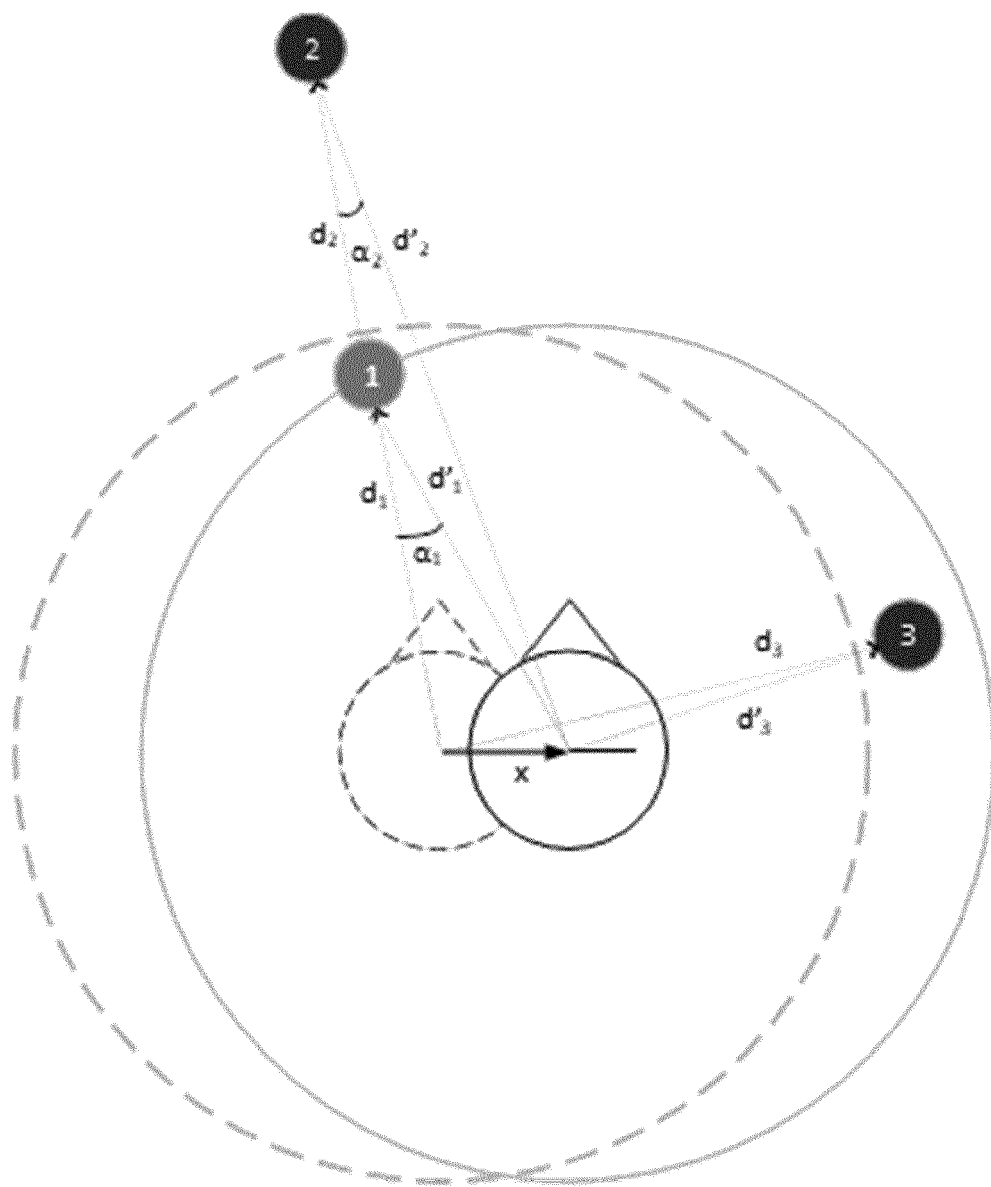


FIG. 5

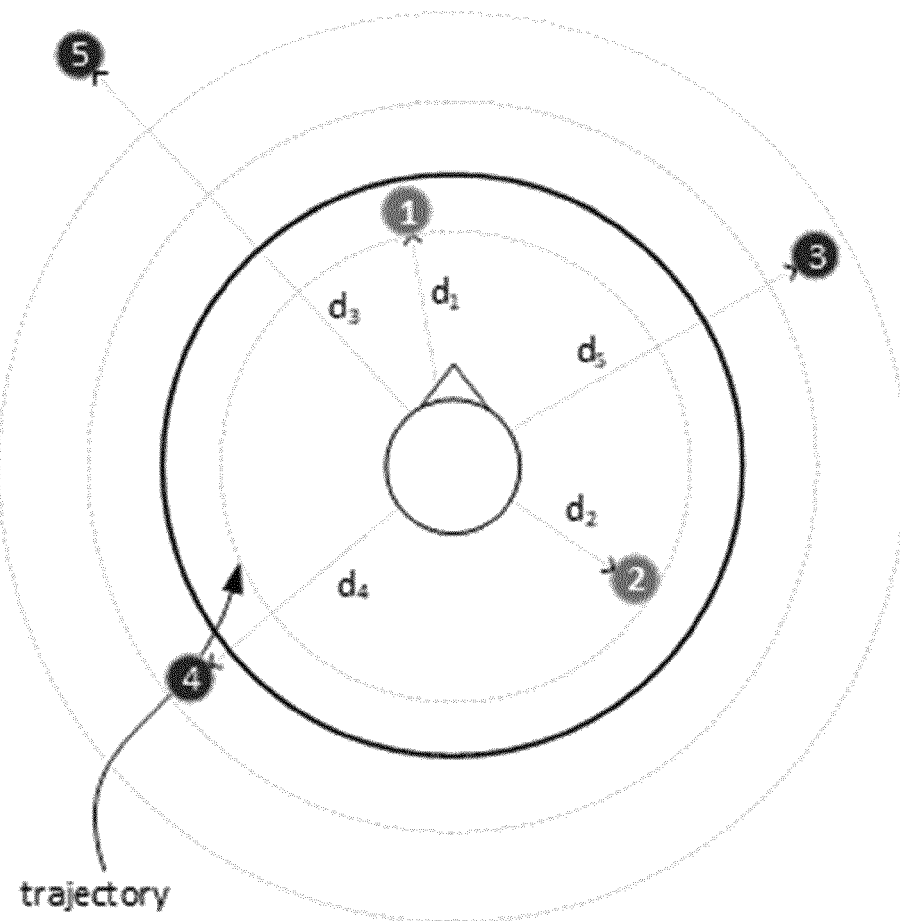


FIG. 6

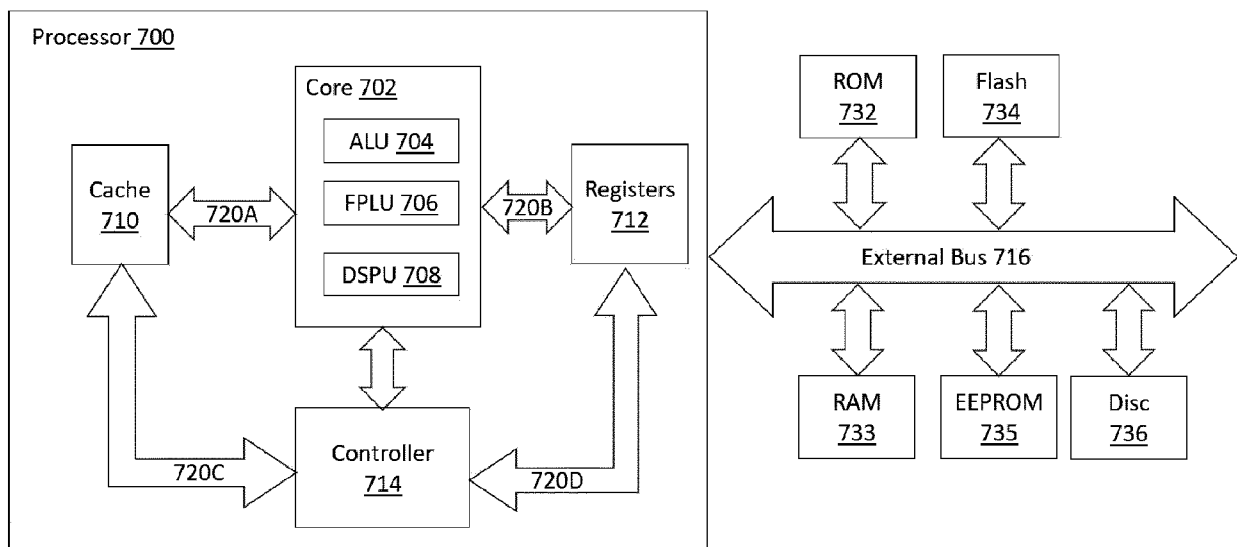


FIG. 7



EUROPEAN SEARCH REPORT

Application Number

EP 23 20 1344

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 2017/366913 A1 (STEIN EDWARD [US] ET AL) 21 December 2017 (2017-12-21) * paragraph [0003] - paragraph [0209]; claims 1-20; figures 1-24 *	1-15	INV. H04S7/00
A	US 2022/053282 A1 (DE BRUIJN WERNER PAULUS JOSEPHUS [SE] ET AL) 17 February 2022 (2022-02-17) * paragraph [0019] - paragraph [0172]; claims 1-29; figures 1-4 *	1-15	
A	US 2022/174447 A1 (KOPPENS JEROEN GERARDUS HENRICUS [NL]) 2 June 2022 (2022-06-02) * paragraph [0019] - paragraph [0177]; claims 1-21 *	1-15	
			TECHNICAL FIELDS SEARCHED (IPC)
			H04S G10L G06F G06T
The present search report has been drawn up for all claims			
Place of search		Date of completion of the search	Examiner
The Hague		1 March 2024	Timms, Olegs
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document			

EPO FORM 1503 03.82 (P04C01)

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 23 20 1344

5

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

01-03-2024

10

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2017366913 A1	21-12-2017	CN 109891502 A	14-06-2019
		EP 3472832 A1	24-04-2019
		JP 7039494 B2	22-03-2022
		JP 2019523913 A	29-08-2019
		KR 20190028706 A	19-03-2019
		TW 201810249 A	16-03-2018
		US 2017366912 A1	21-12-2017
		US 2017366913 A1	21-12-2017
		US 2017366914 A1	21-12-2017
		US 2019215638 A1	11-07-2019
		WO 2017218973 A1	21-12-2017
US 2022053282 A1	17-02-2022	BR 112021004495 A2	08-06-2021
		CN 112740150 A	30-04-2021
		EP 3623913 A1	18-03-2020
		EP 3850470 A1	21-07-2021
		JP 2022500917 A	04-01-2022
		US 2022053282 A1	17-02-2022
US 2022174447 A1	02-06-2022	WO 2020053182 A1	19-03-2020
		CN 113614685 A	05-11-2021
		EP 3712788 A1	23-09-2020
		EP 3942400 A1	26-01-2022
		JP 2022525902 A	20-05-2022
		US 2022174447 A1	02-06-2022
		WO 2020187807 A1	24-09-2020

15

20

25

30

35

40

45

50

55

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82